Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

October/November 2021

Outline

- Part 1 - Probability Concepts and Their Relevance to Forensic Science
  - review of probability concepts
  - conditional probability and independence
  - Bayes' Theorem and likelihood ratio
- Part 2 - Sampling, Statistics and Forensics
  - collecting data
  - variability, reliability and accuracy
  - estimation and inference
- **Part 3 - Assessing the Probative Value of Forensic Evidence**
  - **forensic examinations as expert opinion**
  - **two-stage approach (significance test/coincidence probability)**
  - **likelihood ratio / Bayes factor**

Learning Objectives for Part 3

- Understand tools for evaluating forensic examinations as expert opinion
- Understand the two-stage approach to assessing forensic evidence
- Understand how the likelihood ratio approach might be applied to forensic evidence evaluation
- Understand strengths and limitations of the different approaches

Brief Review of Probability and Statistics

- Probability
  - language for describing uncertainty
  - assigns numbers between 0 and 1 to events
  - depends on information available
    (information conditioned upon)
  - useful for deducing likely observed values 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, substance ID, 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 / different 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

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 and its Role in Court

- 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
- Some states use Frye standard
- FRE 702 and its focus on reliable (trustworthy) methods
- NRC and PCAST reports

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

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 :)
- Recent DOJ statement argues that "black box" studies recommended by PCAST are not required
- Others argue that "black box" error rate estimates are of limited use in a particular case
- My argument would be that there is little doubt "black box" evaluations provide extremely useful information

Forensic Conclusions as Expert Opinions

- 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?)
- A reminder about interpretation
  - false positive rate is Pr(examiner makes identification | non-mates)
  - some studies look at Pr(non-mates | examiner makes identification)
  - both are interesting quantities .... but not the same

Forensic Conclusions as Expert Opinions

- False positive or false identification rate is Pr(examiner makes identification | non-mates)
- Now introduce a new concept often used with diagnostic tests, the positive predictive value (PPV) defined as Pr(mates | examiner makes identification)
- Some forensic studies look at Pr(non-mates | examiner makes identification); This is 1-PPV and is sometimes known as the false discovery rate
  - The false positive rate is a traditional error rate, it relies on knowing the ground truth of the case
  - The other quantity, 1-PPV, depends on the mix of cases that are shown to the examiner
    - PPV = Pr(mates | examiner ID) = True positive*Pr(mates) / (True positive*Pr(mates) + False positive*Pr(non-mates))
    - If an examiner only sees mated pairs, then positive predictive value is 1 and Pr(non-mates | examiner ID) = 0
    - But this does not mean error rate would be zero if they saw non-mates

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
  - But remember that "black box" studies speak to field rather than to a specific case
  - 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

Test yourself

Forensic conclusions as expert opinion

- Which of the following statements related to evaluating expert forensic conclusions are true?
  - If a black box study finds high validity (accuracy) then it must also show high reliability
  - If a black box study finds high reliability then it must also show high validity
  - A false identification error occurs when a known set of non-matching items are mistakenly identified as coming from the same source
  - We can estimate the false identification error rate of a forensic discipline by looking at all of the identifications made by examiners and then determining the fraction of non-matching pairs among those cases

Test yourself

Forensic conclusions as expert opinion - answer

- Answer not provided in this file

Test yourself

Forensic conclusions as expert opinion - answer

- Answer not provided in this file

Test yourself
Inconclusives

- In carrying out a black box study, there are always questions about how to treat inconclusives. Which of the following approaches do you support?
  - Inconclusives are always errors because the pair must either be a same source pair or a different source pair
  - Inconclusives should always be marked as correct because they are not errors
  - Inconclusives should be omitted from the error rate calculation
  - We need more than one summary of the study to accurately convey examiner performance
  - A well-designed study should include some cases for which expert consensus would indicate that inconclusive is the right answer

Test yourself
Inconclusives - answer

- Answer not provided in this file

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 may be described as "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
  - We will see (as noted earlier in the course) that there are conceptual problems with this approach
  - Note also 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 known as 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.$=.0435/\sqrt{5} = .019$
- Sample: mean $= .728$, std.err.$=.0230/\sqrt{5} = .010$
- 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 suggest rejecting this hypothesis
    - In this setting a Type I error is a false exclusion (not a "false positive" as typically assumed)
    - In this setting a Type II error is a false inclusion (not a "false negative" as typically assumed)
  - It seems we have the wrong null hypothesis, i.e., the wrong "starting point". In forensics, the null should perhaps be that the samples are distinguishable.
  - I will avoid using Type I and Type II errors in this discussion

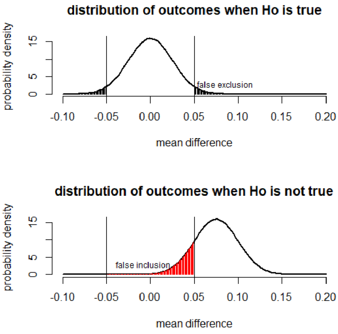
The Two-Stage Approach

- 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 false exclusion 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 false inclusion 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

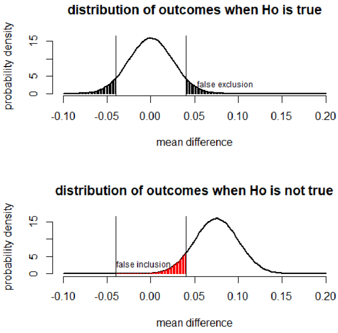
The Two-Stage Approach - thresholds / error rates

(Black=false exclusion; Red=false inclusion)

Initial threshold →
low false exclusion rate; higher false inclusion rate



Lowering the threshold →
increases false exclusion rate; lowers false inclusion rate



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 stages where the first stage may end the analysis is not optimal
  - Much recent attention to this issue in the statistics community
    - failing to find a significant difference is not the same as finding that the null hypothesis is true
    - finding a statistically significant difference may not be practically important
    - the major U.S. statistical association recently recommended not using the term ”statistically significant”
  - This issue is not unique to forensics and can also be found in medicine

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 Δ

$$H_0 : |\mu_{scene} - \mu_{suspect}| > \Delta$$
$$H_A : |\mu_{scene} - \mu_{suspect}| < \Delta$$

    - 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

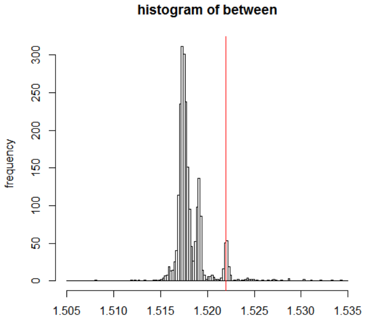
- Now return to the usual significance testing approach and assume we have found a “match” or found the glass to be ”indistinguishable” (i.e., could not reject the null hypothesis)
- Stage 2 - Identification
  - Assess the probability that two samples from different sources would ”match” or be found ”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

- Stage 2 ... the idea for 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)

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

- 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 types of glass (i.e., not just three)
- Stage 2 coincidence probability will depend on the distribution of the types of glass
- 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

Test yourself
Stage 1

- Applying the two-stage approach requires that we select a cutoff at the first stage to determine whether two samples are indistinguishable. Which of the following are true?
  - A high cutoff is best because it will make it difficult to eliminate a suspect
  - A low cutoff is best because it gives the benefit of the doubt to the suspect
  - We should choose a cutoff so that there no type I or type II errors
  - The statistician should get to choose the cutoff
  - Figuring out where to put the cutoff is a hard problem

Test yourself

Stage 1 - answer

- Answer not provided in this file

Test yourself
Stage 2

- Which of the following statements about stage 2 would you support?
  - Stage 2 is very important because we need to know how unusual it is to find indistinguishable samples
  - Stage 2 is not very important because once you have found a match, that is all you need to know
  - Stage 2 is difficult because the relevant population will vary from case to case
  - Stage 2 is not necessary; it should be left to the jury
  - I'm getting tired of this type of question

Test yourself

Stage 2 - answer

- Answer not provided in this file

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

- The goal for the trier of fact in courtroom setting is a decision about the relative likelihood of two hypotheses (e.g., same or different source) given data
- In statistical terms this suggests 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**
- Bayes' rule points us towards the likelihood ratio approach

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 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 technicalities
  - 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 but this would be very challenging
  - other available information (e.g., background) can be incorporated into the LR
  - confusion about the terms likelihood ratio / Bayes factor / Bayesian approach
    - both the likelihood ratio and Bayes factor are relevant in a Bayesian approach to evidence
    - distinction between LR and BF is quite technical and depends on how statistical parameters are treated
    - for our purposes ... the LR and BF are the same

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!

Likelihood Ratio Approach

A simple example

- Suppose evidence is blood types for a crime scene sample (y) and suspect sample (x)
- The source of crime scene sample is unknown (random); we want to assign probabilities to y given observed data and different hypotheses
- Information about the distribution of blood types in the U.S.

Type	A	B	AB	O
U.S. Freq	.42	.10	.04	.44

- 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

TH01	4	5	6	7	8	9	9.3	10	11
Freq.	.001	.001	.266	.160	.135	.199	.200	.038	.001

- For TH01 agreeing on alleles 7, 9, the probability of a random agreement is $2 * .16 * .199 = .064$ so LR ≈ 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)

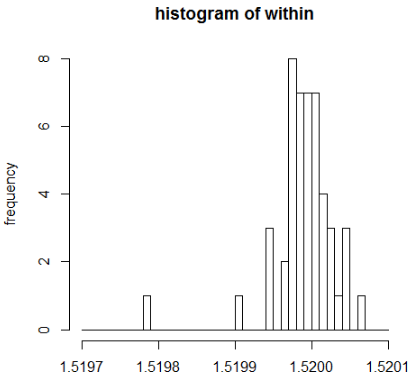
Likelihood Ratio Approach

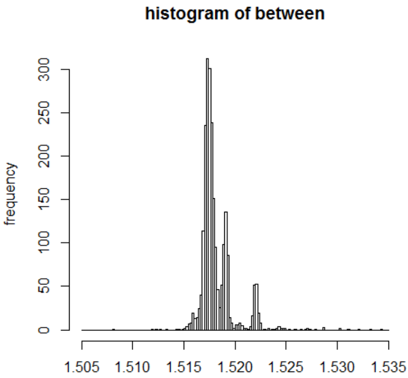
Where it can work .... Trace evidence

- 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

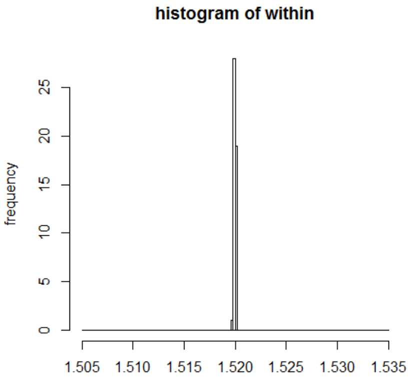
Where it can work .... Trace evidence

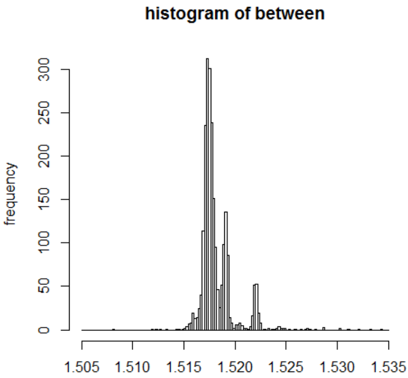




Likelihood Ratio Approach

Where it can work .... Trace evidence





Likelihood Ratio Approach

Continuous measures for trace data - example

- Conceptual discussion of the approach described by Aitken and Lucy (Applied Statistics, 2004)
- 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)
- Note that here the source of crime scene sample is known and the randomness is about whether y is from the same source
- 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 $\bar{S}$
  - 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)

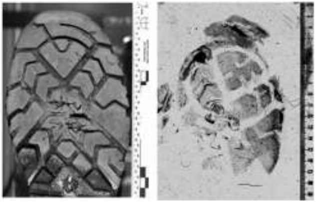
Likelihood Ratio Approach

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

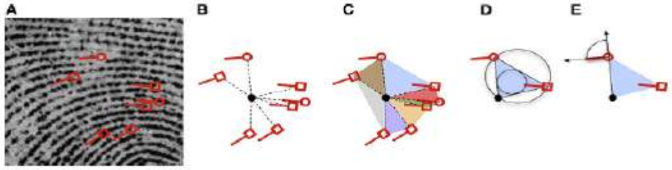
Where it might work .... Pattern evidence

- 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 $\Pr(E \mid \bar{S})$
  - 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 $\Pr(E \mid S)$ and $\Pr(E \mid \bar{S})$
- This is a very hard problem!!
    - Need to assign probabilities to all possible observations 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

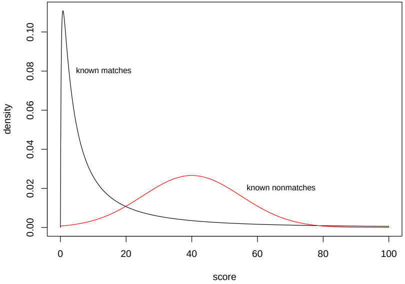
- LRs reported as numbers or as verbal equipments
- Verbal equivalents are less precise, but may be easier to understand

Value of LR	Verbal equivalent: "The forensic findings ...
1	do not support one proposition over the other
2 – 10	provide **weak support** for the same source proposition relative to the different source proposition
10 - 100	provide **moderate support** for the same source proposition relative ...
100 - 1000	provide **moderately strong support** for the same source proposition relative ...
1000 - 10000	provide **strong support** for the same source proposition relative ...
10000 – 1 mill.	provide **very strong support** for the same source proposition relative ...
1 million +	provide **extremely strong support** for the same source proposition relative ...

Likelihood Ratio Approach

Score-based likelihood ratios

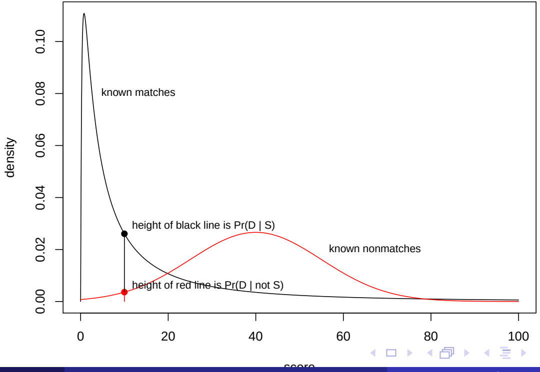
- Given the challenge in developing LRs for pattern evidence there has been recent work developing score-based approaches
- Define a score measuring the "difference" between the questioned and known samples (let's call the score D)
- Obtain the distribution of scores for a sample of known matches (i.e., 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 ($Pr(D \mid \bar{S})$)
  - Score-based likelihood ratio if we observe score D is $SLR = Pr(D \mid S)/Pr(D \mid \bar{S})$



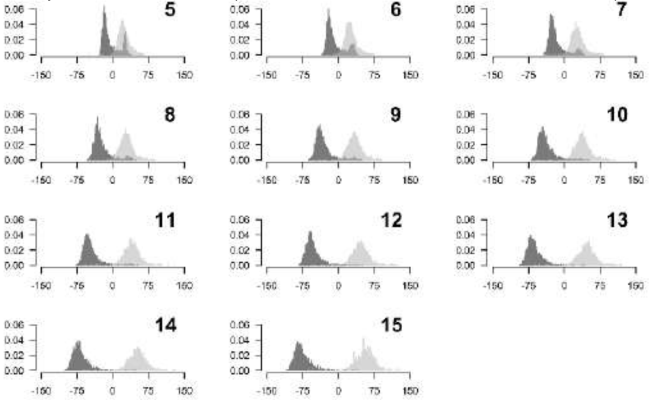
Likelihood Ratio Approach

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

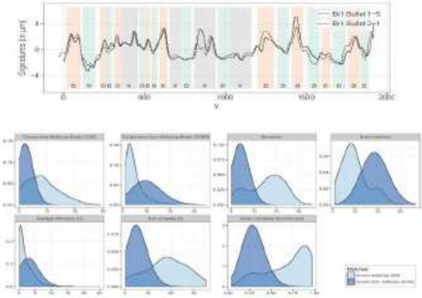
- FRSTATS (Swofford et al., Forensic Science International, 2018)



Likelihood Ratio Approach

Score-based likelihood ratios

- Often no single score is found that distinguishes well between same-source and different-source pairs
- Example - bullet land signatures

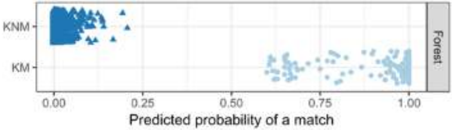


- A possible solution here is to combine them using machine learning techniques trained to distinguish between same-source and different-source pairs (Hare et al., AOAS, 2017)

Likelihood Ratio Approach

Score-based analyses - another approach

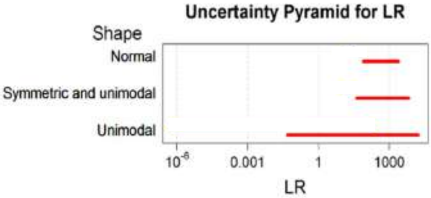
- Another approach is to use the scores (D) within the two-stage framework
  - If D (difference score) is above a threshold, then declare different source
  - If D (difference score) is below a threshold, then declare same source
  - Compute error rates for this type of procedure (perhaps for different 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"



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

Test yourself

Likelihood ratios - the numerator

- The numerator of the likelihood ratio ($\Pr(E \mid S)$) measures
  - The probability that the suspect committed the crime
  - The probability that the expert witness is trying to confuse the jury
  - The probability of observing evidence like the evidence in this case if the two samples came from the same source
  - The probability that the two samples came from the same source given the observed evidence

Test yourself

Likelihood ratios - the numerator - answer

- Answer not provided in this file

Test yourself

Likelihood ratios - denominator

- The denominator of the likelihood ratio ($Pr(E \mid \text{not } S)$) measures
  - The probability that the suspect did not commit the crime
  - The probability that the expert witness is trying to confuse the jury
  - The probability of observing evidence like the evidence in this case if the two samples came from different sources
  - The probability that the two samples came from different sources given the observed evidence

Test yourself

Likelihood ratios - denominator - answer

- Answer not provided in this file

Test yourself

Likelihood ratios

- Which of the following statements about likelihood ratios are true
  - Likelihood ratios provide a continuous measure of the strength of the evidence
  - Likelihood ratios are the greatest invention since sliced bread
  - The likelihood ratio is intriguing but seems very difficult to apply in practice
  - Only statisticians are interested in the likelihood ratio
  - Likelihood ratios can be very sensitive to the assumptions made

Test yourself

Likelihood ratios - answer

- Answer not provided in this file

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

Workshop Summary / Conclusions

- Analysis of forensic evidence requires some familiarity with concepts from probability and statistics
- Course reviewed basics of probability and statistics
- Discussed statistical considerations relevant to current practice (expert opinion)
- 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