Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

October / November 2022

Outline

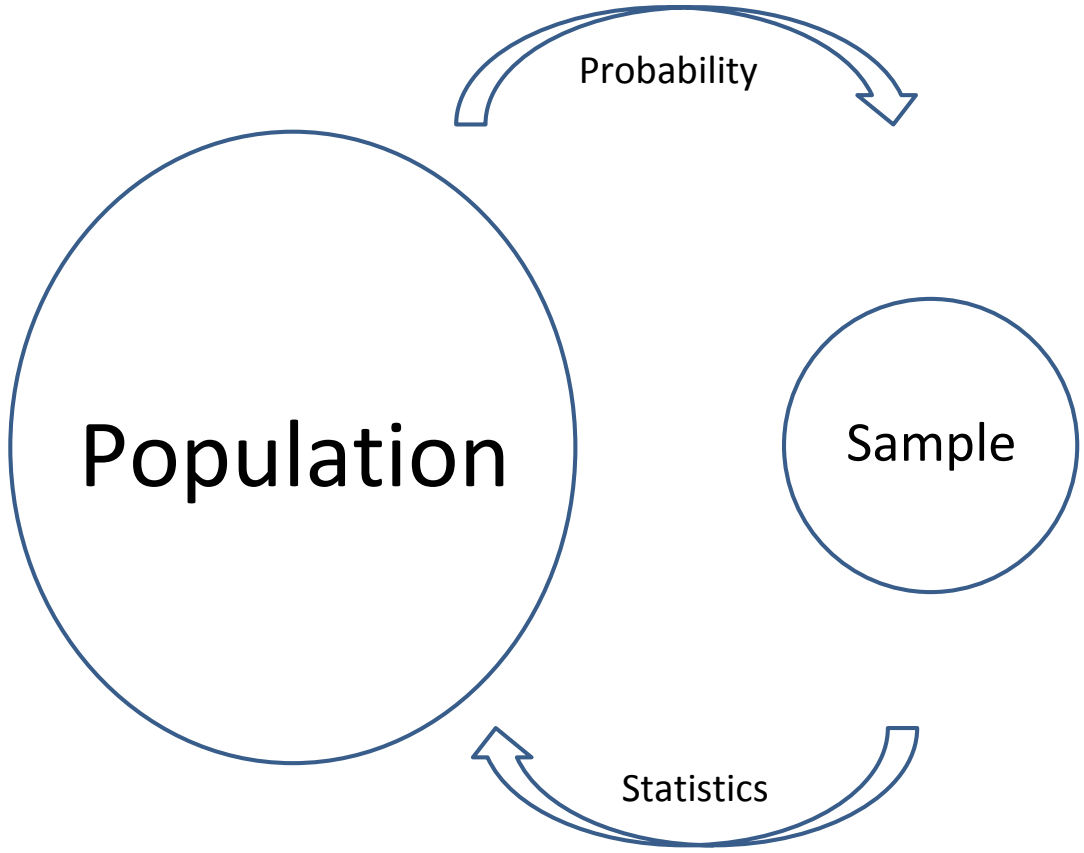
- Part 1 - Probability Concepts and Their Relevance to Forensic Science
  - review of probability concepts
  - conditional probability and independence
  - Bayes' Theorem and likelihood ratio
- **Part 2 - Data, Measurement, Reliability and Expert Opinion**
  - **collecting data**
  - **measurement, variability, reliability and accuracy**
  - **forensic evidence evaluation as expert opinion / black box studies**
- Part 3 - Statistical Inference and the Two-Stage Approach to Evidence
  - estimatiion, confidence intervals, significance tests
  - two-stage approach (significance test/coincidence probability)
- Part 4 - The Likelihood Ratio Approach - Strengths and Weaknesses
  - introducing the likelihood ratio
  - examples – the good, the bad, and the ugly

Learning Objectives for Part 2

- Understand statistical concepts associated with collecting data (sampling and study design)
- Understand the concepts of measurement uncertainty, reliability and validity
- Understand the argument for and basic principles of black box studies
- Understand the limitations of black box studies

Review of Part 1

"The Big Picture"



- Population = universe of objects of interest
  Sample = objects available for study
- Probability: population → sample (deductive)
- Statistics: sample → population (inductive)

Review of Part 1

A short recap of probability

- 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 ...
  probability of an event depends on what information is considered
- The need for careful interpretation ...
  Pr(evidence | hypothesis) vs Pr(hypothesis | evidence)
- Bayes' Rule is a mathematical result showing how we should update our probabilities
  - leads to thinking about the likelihood ratio as a summary of the evidence

Data Collection

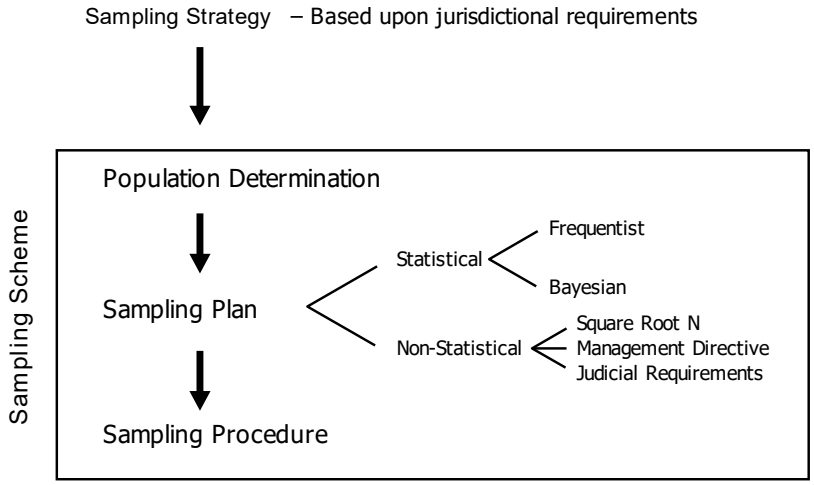
- Data are central to the analysis of forensic evidence
  - For quantitative analysis of the evidence
  - To develop and validate forensic procedures
- Assessment and validation of forensic procedures requires careful study
- Where do data come from and how do we carry out convincing studies?
- Statistics has important ideas to contribute
- Two fundamental ideas
  - sampling - getting a subset of the population of interest to study
  - experimentation - carrying out a study of a procedure/method (e.g., a black-box study)

Data Collection

- We usually collect data for purposes of drawing inferences about a population or process
- Examples
  - Learning something about a population
    (e.g., what is the frequency of a certain size or brand of shoe?)
  - Understanding variation across measurement protocols
    (e.g., calibrating a new instrument)
  - Determining causal effects
    (e.g., does a certain type of training improve performance?)
- The type, quantity and quality of data we collect determines the kind of information we can extract
- Important to plan data collection carefully

A Standard for Sampling - 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

A Standard for Sampling - 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.

Data Collection

Sampling from the population

- Sampling
  - Sampling refers to selecting a subset of the items (e.g., persons, guns, shoes) from a population of interest
  - The idea, per the earlier picture, is to use the sample to make inferences about the population
  - Why sample?
    - We sample because it is too costly or time-consuming to study the entire population
  - There are two major sampling paradigms
    - Probability sampling - items are selected according to a specified probabilistic/random approach
    - Non-random sampling - includes systematic samples, ad hoc samples, convenience samples, etc.

Data Collection

Probability-based sampling

- Starting point is the samping frame - a list of all of the items in the population
- Probability is used in choosing the sample
- Probability-based samples allow us to use the laws of probability to describe how certain we are that calculations based on our sample will reflect the population
- The simplest version of a probability sample is a simple random sample in which every set of items has the same probability of being selected
- Alternatives include stratified random samples (random samples from different categories - M/F, Age groups) and cluster samples (sampling regions and then individuals within regions)

Data Collection

Probability-based sampling

- Probability-based sampling is a very powerful concept
- Example: Literary Digest (LD) Poll of 1936
  - LD polls since 1916 had always gotten the "right" answer
  - 1936 poll of 10 million individuals (with 2.3 million respondents) predicted Landon landslide over Roosevelt (wrong by 20%!!)
  - What happened?
    - LD surveyed subscribers, auto owners, telephone users – // groups not representative of the population during the depression
    - Non-response bias – anti-Roosevelt individuals were more likely to participate
  - A big boost for the Gallup organization
    - Correctly predicted the 1936 election using probability-based sampling
    - Using probability-based sample was also able to obtain incorrect Literary Digest result with a much smaller sample

Data Collection

Non-probability-based sampling

- Useful when collecting a probability-based sample is not practical
  - No complete list of the population is available
  - Sampling individuals who do not wish to be found (e.g., undocumented individuals)
- Examples include call-in surveys, open web surveys, quota sampling
- Limitation is that it is not generally possible to make accurate statements about the population
  - Can have bias due to self-selection (individuals choose to participate)
  - Or more generally because the sample is not representative of the population
- Many famous failures with non-probabilistic sampling (e.g., Truman vs Dewey election)

Data Collection

Relevance to forensic science

- Consider a seizure of a large shipment of baggies with white powder
  - If we want to know if any baggies contain illegal drugs, then non-probability sampling may be sufficient
  - If we want to know how much illegal drugs are in the shipment, then probability-based sampling would be necessary
- Other examples where sampling comes up ....
  - How should we construct a shoe database for asessing footwear impression evdience (sample of manufactured shoes? police database?)
  - How to sample automobile windshield glass to estimate elemental concentrations?
- Regardless of approach, there are numerous issues to consider including sample size determination, non-response, biased responses

Data Collection

Experimental design

- Statistics also contributes to science through principles of study design
- Frequently we perform a study to assess performance (e.g., black box study) or understand the relationship of two or more variables (e.g., comparing measurement protocols or training programs)
- Studies can be observational or experimental (experiments involve some manipulation/intervention)
  - Observational studies gather data on a subset of the population but do not intervene. Thus we might compare two training programs by comparing the performance of graduates from the two programs.
  - Randomized controlled experiment - Participants in the study are randomly allocated to treatments (e.g., program a vs program b) and then outcomes are measured
- Randomized controlled trials are considered the gold standard for determining cause and effect. Random assignment ensures that the treatment groups are similar on all characteristics other than the assigned treatment.

Data Collection

Challenges of Causal Inference

- It is not always possible to carry out a randomized controlled experiment (e.g.,to learn about the impact of smoking)
- In such cases we may use observational studies to compare the outcomes of two groups
- Need for great care in drawing causal conclusions from observational data
- UC Berkeley admissions - a famous example
  - UC Berkeley graduate admissions in Fall 1973: Admission rates - Male=44%, Female=35%
  - Very large sample of applicants
  - These data suggest possible discrimination

Data Collection

Challenges of Causal Inference

- UC Berkeley graduate admissions in Fall 1973 (cont'd):
  - Overall admission rates - Male=44%, Female=35%; Suggests possible discrimination
  - Program-by-Program examination shows similar admit rates (A: 62% M vs 82% F; B: 63% vs 68%; C: 37% vs 34%; D: 33% vs 35%; E: 28% vs 24%; F: 6% vs 7%)
  - What happened?
    - Females applied disproportionately to programs with low admission rates (C, D, E, F); Males applied disproportionately to programs with high admission rates (A, B)
    - The aggregate analysis ignores two critical factors: (1) differences in departments where groups applied; and (2) differences in selectivity of the departments.
    - This type of difference between aggregate data and group-level data is known as Simpson's Paradox.

Data Collection

Principles of effective study design

- Study design has a large impact on the validity and relevance of results
- Key study design principles
  - compare treatments to a control (e.g., current practice)
  - randomly assign treatments to units
  - make sure sample size is large enough to draw reliable conclusions
  - make environment as realistic as possible
  - use blinding where possible 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 suggest integrating test cases with actual casework
  - a key issue in PCAST report (and the DOJ response) – more later

Test yourself

Data collection and study design

- Consider a black box study for packing tape comparisons
  - A sample of 50 volunteer forensic examiners who make packing tape comparisions are used in the black box study.
  - Each is given 10 pairs of questioned/known pairs and asked to assess whether the questioned tape came from the same roll as the known sample.
  - Researchers know ground truth for each questioned/known pair
  - Some of the examiners do not complete all 10 assigned pairs.

Test yourself

Data collection and study design

- Identify the key disadvantage of using volunteers in the study.
  - The volunteers don't get paid
  - The volunteers may not be representative of the population
  - The volunteers are likely more experienced
  - The volunteers are likely from bigger laboratories
- Some examiners do not complete all 10 assigned pairs. Identify which statement is true.
  - This is not a problem as it is a lot of extra work
  - This is a problem because the sample size is too small
  - This is not a problem we can assume they would have gotten the others correct
  - This is a problem because the pairs they skipped may differ in some way from the ones they did

Test yourself

Data collection and study design - answer

- Black box study of packing tape comparisons – A sample of 50 volunteer forensic examiners who make packing tape comparisions are used in a black box study. Each is given 10 pairs of questioned/known pairs and asked to assess whether the questioned tape came from the same roll as the known sample. Some of the examiners do not complete all 10 assigned pairs.
  - The key disadvantage of volunteers is that they may not be representative of the population. They could be more experienced or less on average ... but either is a potential problem.
  - Note though that using volunteers may be the only way to do the study.
  - Examiners not completing all the assignments is a problem. Here too, we don't have representative data. Examiners may have skipped harder ones or easier ones ... but either is a potential problem.

Measurement, Variability, Uncertainty

- Once data have been collected that are relevant to the scientific question of interest, the focus shifts to measurement and analysis
- Motivation: ASTM 2927-16 Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using LA ICP-MS for Forensic Comparisons
  - **1. Scope** - This test method covers a procedure for the quantitative elemental analysis of the following seventeen elements: lithium (Li), magnesium (Mg), aluminum (Al), potassium (K), calcium (Ca), iron (Fe), titanium (Ti), manganese (Mn), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba), lanthanum (La), cerium (Ce), neodymium (Nd), hafnium (Hf) and lead (Pb) through the use of Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS) for the forensic comparison of glass fragments. The potential of these elements to provide the best discrimination among different sources of soda-lime glasses has been published elsewhere (1-5).

Measurement, Variability, 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 (e.g., Bush vs Gore recount)
- This reflects natural variability in the measurement process, environmental factors or other contributors
  (these are often referred to as "noise")
- A measure of the resulting uncertainty should be provided to the user

Measurement, Variability, Uncertainty

Types of data

- Statisticians distinguish between different types of data
- The different types require different measurement and analysis methods
  - qualitative data
    - categorical (blood type: A,B,AB,O)
    - ordinal (grades: A, B, C, D, F)
  - quantitative data
    - discrete (consecutive matching striae)
    - continuous (refractive index of a glass fragment)
- For any type of data it is critical to understand the uncertainty associated with the observation

Measurement, Variability, Uncertainty

The role of probability

- Scientists focused on physical measurements often use uncertainty to refer to the intrinsic uncertainy 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 can often be addressed using probabilities, 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 with 95% confidence

Measurement, Variability, Uncertainty

Variability

- 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

Measurement, Variability, Uncertainty

Measures of central tendency

- With repeated measurements, we often summarize the data by presenting a measure of central tendency and a measure of variability
- Example: Suppose we want to study a new method using a subject's breath to measure their blood alcohol. We take multiple breath-based measurements from an individual whose BAC is known to be .08 or 80 mg/100ml.
  - Observations (sorted) in mg/100ml are:
    69, 74, 77, 79, 80, 80, 81, 85, 89, 120
- Common measures of central tendency are the mean and median
  - sample mean (average) = 83.4
  - sample median (middle) = 80
  - median is not effected by unusual observations

Measurement, Variability, Uncertainty

Measures of variability

- Variability associated with a set of measurements is often described with quantities like the standard deviation, interquartile range or range
  - standard deviation = "typical" deviation from the mean = 14.0 (square root of the average squared deviation from the mean)
  - range = maximum value - minimum value = 51
  - interquartile range = 75%ile of the set - 25%ile of the set = 8
- Note these are very different from each other; each conveys different information about the distribution
- We will see that the standard deviation is useful for learning about the population mean

Measurement, Variability, Uncertainty

Reliability

- Variability is related to the concept of reliability. Reliability plays a large role in ongoing discussions about forensic science
- Reliability refers to the consistency of a measurement or a measurement protocol, i.e., will we get the same answer if a process is repeated
- Variability and reliability are related concepts but not the same. Consider using a scale to determine the weight of an object.
  - A scale with low variability (small standard deviation) gives reliable measurements
  - An imprecise scale (large variability) provides measurements with considerable uncertainty (low reliability)
  - But we may be able to get a reliable esimate of the weight by averaging a number of readings from the scale
  - So a single scale reading is not a reliable measurement approach but the average of a number of readings is a reliable measurement approach

Measurement, Variability, Uncertainty
Reliability

- There are several aspects of reliability
  - **repeatability** refers to whether a measurement or decision would be the same in two instances using the same item and the same examiner. It is an intra-examiner assessment.
  - **reproducibility** refers to whether a measurement or decision would be the same in two instances using the same item and different examiners. It is an inter-examiner assessment.

Measurement, Variability, Uncertainty

Reliability

- Questions about reliability are central to thinking about forensic evidence
- Example: It is believed that signature complexity is relevant to the assessment of signature evidence
  - More complex signatures may allow an examiner to have more confidence in their conclusion
- But .... Before we can verify that belief we need to know how reliably can we measure complexity!

Measurement, Variability, Uncertainty

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)

ID	FDE1	FDE2	FDE3	FDE4	FDE5
001	4	4	5	3	4
002	4	5	5	4	5
003	3	4	4	4	3
004	4	4	5	4	4
005	2	2	2	3	3
...	...	...	...	...	...

  - 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

Measurement, Variability, Uncertainty
Validity

- There are many meanings of validity
- We are interested in validity in the sense of accuracy
- A measurement or decision is **valid** if it matches a known truth
- Note that validity is also used for describing psychological measures (e.g., depression scales)
  - convergent validity - it correlates with other measures of the same concept
  - discriminant validity - it doesn't correlate with measures of other concepts
- Note that validity is also used in evaluating study designs
  - internal validity - the study is well designed and leads to appropriate conclusions for study population
  - external validity - the study can be expected to generalize to other populations

Measurement, Variability, Uncertainty

Validity and Reliability

- Validity is different than reliability
- High reliability is required for us to have a valid/accurate procedure
  - If the measurement is valid, then repeated measurements will give the same (correct) answer; this means they must agree with each other (hence reliable)
- But high reliability itself does not guarantee a valid/accurate procedure
  - Why not?
  - A scale can be very reliable in that it always gives the same reading for a person .... but it may give incorrect weights
  - All forensic examiners may agree in their decisions .... but be wrong

Test yourself

Reliability

- For each item, indicate whether the statement is true or false.
  - Repeatability and reproducibility are both components of reliability
  - Repeatability is a between-examiner assessment
  - High repeatability and high reproducibility guarantee high accuracy
  - A highly accurate forensic discipline will also be found to have high reproducibility

Test yourself

Reliability - answer

- For each item, indicate whether the statement is true or false.
  - TRUE - Repeatability and reproducibility are both components of reliability
  - FALSE - Repeatability refers to within-examiner assessments. Reproducibility refers to between-examiner assessments.
  - FALSE - High repeatability and high reproducibility only guarantee consistent measurements. They may or may not be correct.
  - TRUE - A highly accurate forensic discipline will also be found to have high reproducibility

Applying our Statistical Concepts to Forensic Science

- Goal of the short course is to connect statistical concepts to forensic science
- For today's topics (study design, measurement) it is natural to talk about black box studies
- Before doing so we set a context that will apply for the remainder of the short course

The Forensic Examination

- There are a range of questions that arise in forensic examinations - source conclusions, timing of events, substance ID, cause/effect
- Focus for the course is on source conclusions
  - 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 the same 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

- A wide range of 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
- We focus in the short course 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
- Today we focus on forensic conclusions as expert opinion
- Part 3 will talk about statistical inference and the two-stage approach
- Part 4 will talk about the likelihood ratio approach

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
- For example, handwriting examiners may use statements like: 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, as per FRE 702, to establish that testimony is
  - "based on sufficient facts or data"
  - "the product of reliable principles and methods"
- Our discussion of reliability and validity is relevant here
  - 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)?
  - Do analysts get the right answer in studies where ground truth is available (accuracy / validity)?
- These three questions led PCAST to suggest "black box" studies

Forensic Conclusions as Expert Opinions

PCAST Report

- PCAST-style "black box" studies of performance can be used to assess reliability and validity of a field
  - 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 identified several requirements for studies
  - Study should include a large number of examinations/examiners
  - Examiners in study should be representative of the population
  - Samples with known ground truth
  - Samples should be representative of casework
  - Study overseen by independent party
  - Study design and results should be peer reviewed
  - Data, samples, results made publicly available
  - Need more than one study for a discipline

Forensic Conclusions as Expert Opinions
PCAST Report

- PCAST report is controversial :)
- DOJ statement (January 2021) argues that "black box" studies recommended by PCAST are not required
- Specifically argued that not all of the PCAST requirements are needed for good scientific studies
- Others argue that "black box" error rate estimates are of limited use in a particular case
- My view
  - PCAST requirements are clearly desirable though perhaps not always attainable
  - "Black box" evaluations provide extremely useful information in understanding how to interpret forensic evidence

Forensic Conclusions as Expert Opinions

Latent print black box study

- Ulery et al. (2011) fingerprint study
  - 169 examiners
  - 744 pairs (latent, known)
  - known truth (mates, non-mates) for each pair
  - each examiner assessed 100 pairs
- Accuracy results

	Total	Exclusions	Inconclusives	Individualizations
Mated Pairs	5969	450	1856	3663
NonMated Pairs	4083	3622	455	6

  - False negative error rate = false exclusion rate = $450/5969 = 7.5\%$
  - False positive error rate = false ID rate = $6 \ / \ 4083 = 0.15\%$
  - Note that inconclusives are essentially treated as correct here. More on this below

Forensic Conclusions as Expert Opinions

Latent print black box study

- A note about the false positive (false ID) error rate
  - false positive rate is Pr(examiner makes identification | non-mates)
  - There is another 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
    - Suppose an examiner only sees mated pairs and correctly reports IDs every time
    - For this person 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

Firearms black box study

- Baldwin et al. (2014) cartridge case study accuracy results

	Total	Eliminations	Inconclusives	Individualizations
Mated Pairs	1090	4	11	1075
NonMated Pairs	2178	1421	735	22

  - False negative error rate = false exclusion rate = $4/1090 = 0.36\%$
  - False positive error rate = false ID rate = 22 / 2178 = 1.0%
  - Inconclusives:
    - An especially big issue for nonmated pairs with 33.7% inconclusive
    - Above false ID rate treats inconclusives as correct. Not quite right
    - Can look at false positive error rate out of decisions = $22/1443 = 1.5\%$ (this ignores inconclusives)
    - Two summaries? 33.7% inconclusive; 1.5% false IDs among decisions
    - Some are arguing that inconclusives are errors which would make error rate $757/2178 = 34\%$. Again, not quite right
    - Some inconclusives may be "correct", others may be "errors". How do we decide?
    - Much current controversy about inconclusives in firearms !!

Forensic Conclusions as Expert Opinions

Back to latent prints

- Ulery et al. also studied reliability (2012 paper)
- Repeatability (**same examiner**, 7 months apart)

	First Decision	Total	Second Decision Exclus	Inconcl	Individ
Mated Pairs	Exclusion	226	30%	49%	21%
	Inconclusive	527	3%	91%	6%
	Individualization	265	3%	8%	89%
NonMated Pairs	Exclusion	470	91%	9%	0%
	Inconclusive	175	27%	73%	0%
	Individualization	3	67%	33%	0%

Forensic Conclusions as Expert Opinions

Back to latent prints

- Ulery et al. also studied reliability (2012 paper)
- Reproducibility (**different examiners**)

	First Examiner	Total	Second Examiner Exclus	Inconcl	Individ
Mated Pairs	Exclusion	3,194	17%	57%	26%
	Inconclusive	32,224	6%	86%	8%
	Individualization	15,962	5%	16%	79%
NonMated Pairs	Exclusion	13,735	87%	13%	0.2%
	Inconclusive	5,263	34%	66%	0.06%
	Individualization	28	89%	11%	0%

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
  - "Black box" studies are helpful (my opinion) but have limitations
    - They may not represent practice
      (people know they are being studied, there is no validation step)
    - They 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

- If a black box study finds high validity (accuracy) then it must also show high reliability
- TRUE - If all examiners get the correct answers, then they must all agree (high reliability)
- If a black box study finds high reliability then it must also show high validity
- FALSE - All examiners could agree (high reliability) but they may not be right very often
- A false identification error occurs when a known set of non-matching items are mistakenly identified as coming from the same source
- TRUE - This is the definition; it focuses on the subpopulation of non-matching items that are analyzed

(cont'd)

Test yourself

Forensic conclusions as expert opinion - answer

- 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
- FALSE - This is a bit confusing. These are false identifications but this approach doesn't provide a reliable estimate of the false identification rate. This approach depends on the mix of cases that examiners see. If investigators do a great job and only present evidence of guilty suspects, then we will see very few false identifications. That does not however prove that examiners would make few false identifications if they were given more opportunities. Important to think about.

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

- In carrying out a black box study, there are always questions about how to treat inconclusives. Which of the following approaches do you support?
- My own view (**in parentheses**):
  - Inconclusives are always errors because the pair must either be a same source pair or a different source pair (**This does not appeal to me. But some statisticians support.**)
  - Inconclusives should always be marked as correct because they are not errors (**This definitely seems like a bad idea to me.**)
  - Inconclusives should be omitted from the error rate calculation (**This is a reasonable approach.**)
  - We need more than one summary of the study to accurately convey examiner performance (**I believe this is the right way to think about it.**)
  - A well-designed study should include some cases for which expert consensus would indicate that inconclusive is the right answer (**A good idea. But likely challenging to do.**)

Data, Measurement, Reliability and Expert Opinion

A short recap

- Random samples allow for generalization to the population
- Controlled experiments are best for cause/effect conclusions
- Understanding uncertainty of measurements / decisions is crucial (e.g., ISO standard)
  - reliability refers to the consistency of measurements / decisions
  - validity refers to the accuracy of measurements / decisions
- Black box studies provide useful "discipline"-wide metrics regarding the use of expert opinion to summarize evidence
  - Challenging to execute these studies (but recent publications in handwriting, bloodstain pattern, footwear)
  - Need better approaches to dealing with "inconclusives"
  - Do black box studies reflect actual practice? Can we embed occasional black box examples in casework?
  - Don't explicitly address what happens in an individual case