An Introduction to Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

Center for Statistics and Applications in Forensic Evidence

April 2017

OUTLINE

- Part I - Statistical Preliminaries - pg. 3
  - review of probability concepts
  - review of statistical inference concepts
- Part II - Statistics for Forensic Science - pg. 65
  - significance test/coincidence probability
  - likelihood ratio
  - forensic examination as expert opinion

Statistical Preliminaries

Outline and Overview

- Outline
  - probability
  - statistical inference
- Overview - "The Big Picture"
  - Population = universe of objects of interest
    Sample = objects available for study
  - Probability: population $\rightarrow$ sample (deductive)
  - Statistics: sample $\rightarrow$ population (inductive)
  - Often use both together
    1. build/assume model for population
    2. assess sample using model
    3. refine model; go back to step 2

Statistical Preliminaries
The Big Picture in Practice

- Applications
  - Drug seizure (population = 100 bags; sample chosen for analysis)
  - Glass fragments (two populations = glass from scene and glass from suspect; take samples from each)
  - Forensic accounting
    (population = all transactions; sample chosen for analysis)
- Relevance to pattern evidence?
  - less obvious but we will return to this point

Probability

Basic setup

- Experiment with uncertain outcomes
- Sample space: list of all possible outcomes
- Event: set of outcomes of interest
- Example 1 - Footwear
  - measure size of shoe
  - sample space $= (6, 6.5, 7, 7.5, 8, 8.5, ...)$
  - event = shoe of size 9
- Example 2 - Footwear
  - record type of shoe
  - sample space = (Nike, Vans, ...)
  - event = sneaker
- Example 3 - Firearms
  - consecutive matching striae for pair of bullets
  - sample space = 0, 1, 2, ...
  - event = 10 or more matching striae

Probability
Interpretation

- Probability of an event is a number (between 0 and 1) assigned to the event describing the likelihood it occurs
- Interpretations
  - long run frequency of occurrence of event (must be a repeatable experiment)
  - subjective belief of likelihood of an event (e.g., a suspect was present at a crime scene)

Probability

Axioms/basic laws of probability

- Use $P(Y)$ or $Pr(Y)$ to denote the probability of an event Y
- Probabilities are between 0 and 1
  - P(Y)=0, event never happens
  - P(Y)=1, event always happens
- Sum of probabilities of all possible outcomes is 1
- Define complement of an event Y as the event that Y does not happen, denoted as $\bar{Y}$ (or sometimes "not Y")
  - $P(Y) + P(\bar{Y}) = 1$

Probability

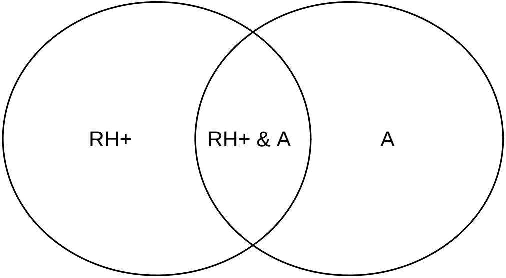
Probability and Odds

- Probabilities are related to odds
  - odds in favor of event R are defined as $O = P(R)/P(\bar{R}) = P(R)/(1 - P(R))$
  - odds against event R are defined as $O = P(\bar{R})/P(R) = (1 - P(R))/P(R)$
  - odds against and probabilities are related since $P(R) = 1/(O + 1)$
  - if $P(R) = .2$, then odds against are $.8/.2 = 4$ (or "4 to 1")
  - if $P(R) = .8$, then odds against are $.2/.8 = .25$ (or "1 to 4")

Probability

Adding probabilities

- Consider two events Y and Z
  - If Y and Z can't happen together then we get a simple addition formula $P(Y \text{or} Z) = P(Y) + P(Z)$
  - In general a bit more complicated because of overlap
- Examples
  - P(type A or type B blood) $=$ P(type A) $+$ P(type B)
  - P(RH$+$ or A blood) $=$ P(RH$+$) $+$ P(A) $-$ P(RH$+$ and A)



Probability
Conditional probability

- Consider an event Y which has probability $P(Y)$
- Suppose we learn that another event Z has occurred
- This may change our opinion about the likelihood of Y occurring
- Key concept: The probability of an event often depends on other information
- This leads to the definition of conditional probability
  - We write $P(Y|Z)$ for the conditional probability that Y occurs given that we know Z has occurred
  - $P(Y|Z) = P(Y \text{ and } Z)/P(Z)$
- Note this gives a multiplication rule for probabilities, $P(Y \text{ and } Z) = P(Z) \times P(Y|Z)$
- Can be helpful to think of all probabilities as being conditional ... it is just a question of what information is assumed known

Probability

Conditional probability - example

- Study of sentencing of black convicted murderers in Georgia (1980s)
- Murderers categorized by race of victim and sentence received in the table below

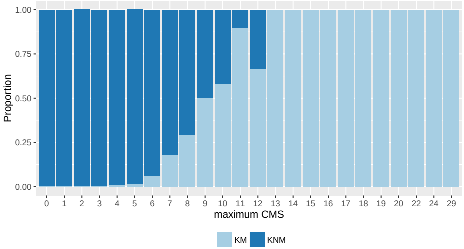
	Death Penalty	No DP	Total
White victime	45	85	130
Black victime	14	218	232
Total	59	303	362

- P(Death Penalty) $= 59/362 = .16$
- P(Death Penalty | White Victim) $= 45/130 = .35$
- P(Death Penalty | Black Victim) $= 14/232 = .06$
- A number of important factors are not included (e.g., context of murder)

Probability

Conditional probability - example

- What is the probability of seeing two known matching bullets (KM) and two known non-matching bullets (KNM) given the maximum number of consecutive matching striae (CMS)? Here we condition on the maximum number of CMS.



- For low CMS values it is much more likely that we have a non-matching pair

Probability

Independent vs. Dependent events

- Definition - if the likelihood of one event is not affected by knowing whether a second has occurred, then the two events are said to be independent (e.g., living in MW and driving a red car)
- Earlier example shows death penalty is not independent of race of victim
- Forensics example: DNA analyses typically use loci on different chromosomes so they will be independent

Probability

Multiplying probabilities

- It follows from the definition of conditional probability that

$$P(A \text{ and } B) = P(B)P(A|B) = P(A)P(B|A)$$

- If A and B are independent, then we get a simple formula $P(A \text{ and } B) = P(A)P(B)$
- Example:
  - P(left handed and from Florida) = P(left handed) P(from Florida)

Probability

Multiplying probabilities ...a cautionary tale

- People vs Collins (California, 1968)
  - Robbery trial
  - Eyewitness describes robbers as "black male with a beard and moustache, and a white female with a blonde ponytail, fleeing in a yellow car"
  - Prosecution provided estimated odds of each characteristic (black man with beard $= 1/10$, black man with moustache $= 1/4$, white woman with pony tail$= 1/10$, white woman with blonde hair $= 1/3$, yellow car$= 1/10$, interracial couple in car $= 1/1000$)
  - Mathematics expert talked about the multiplication rule for probability (yields 1/ 12 million)
  - Conviction was set aside and the statistical reasoning criticized for ignoring dependence among the characteristics

Probability

Probabilities in court

- Let S be event that suspect was present at the scene of a crime and $\bar{S}$ be event that the suspect was not present
- Assume each juror has initial probability for each event
- Witness says saw tall Caucasian male running from scene, defendant is tall Caucasian male
  - jurors update probabilities
- Window broken during the crime and fragments found on the defendant's clothing match the broken window
  - jurors update probabilities
- How should jurors update their probabilities?
- Do jurors actually think this way?

Probability

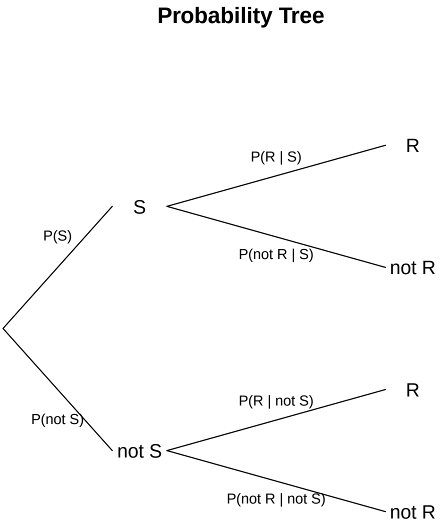
Bayes' Rule (or Theorem)

- Bayes' rule provides an updating formula
  for probabilities that can sometimes be used
- Suppose we have initial estimate for probability of event S
- We learn that event R has occurred and want to update our probability
- Need to know something about the relationship of R and S
- If so, can use Bayes' Rule

$$
\begin{aligned}
P(S|R) &= \frac{P(R \text{ and } S)}{P(R)} = \frac{P(R|S)P(S)}{P(R)} \\
&= \frac{P(R|S)P(S)}{P(R|S)P(S) + P(R|\bar{S})P(\bar{S})}
\end{aligned}
$$

Probability

Bayes' Rule (or Theorem)



Probability

Bayes' Rule (or Theorem)

- Example - diagnostic tests for gunshot residue
  - let G denote the presence of gunshot residue and T denote a positive test
  - often know from validation of the test quantities like $P(T|G)$ (sensitivity) and $P(\bar{T}|\bar{G})$ (specificity)
  - start with some idea of $P(G)$ (prevalence)
  - may want $P(G|T)$
  - use Bayes rule $P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G)+P(T|\bar{G})P(\bar{G})}$
  - key point is that, in general, $P(T|G) \neq P(G|T)$

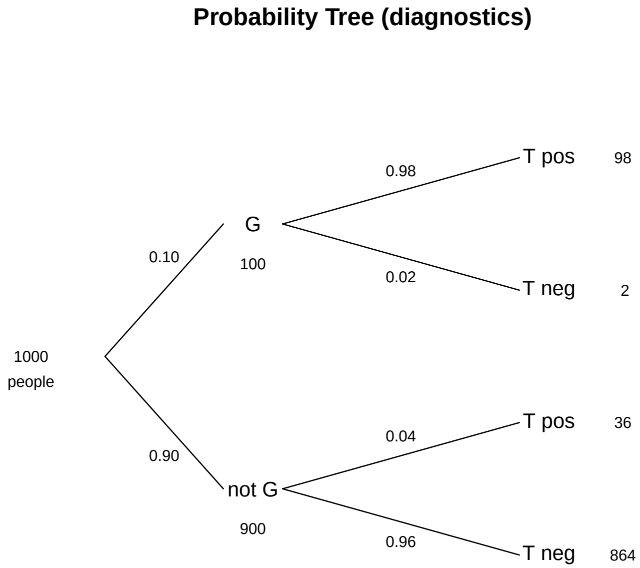
Probability

Bayes' Rule (or Theorem)

- Can sometimes get surprising results
- Recall diagnostic test setup
  - assume $P(T|G) = .98$ (sensitivity)
  - assume $P(\bar{T}|\bar{G}) = .96$ (specificity)
  - assume $P(G) = .9$ (prevalence)
    (i.e., testing in a population where gun usage is common)
    - $P(G|T) = \frac{.98 * .9}{.98 * .9 + .04 * .1} = .995$
  - now assume $P(G) = .1$ (low prevalence)
    (i.e., testing in a population where gun usage is rare)
    - $P(G|T) = \frac{.98 * .1}{.98 * .1 + .04 * .9} = .73$
    - even if test is positive still not very sure about whether residue is present
    - why does this happen?

Probability

Bayes' Rule (or Theorem)



Probability

Bayes' Rule to the likelihood ratio

- In the general forensic setting let S denote "same source" and E denote "evidence"
- Bayes' rule: $P(S|E) = P(E|S)P(S)/P(E)$
- Bayes rule can be rewritten in terms of odds

$$\frac{P(S|E)}{P(\bar{S}|E)} = \frac{P(E|S)}{P(E|\bar{S})}\frac{P(S)}{P(\bar{S})}$$

  - left hand side: odds in favor of S given the evidence (E)
  - last term on right hand side: odds in favor of S before evidence (E)
  - first term on right hand side is known as the likelihood ratio
  - ratio of likelihood of evidence under S to likelihood of evidence under $\bar{S}$
- Likelihood ratio is the factor by which multiply prior odds of same source to get posterior odds of same source

Probability

Bayes' Rule to the likelihood ratio

- Example - gunshot residue (E = evidence = positive test; S = suspect has gunshot residue)
  - $LR = P(E|S)/P(E|\bar{S}) = .98/.04 = 24.5$
  - In high prevalance case, prior odds are 9 : 1 and posterior odds are 220.5 : 1 (posterior probability = .995)
  - In low prevalance case, prior odds are 1 : 9 and posterior odds are 24.5 : 9 (posterior probability = .73)
- Bayes' rule is a mathematical result showing how we should update our probabilities
  - It is not at all clear that this is how people operate in practice

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 also 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 for a family like the Clarks was $1/8500$, so the probability of two SIDS deaths was $1/73$ million ($1/8500^2$)
- Several problems with this argument ... 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 also 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 for a family like the Clarks 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
  - Is 1/8500 correct for "families like the Clarks"
  - The use of multiplication assumes independence for the two children
  - Also need to consider alterative hypotheses (and their likelihood)

Probability to Statistical Inference
Collecting data

- Where do data come from
  - experiment - investigator designs a study
  - sample - study a subset of the population of interest
- Statistics can tell us a great deal about how to design an experiment or choose a sample

Probability to Statistical Inference

Collecting data

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

Probability to Statistical Inference

Collecting data

- 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 our sample answer reflects the population
  - many famous failures (e.g., Truman vs Dewey election)

Probability to Statistical Inference
Collecting data

- Types of data
  - qualitative
    - categorical (blood type: A,B,AB,0)
    - ordinal (eval of teacher: poor, avg, exc)
  - quantitative
    - discrete (consecutive matching striae)
    - continuous (refractive index of a glass fragment)

- 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
- Examples of distributions
  - Binomial: # of success in n trials
  - Poisson: count # of events
  - normal: bell-shaped curve
  - log normal: logarithm of observations follow a normal distribution

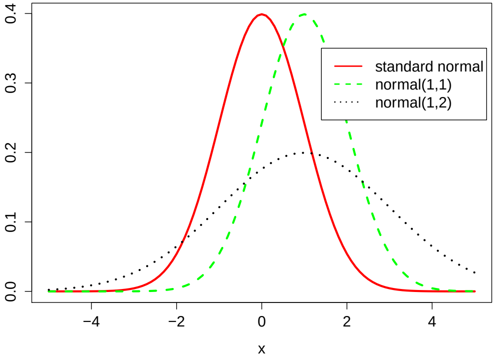
Probability

Probability distributions - normal

- Familiar bell-shaped curve
- Measurement error is often assumed to follow a normal distribution
- Described by two parameters, mean μ and standard deviation σ
- We write $X \sim N(\mu, \sigma)$
  (for measurement error we often assume the mean is zero)
- 95% of values within 2 σ of the mean and over 99% within 3 σ
- For $X \sim N(\mu, \sigma)$, $Z = (X - \mu)/\sigma \sim N(0, 1)$ is a standard normal and tables exist to compute probabilities for Z

Probability

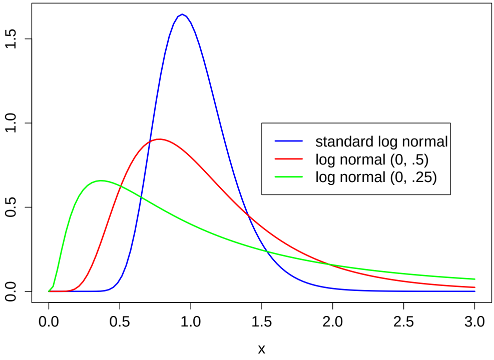
Probability distributions - normal



- Often act as if everything is normally distributed
- Of course this is not true
- Can't be exactly true for a quantity that is certain to be nonnegative (trace element concentration)
- In that case, may believe the logarithm of the quantity is normal, this gives a lognormal distribution for the quantity
- Specify lognormal distribution with two parameters, mean (on log scale) μ and standard deviation (on log scale) σ

Probability

Probability distributions - lognormal



Probability

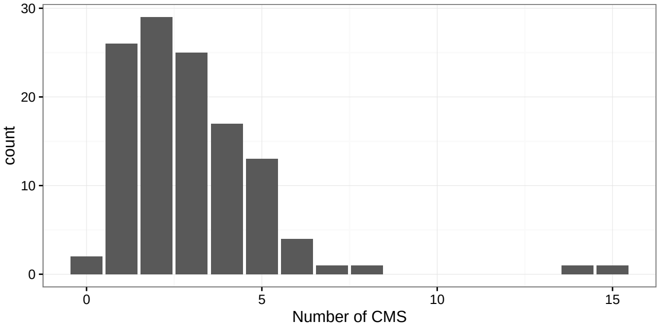
Probability distributions - discrete

- Some quantities take on very few values (discrete data)
- There are two popular discrete distributions
  - binomial
    - binary data (two categories)
    - n independent 'trials'
    - P(success)=p on each trial
    - expected number of successes $= np$
    - example: # correct answers in a cheating case
  - Poisson
    - count data (# events in a fixed time)
    - this distribution has mean = variance
    - variability increases with the mean
    - example: number of calls to 911
      between 10:00 and midnight on Friday nights

Probability

Probability distributions - discrete

- Distribution of the maximum number of CMS for a randomly selected bullet compared to 118 known lands approximately follows a Poisson model



Statistical Inference
Background

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

Statistical Inference
Background

Recall - “The Big Picture”

- Population = universe of objects of interest
  Sample = objects available for study
- Probability: population $\rightarrow$ sample (deductive)
- Statistics: sample $\rightarrow$ population (inductive)
- Often use both together
  1. build/assume model for population
  2. assess sample using model
  3. refine model; go back to step 2

Statistical Inference

Example

What if we are interested in the average height of adult U.S. population?

- Population = all U.S. adults
  Sample = everyone in this room
- We take the average height of everyone here and use this statistic to make inference about the population mean for all U.S. adults
- This assumes of course that our sample is a random sample from the population. This assumption may be questionable.

Statistical Inference
Background

- Goal: inference about a parameter
- Possible parameters
  - mean
  - variance
  - proportion
- Different kinds of inferential statements
  - single estimate of parameter (point estimate)
  - range of plausible values for parameter (interval estimate)
  - perform test of a specific hypothesis about parameter

Statistical Inference
Background

- Summarize qualitative data by recording frequency of various outcomes
- Summaries of quantitative data
  - mean or average
  - median = middle value
  - standard deviation (s.d.) = measure of spread
  - percentiles
- Example: if data $= (19, 20, 21, 22, 23)$, then mean = median = 21, s.d. = 1.58
- Example: if data $= (19, 20, 21, 22, 93)$, then mean = 35, median = 21, s.d. = 32.4

Statistical Inference

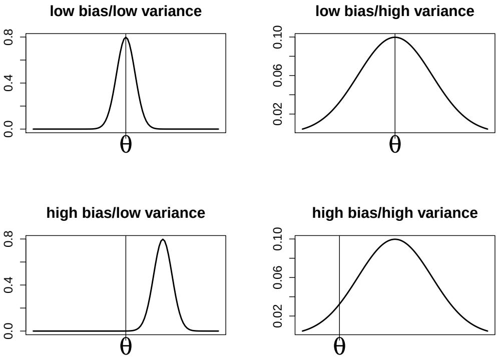
Point estimation

- 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
- For population mean, might use sample mean as an estimator
  - no bias
  - low variability if sample is large

Statistical Inference

Performance of diifferent estimators for unknown θ

Figures below show what would happen in many repeated attempts



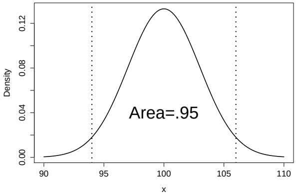
Statistical Inference
Standard errors

- A limitation of point estimators is that they don't provide any indication of accuracy
- The standard error of an estimator describes how far from the correct population answer the estimate is likely to be
- Recall that standard deviation is a measure of the spread (variability) in a sample or in a population
- When we look at a summary statistic (mean, median, extreme value) it is also a random quantity (would give diff't value in diff't samples)
- Standard error of an estimate is how we measure its variability

Statistical Inference

Standard errors

- Consider a population with mean 100 and s.d. 15
  - review meaning of s.d.: expect 95% of observations to be between 70 and 130
  - now suppose we compute mean of 25 responses
  - standard error is $15/\sqrt{25} = 3$
  - sample mean should be around 100
  - 95% of the time it will be between 94 and 106

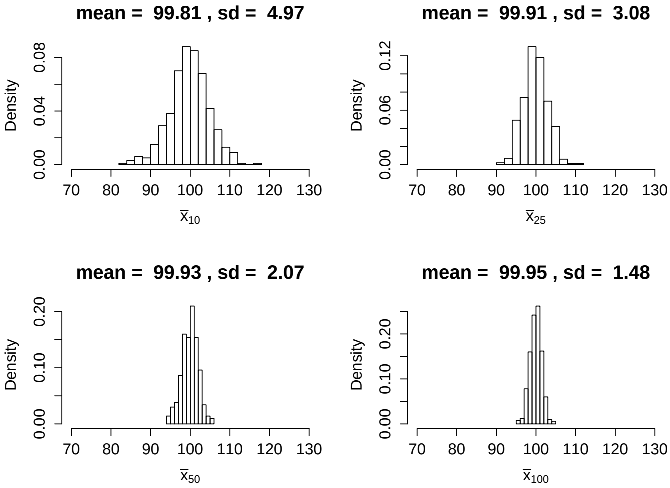


Statistical Inference

Standard errors and sample size

- Again consider a population with mean 100 and standard deviation 15
- The size of a sample plays a critical role in determining how accurate we can be
- On the same size graph we show below: the distributions of the means computed from samples of size 10, 25, and 50
- Note that distributions of means are tighter around the expected value as the sample size increases
- The standard error decreases in proportion to the square root of the sample size

Statistical Inference
Standard errors and sample size



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 a 95% confidence interval for a quantity by taking point estimate $\pm$ 2 std errors

Statistical Inference

Interval estimation

- Example: 10 glass fragments from crime scene
- measure concentration of aluminum
- mean = 0.730, standard devation = 0.04
- standard error = 0.040 / sqrt(10) = 0.013
- approximate 95% confidence interval for the mean aluminum concentration in the crime scene windo is 0.73 +/- 2*0.013 = (.704,.756)
- interpretation

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, usually means the status quo. We look for evidence against the null
- There is an alternative (or research) hypothesis that helps us to design the test
- If we reject the null hypothesis then we say we have a statistically significant result

Statistical Inference

Hypothesis testing

- Two types of errors
  - type I: reject the null hypothesis when it is true
  - type II: fail to reject the null when it is false
- Type I error often considered more serious: we only want to reject the null if strong evidence against it
- For juries: null hypoth=innocent, alternative=guilty
  - type I error is to say guilty when innocent
  - type II error is to say innocent when guilty

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
  (point estimate - null hypothesis value)/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

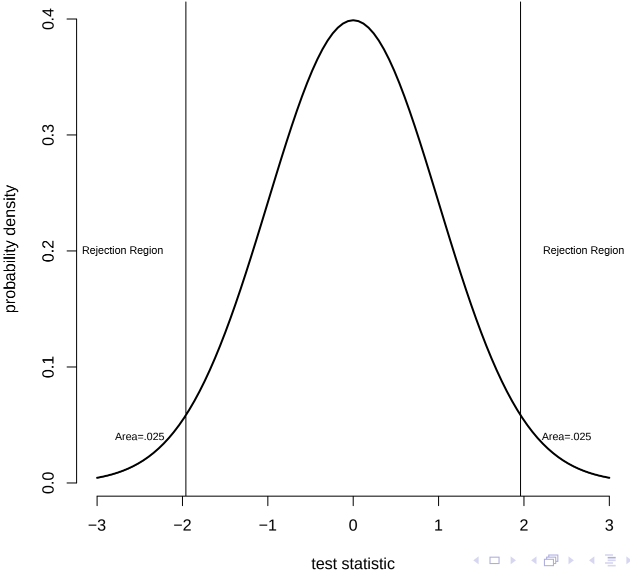
- Summarize test by attaching a probability to the test statistic
- Definition: a *p*-value gives the probability that we would get data like that we have in the sample (or something even more extreme) given that the null hypothesis is true
- Small *p*-values mean unusual data that lead us to question 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

Procedures for normal data

- Most well established procedures are those related to drawing conclusions about the mean of a normal population
- Suppose that we have acquired a random sample of n observations from the population
- Natural point estimate of population mean is the sample mean $\bar{X} = \sum X_i/n$
- 95% confidence interval is: $\bar{x} \pm 1.96SE(\bar{x})$
  - $SE(\bar{X})$ is the standard error of the sample mean
  - $SE(\bar{X}) = SD(\text{population})/\sqrt{n}$
- Can test a hypothesis about μ say $H_0 : \mu = \mu_0$ using the test statistic, where p-value is obtained from a t table
- Key result is that these procedures work well even if population is not normally distributed as long as the sample size is large

Statistical Inference
Procedures for normal data



Statistical Inference

Procedures for normal data - example

- Want to estimate the mean amount of a trace element for population= all bullets in Southern Iowa
- Get a random sample of 500 bullets from the area
    - sample mean is 55, standard deviation is 22
    - standard error of the mean is $22/\sqrt{500} = 0.98$
- 95% confidence interval: $55 \pm 1.96 * .98 = (53.1, 56.9)$
- Suppose we have reason to believe that Remington (mean=58) is main producer in this area, can check with hypothesis test
    - hypothesis test: $H_0 : \mu = 58$
    - test statistic, $t = (55 - 58)/.98 = 3.06$
    - estimate is 3 standard errors from the mean under the null hypothesis
    - p-value = .00135: if null hypothesis is true, then observe a value 3 standard errors or more .1% of the time, hence reject null hypothesis

Statistical Inference

Procedures for a proportion - example

- The same approach works in cases where we are not focused on the mean of a normal population
- Another common situation is that we are interested in testing the proportion of a population having some characteristic
- Example:
  - Suppose nationally, 35% of all athletic shoes are Nike ($p = .35$).
  - We take a random sample of 100 individuals in LA, and 42 are wearing Nike's ($\hat{p} = .42$).
  - Can we still use the national proportion as a reference for LA or is LA different?

Statistical Inference

Procedures for a proportion - example

- Suppose nationally, 35% of all athletic shoes are Nike ($p = .35$).
- We take a random sample of 100 individuals in LA, and 42 are wearing Nike's ($\hat{p} = .42$).
- Can we still use the national proportion as a reference for LA or is LA different?
- The central limit theorem is a powerful theorem that ensures sample proportions behave similar to normal sample means, $\hat{p}$ has a normal distribution with mean p and standard error $\sqrt{p(1-p)/n}$
- Test:

$$z = \frac{(\hat{p} - p)}{\sqrt{\frac{p(1-p)}{n}}} = \frac{(.42 - .35)}{\sqrt{\frac{.35(1-.35)}{100}}} = 1.47$$

- p-value=.0705. Therefore, we fail to reject the null hypothesis (LA is not significantly different)

Statistical Inference

Comparing two means

- Previous discussion addresses methods for one sample
- In practice most 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 μ_{scene} to be mean trace element level for the "population" of glass on the scene
    - define $\mu_{suspect}$ to be the mean 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

Statistical Inference

Comparing two means - example

- Take null hypothesis to be that evidence agrees, $\mu_{scene} = \mu_{suspect}$
- Take alternative hypothesis to be that the evidence does not agree, $\mu_{scene} \neq \mu_{suspect}$
- Suppose 10 glass fragments are taken from glass at the scene (Y) and 9 fragments are found on the suspect (X).
  - $\bar{X} = 5.3, s.d. = 0.9, SE(\bar{X}) = 0.9/\sqrt{10} = .28$
  - $\bar{Y} = 5.9, s.d. = 0.85, SE(\bar{X}) = 0.85/\sqrt{9} = .28$
  - 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
- Interpretation is a key issue (can't reject the hypothesis of common source)

Statistical Inference
Equivalence Testing

- Recall that the hypotheis test assumes the null hypothesis is true until proven otherwise
- There is an alternative approach to the previous test
- Equivalence testing assumes the population means are different (this becomes the null hypothesis)
- Requires us to specify a "practically" important difference Δ

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

- This requires we test two different hypotheses: (1) the means differ by more than Δ vs the alternative that they don't; (2) the means differ by less than $-\Delta$ vs the alternative that they don't
- We reject this null hypothesis (the means are different) and conclude the samples are equivalent ONLY if we get a small p-value for both hypothesis tests

Statistical Inference

Equivalence testing - numerical example

- Return to our example of comparing two glass samples
- Previous research suggests a difference of 1 is considered significant
- We first test whether there is evidence that $\mu_y - \mu_x > 1$ vs the alternative that $\mu_y - \mu_x < 1$
  - Observed difference is 0.6, standard error is 0.4. Observed difference is one standard error away from the hypothesis so p-value $= .32$ (can't reject the possibility that the true difference is bigger than 1 even though our observed difference is less than 1)
- We next test whether there is evidence that $\mu_y - \mu_x < -1$ vs the alternative that $\mu_y - \mu_x > -1$ (no details provided)
- Because we need to reject both tests to reject the null and didn't reject in the test above, we fail to reject the null hypothesis that the mean trace levels are different between the scene and suspect

Statistical Inference

Hypothesis testing - discussion

- Hypothesis testing does not treat the two hypotheses symmetrically (null is given priority)
- P-values depend heavily on the sample size
  - If you have the same means and standard devations 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
  - Failing to reject the null hypothesis does not mean that the null hypothesis is true

Statistical Preliminaries

- Reviewed basics of probability
  - probability is the language of uncertainty
  - important to understand what is being assumed when talking about probability (probability of having disease given a positive test is different than the probability of having a positive test given the disease)
  - probability distributions describe the variability in a population or in a series of measurements
- 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
Part II - Outline

- Brief review of probability/statistics
- The forensic examination
- Significance test/coincidence probability approach
- Likelihood ratio approach
- Forensic conclusions as expert opinion

- Reference: *Statistics and the Evaluation of Evidence for Forensic Scientists*, C. G. G. Aitken

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
- 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., race of criminal, evidence substrate)
- Two hypotheses
  - S - items from crime scene and suspect have common source (or suspect is source of crime scene item)
  - $\bar{S}$ - no common source
- Goal: assessment of evidence
  - do items appear to have a common source
  - how unusual is it to find 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
  - discrete/continuous variables
  - information about the probability distribution of measurements
  - existence of reference database
  - role of manufacturing process

The Forensic Examination

- Daubert factors relevant for establishing validity of scientific testimony
  - 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

- 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"
  - validity as applied - information about specific examiner

The Forensic Examination

- A community in transition
  - National Commission on Forensic Science
    (advisory to Attorney General) - NOW OUT OF BUSINESS
  - Organization of Scientific Area Committees (NIST)
  - Research
    - NIJ-funded projects
    - NIST Forensic Sciences (intramural program)
    - NIST Center of Excellence
      (CSAFE = Center for Statistical Applications in Forensic Evidence)

Common Approaches to Assessing Forensic Evidence

- Expert assessment based on experience, training
- Statistical approaches
  - Statistical-test based approach
  - Likelihood ratio
- Ultimately perhaps a combination of the two?

Signif. Test/Coincidence Prob. Approach

- One common statistical approach solves the forensic problem in two stages
- Step 1: determine if the crime scene and suspect objects agree on characteristic of interest
  (typically using a hypothesis/significance test)
- Step 2: assess the significance of this aagreement by finding the likelihood of such agreement occurring by chance
- Also known as the comparison/significance approach
- Note: DNA analysis can be categorized in this way but is usually thought of as a likelihood ratio approach

Signif. Test/Coincidence Prob. Approach

- Determining agreement is straightforward for discrete data like blood type, gender
  - note that there are still limitations in these cases due to the possibility of laboratory or measurement error
  - usually more straightforward to think about discrete data in terms of the likelihood ratio
- Statistical significance tests can be used for continuous data like trace element concentrations (e.g., in glass fragments)

Signif. Test/Coincidence Prob. Approach

- Testing procedure
  - characterize each object by mean value (e.g., mean trace element concentrations in set 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 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 no agreement
  - otherwise can't reject the hypothesis that the two means are equal (.... but is this evidence that they came from the same population?)

Signif. Test/Coincidence Prob. Approach

- Example 1: Two glass samples (from Curran et al. 1997)
- Five measurements of aluminum concentration in crime scence 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
- In fact, these are 10 measurements from same bottle

Signif. Test/Coincidence Prob. Approach

- Example 2: 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

$$.929, .859, .845, .931, .915$$

- Control: mean $= .730$, std.err.$=.0435/\sqrt{5} = .019$
- Sample: mean $= .896$, std.err.$=.0408/\sqrt{5} = .018$
- Test statistic $= \frac{.896-.730}{\sqrt{.019^2+.018^2}} = \frac{.166}{.026} = 6.38$
- p-value $= .00015$ ..... reject hypothesis of equal means
- In fact, these are from two different bottles

Signif. Test/Coincidence Prob. Approach

- Alternative methods exist
  - 3-sigma methods create interval for each element in each sample (mean conc. +/- 3 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)

Signif. Test/Coincidence Prob. Approach

- Before moving to assessing the probability of a coincidental match it is important to understand some objections to testing
  - significance tests do not treat the two hypotheses (equal mean, unequal mean) symmetrically
    - null hypothesis (equal mean) is assumed true unless the data rejects
    - acceptance of null is taken as evidence against suspect (asymmetry of null/alternative is an issue here)
  - binary decision of match/no-match requires an arbitrary cutoff (e.g., why 3 sigma rather than 2.5 sigma)
  - separation of match decision from assessment of the significance of the match

Signif. Test/Coincidence Prob. Approach

- Some more technical objections
  - many test procedures (t-test, Hotelling's test) require assumptions about the probability distribution of the data
  - univariate procedures are repeated on multiple elements which must be accounted for
  - univariate procedures ignore information in the correlation of elements
  - multivariate procedures require large samples

Signif. Test/Coincidence Prob. Approach

- Alternatives to significance tests
  - Equivalence testing instead of significance testing (changes the null hypothesis)
  - Bayesian approach and the likelihood ratio (more on this later)

Signif. Test/Coincidence Prob. Approach

- Second stage of analysis is assessing the "significance" of a match
- Because significance has a formal statistical meaning we try not to use that term here
- Other terms - strength of evidence, quality of evidence, usefulness of evidence, probative value
- Examples: we know that suspect and criminal
  . . . are both male - limited usefulness
  . . . are both one-armed males - more useful
- Key idea is to determine probability of a match by coincidence
- This step is crucial for courtroom setting

Signif. Test/Coincidence Prob. Approach

- Discrete data (e.g., blood type, DNA)
  - want to find 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 Chinese, etc.)
    - where do data come from (population records, convenience sample)
    - discussed further in the context of likelihood ratios later

Signif. Test/Coincidence Prob. Approach

- Continuous data
  - typically a bit harder to do
  - need likelihood that objects (e.g., glass fragments) selected at random would match crime scence (control) sample
  - basic idea (described in terms of t-test)
    - 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 to the "control" sample
    - then total coincidental agreement probability is an average over all possible choices for the random source

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

  - this is technically challenging but it can be done
- Key question - where does information about the population come from?

Signif. Test/Coincidence Prob. Approach

- Take $\bar{X} = .730, s.d. = .04, n = 5$ as in Curran et al. glass example
- Use cutoff corresponding to a p-value roughly .05
- Suppose that means of different glass samples in the population can be described as a normal distn
- Results
  - if popul. mean is .73 and s.d. .20 then coincidence prob is .20
  - if popul. mean is .73 and s.d. .10 then .... .37
  - if popul. mean is .73 and s.d. .05 then .... .65
  - if popul. mean is .83 and s.d. .20 then .... .37
  - if popul. mean is .83 and s.d. .10 then .... .24
  - if popul. mean is .83 and s.d. .05 then .... .17
  - if popul. mean is .93 and s.d. .20 then .... .12
  - if popul. mean is .93 and s.d. .10 then .... .06
  - if popul. mean is .93 and s.d. .05 then .... .002

Signif. Test/Coincidence Prob. Approach

- Probability of a coincidental match (i.e., related to false match probability) is high when:
  - small difference between control sample and population of samples we are likely to find (i.e., control sample is "ordinary")
  - large amount of heterogeneity among samples in the population
  - large amount of variability in the measurement process

Likelihood Ratio Approach
Introduction

- Real goal in courtroom setting is statement about relative likelihood of two hypotheses (prosecution, defense) given data
- In statistical terms this is a Bayesian formulation (asks for probabilities about the state of the world given observed data)
- Review of Bayes' rule (or Bayes' Theorem):
  Given two events A and B we have

$$\Pr(A|B) = \Pr(A)\Pr(B|A)/\Pr(B)$$

- This is a way of reversing direction of conditional probabilities ...
  we go from statements about likelihood of
  evidence given hypothesis to statements about
  likelihood of hypothesis given evidence

Likelihood Ratio Approach
Introduction

- Formally using E (evidence) and S (same source)

$$\Pr(S|E) = \frac{\Pr(S)\Pr(E|S)}{\Pr(E)} = \frac{\Pr(S)\Pr(E|S)}{\Pr(S)\Pr(E|S) + \Pr(\bar{S})\Pr(E|\bar{S})}$$

- Bayes' rule can be rewritten 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
- Thus likelihood ratio (sometimes known as the Bayes factor) is a ("the"?) measure of the value of the evidence (does not depend on prior beliefs)

Likelihood Ratio Approach
Introduction

- Reminder: $LR = \frac{\Pr(E|S)}{\Pr(E|\bar{S})}$
- Some 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 (more on this later)
  - other available information (e.g., background) can be incorporated into the LR (more on this later)
- Interpretation
  - Derivation shows that LR is factor we should use to change our odds
  - There are some proposals for scales (e.g., ENFSI) that map LRs to words:
    2-10: weak support; 10-100: moderate support; etc.

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 poterior odds
- LR vs Bayes Factor
  - Distinction is technical and has to do with how statistical parameters are treated

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

- Makes explicit the need to consider the evidence under two different hypotheses
- Separates "objective" information about evidence from "subjective" assessments of hypothesis (S)
- There is some 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
  - Defense attorney's fallacy: other misinterpretations of $\Pr(E|\bar{S})$
    - If $\Pr(E|\bar{S}) = 1/1000000$, then there are 300 other people who could have done it

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 crime scene and x is measurement of evidence from suspect
- Note: sometimes role of x and y are reversed
- Replace Pr() by $p()$ to cover discrete and continuous cases
- Then we have from laws of probability

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

- Often likelihood of x is same for S and $\bar{S}$, (i.e., $p(x|S) = p(x|\bar{S})$) i.e.., distribution of suspect data doesn't depend on who committed the crime
- If so ..... $LR = \frac{p(y|x,S)}{p(y|x,\bar{S})}$

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
    (Does zero really mean zero? We should consider lab error or other explanations)
  - Denominator is probability of a coincidental match
- Continuous case
  - Numerator is measure of how likely it is to observe the numbers x, y if they represent multiple measures from the same source
  - Denominator is a measure of how likely it is to observe the numbers x, y if measures from different sources

Likelihood Ratio Approach

A simple example

- Suppose evidence is blood types for a crime scene sample and suspect sample
- We have information about the distribution of blood types in the population

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

- Suppose both samples are observed to be of blood type O
- $\Pr(y = O|x = O, S) \approx 1$
  (we'd expect to see the same blood type if S is true)
- $\Pr(y = O|x = O, \bar{S}) = 0.44$
  (type O blood is relatively common in the U.S.)
- $LR \approx 1/0.44 \approx 2.3$
- Evidence provides weak support for "same source"
- Blood type AB is rare in the U.S.. If the two samples were both AB, then LR would indicate stronger evidence (LR would be $1/ .04 = 25$)

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 pre-2017)
- Locations on different chromosomes are independent
- Recall that if independent we can multiply probabilities (which basically means multiplying likelihood ratios)
- A match at all location 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

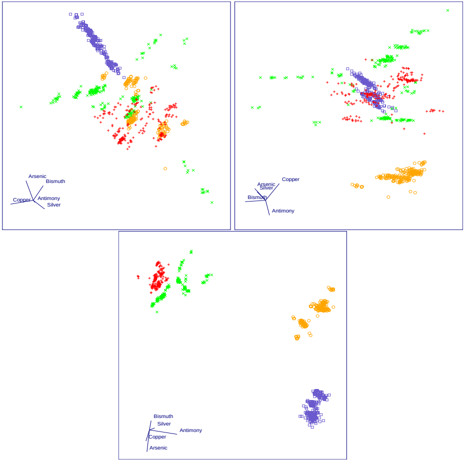
Where it can work .... Trace evidence

- 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 from elemental analyses of bullet lead and refractive indices of glass

Likelihood Ratio Approach

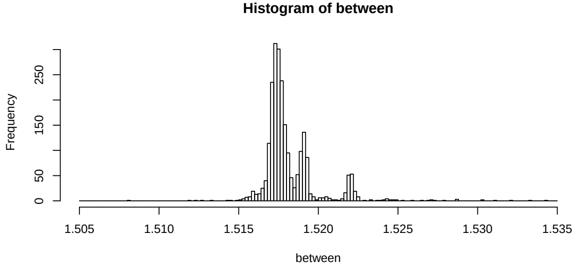
Where it can work .... Trace evidence

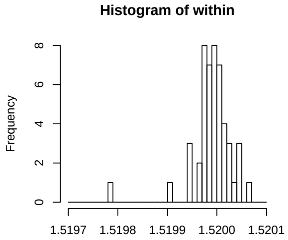
Three projections of 5-dimensional trace element data

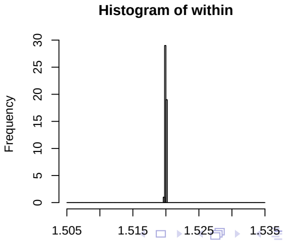


Likelihood Ratio Approach

Where it can work .... Trace evidence







Likelihood Ratio Approach

Continuous measures for trace data - example

- Take y and x to be trace element concentrations for a single element from glass fragments at the scene (y) and on the subject (x)
- Can generalize to multiple x, y measurements
- Assume normal distn for trace element concentrations (may be more reasonable for logarithms)
- Within a single source (e.g., sheet of glass) assume that $x \sim N(\mu_s, \sigma^2)$ where μ_s is the mean concentration in the sheet and σ^2 is the (presumably small) variance
- Two different sheets will have different μ_s's
- Assume sheet-to-sheet variation is described by normal distribution, $N(\mu, \tau^2)$, where μ is a "manufacturer" mean and τ^2 is sheet-to-sheet variance
- Implication of the two normal assumptions: distn of a random fragment is $N(\mu, \tau^2 + \sigma^2)$

Likelihood Ratio Approach

Continuous measures for trace data - example (cont'd)

- The following is not precise (ignores transfer, makes some simplifying approximations)
- Numerator $p(y|x, S)$
  - assume σ^2 is known
  - use x (or avg of x's) to estimate μ_s
  - $p(y|x, S) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y-x)^2/2\sigma^2\}$
  - not exact (treats x as measured precisely)
- Denominator $p(y|x, \bar{S})$
  - assume σ^2 and τ^2 are known
  - assume μ (company mean) is known, otherwise use x to estimate
  - $p(y|x, \bar{S}) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} \exp\{-(y-\mu)^2/2(\sigma^2+\tau^2)\}$
- Likelihoods depend on some parameters (μ, σ, τ) which must be estimated from available data (once again the challenge of finding representative population data)

Likelihood Ratio Approach

Continuous measures for trace data - example (cont'd)

$$LR = \sqrt{1 + \frac{\tau^2}{\sigma^2}} \exp\left\{ \frac{(y-\mu)^2}{2(\sigma^2+\tau^2)} - \frac{(y-x)^2}{2\sigma^2} \right\}$$

- LR is small if $y - x$ is large (no match)
- LR is big if $y - \mu$ is large and $y - x$ is small (a match on an unusual value)
- Evett's work supplies values of parameters for refractive index, typical LRs are order of magnitude 100
- LR idea works in same basic way for other non-normal distributions

Likelihood Ratio Approach

Where it can might 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 ...
  - Likelihood ratios can be very sensitive to assumptions
  - Assessing the "population" is hard (and may vary from case to case)

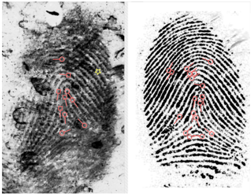
Likelihood Ratio Approach

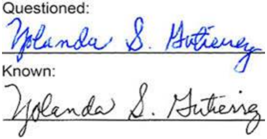
Where it might work .... Pattern evidence

- Many forensic disciplines are focused on comparing a sample (mark) at the crime scene ("unknown" or "questioned") and a potential source ("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
  - Firearms
  - Tool marks

Likelihood Ratio Approach

Where it might work .... Pattern evidence





Likelihood Ratio Approach

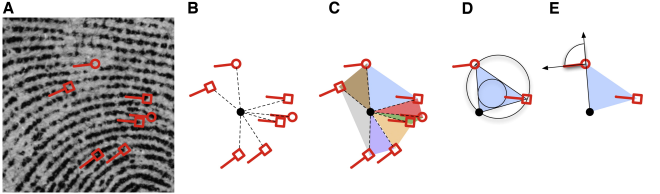
Where it might work .... Pattern evidence

- A number of challenges in constructing likelihood ratios
  - Data are very high dimensional (often images)
  - Flexibility in defining the numbers/types of features to look at
  - Lack of probability models for multivariate features / patterns
  - Need to study variation across a relevant population
- Very hard, but there is work under way

Likelihood Ratio Approach

Where it might work .... Pattern evidence

- Neumann et al. (2015) approach

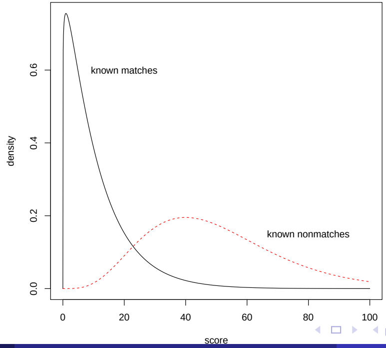


- Each minutiae is characterized by direction/angle, type of minutiae, shape/configuration
- Separate likelihood ratio for each of these characteristics
- Numerator based on variation in same finger (obtained from distortion model)
- Denominator based variation across diff't fingers (uses nearest non-match from a database search)

Likelihood Ratio Approach

Score-based likelihood ratios

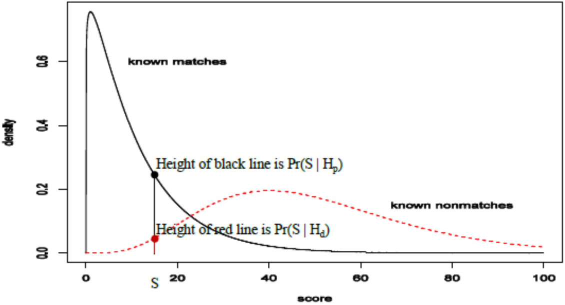
- Given the challene in developing complete LR for pattern evidence there is some interest in an empirical score-based approach
- Define a score measuring the "difference" between the questioned and known sample
- Obtain distribution of scores for a sample of known matches
- Obtain distribution of scores for a sample of known non-matches



Likelihood Ratio Approach

Score-based likelihood ratios

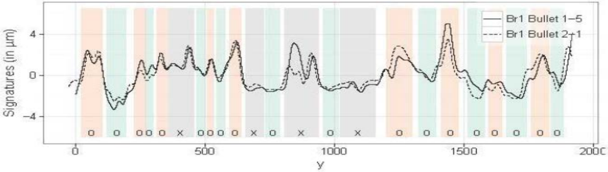
- The idea
  - Fit a probability distribution to the scores of known matches ($Pr(S|H_p)$)
  - Fit a probability distribution to the scores of known nonmatches ($Pr(S|H_d)$)
  - Score-based likelihood ratio if we observe score S is $SLR = Pr(S|H_p)/Pr(S|H_d)$

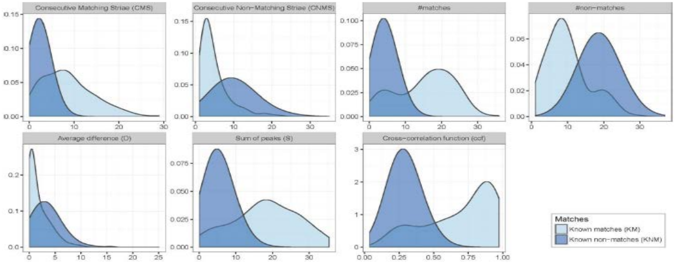


Likelihood Ratio Approach

Score-based likelihood ratios

- Example - bullet land signatures





Likelihood Ratio Approach

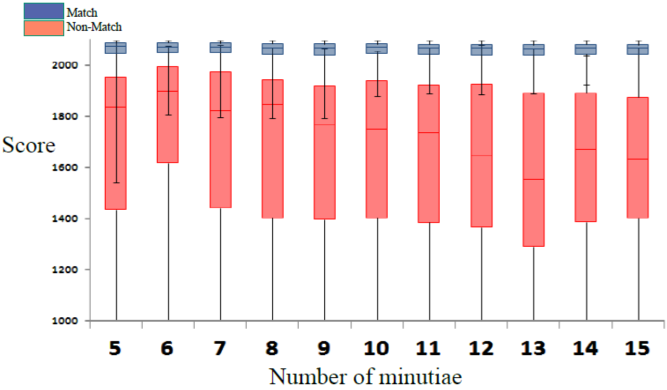
Score-based likelihood ratios - challenges

- Across a number of existing examples the score distribution for known matches seems relatively straightforward to characterize
- There are challenges though in definiing 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?

Likelihood Ratio Approach

Score-based likelihood ratios - challenges

Example from DFSC's evaluation of a latent print score function



Likelihood Ratio Approach
Complications

- It is not hard to find issues that complicate the calculation of LR
  - accounting for transfer evidence with glass or fibers
  - bullets in a box are heterogeneous (containing representatives from a number of different manufacturing lots)
  - usage/lifetime of products (e.g., sneakers)
- Though good work is being done it will take some time before LRs are available for pattern evidence

Likelihood Ratio Approach

Contextual information

- There is some discussion in the community about the role of contextual bias, task-relevant/task-irrelevant information, etc.
- Likelihood ratio framework can accommodate this discussion
- $LR = \Pr(E|S, I)/\Pr(E|\bar{S}, I)$
  where I refers to other information that is being considered
- I should include task-relevant information (e.g., substrate)
- I should not include results of other forensic examinations, other case information

Likelihood Ratio Approach

Multiple sources of evidence

- Likelihood ratio can accommodate multiple types of evidence
- $LR = \Pr(E_1, E_2|S)/\Pr(E_1, E_2|\bar{S})$
  where E_1, E_2 are two evidence types
- If evidence types are independent, then this simplifies considerably
  $LR = \frac{\Pr(E_1|S)}{\Pr(E_1|\bar{S})} \times \frac{\Pr(E_2|S)}{\Pr(E_2|\bar{S})}$
- If dependent, then the joint analysis can be tricky

Likelihood Ratio Approach
ENFSI Guideline for Evaluative Reporting

- Reporting requirements
  - balance - should consider two propositions
  - logic - should focus on likelihood of evidence given hypotheses
  - robustness - should withstand scrutiny
  - transparency - clear case file and report
- Propositions (different kinds, etc.)
- Assignment of likelihood ratio
  - data and/or expert knowledge used to assign probabilities required for likelihood ratio
  - subjective elements can be used
  - avoid undefined qualifiers (e.g., rare)
  - account for uncertainty
- LR forms basis for evaluation (verbal equivalents)

Likelihood Ratio Approach
Summary

- Advantages
  - explicitly compares the two relevant hypotheses/propositions
  - provides a quantitative summary of the evidence
  - 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
- Disadvantages
  - requires assumptions about distributions
  - 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

Forensic Conclusions as Expert Opinion

- Forensic evidence enters the courtroom through expert testimony
- Expert analysis in most domains is based on experience, training, use of accepted methods
- Issues to consider
  - range of conclusions
    - identification, inconclusive, exclusion
    - multi-point scales (some support, strong support, very strong support, etc.)
  - testifying in this way requires validation of the expert opinion (PCAST)

Forensic Conclusions as Expert Opinions

- Statistical methods are relevant to carrying out reliabaility and validation studies
- Reproducibility - how often would the same examiner reach the same conclusion for given evidence
- Reliability - how often would different examiners reach the same conclusion for given evidence
- "White Box" Studies - studies of repeatability and reproducibility of different aspects of the forensic examination
- Validation studies
  - "Black Box" Studies of performance - examiners given cases with known "ground truth" to assess frequency of different types of errors e.g., Ulery et al. 2011 in PNAS for latent prints
  - One way to think about this is that now E = examiners conclusion; Need to assess $P(E|S)$ and $P(E|\bar{S})$
  - Importance of earlier discussion regarding study design

Forensic Conclusions as Expert Opinions

- There will always be unique situations (e.g., did this typewriter produce this note?) for which there are no relevant validation/reliability studies
- This is not a problem ...
  But the conclusions expressed by the expert must acknowledge uncertainty about the likelihood of a coincidental agreement

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 testing-based approaches and likelihood ratios to forensic examinations
- Key points
  - Any approach must 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
- Ongoing discussion about how appropriate source conclusions in the OSAC