Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

October / November 2022

Outline

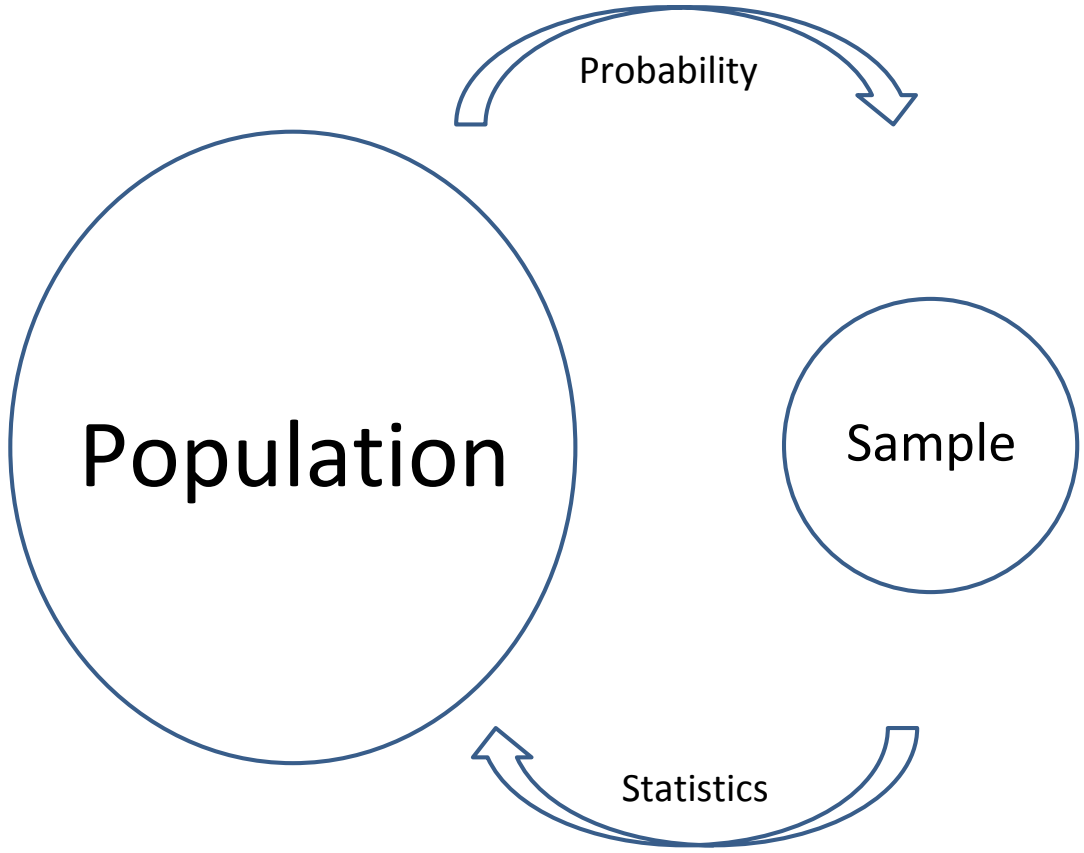
- Part 1 - Probability Concepts and Their Relevance to Forensic Science
  - review of probability concepts
  - conditional probability and independence
  - Bayes' Theorem and likelihood ratio
- Part 2 - Data, Measurement, Reliability and Expert Opinion
  - collecting data
  - measurement, variability, reliability and accuracy
  - forensic evidence evaluation as expert opinion / black box studies
- **Part 3 - Statistical Inference and the Two-Stage Approach to Assessing Forensic Evidence**
  - **estimation, 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 3

- Understand how probability and statistics tools can be used as a basic for inference about a population
- Understand principles of point estimation and interval estimation
- Understand how statistical hypothesis tests work and their limitations
- Understand the strengths and weaknesses of the two-stage approach to assessing forensic evidence

Probability and Statistical Inference

Recall "The Big Picture"



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

Probability

A short review

- Probability is the mathematical language of uncertainty
- Provides a common scale (0 to 1) for describing the chance that an event will occur
- Need to think about where probabilities come from – data, theory, subjective opinion
- Conditional probability is a key concept ...
  the probability of an event depends on what information is considered
- Independent events can be powerful (allows us to multiply probabilities as is common in DNA analysis) ... but the assumption needs to be confirmed
- Important to carefully interpret conditional probability $P(A \mid B)$
  - what events are we assigning probabilities to (the event A)
  - what information are we assuming to be true (the event B)

Data, Measurement, Reliability and Expert Opinion

A short review

- 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

Data, Measurement, Reliability and Expert Opinion

A short review

- 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)

Motivation - ASTM E2927-16

- ASTM E2927-16: Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons
  - Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition)..... **The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass**."
  - The "Big Picture" applies in this situation as well
  - Now **two populations** (one corresponding to known source and one corresponding to questioned source)
  - Question of interest is whether these populations differ in important ways (are distinguishable)

Motivation - ASTM E2927-16

- 11. Calculation and Interpretation of Results

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

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

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

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

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

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

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

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

- This is a statistical inference procedure!

Probability to Statistical Inference

Summarizing data

- For categorical / ordinal data, we usually summarize data with a table
  e.g, US blood type distribution

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

- For discrete data, we may summarize with a table, graph and numerical summaries
  e.g., CMS in a study of known matching bullet groove impressions

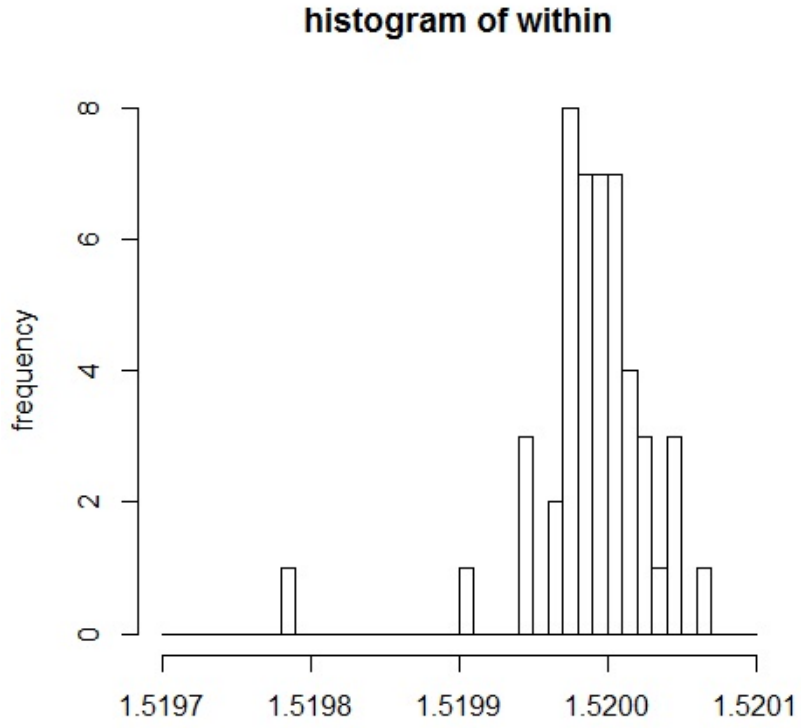
	2	3	4	5	6	7	8	Total
Count	55	54	23	11	2	0	1	146
Proportion	.377	.370	.157	.075	.014	.000	.007	1.000

  - sample mean (average) $= 3.01$, standard deviation $= 1.07$

Probability to Statistical Inference

Summarizing data

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



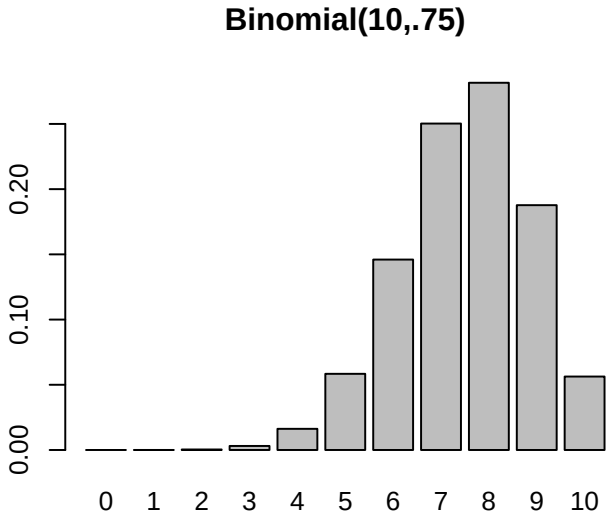
Probability to Statistical Inference

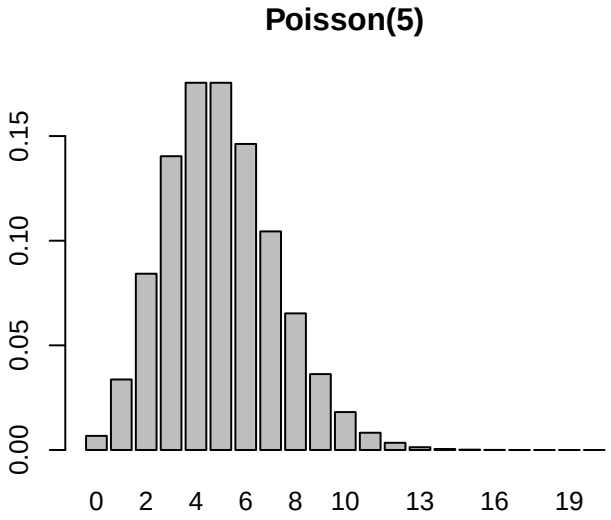
Probability distributions

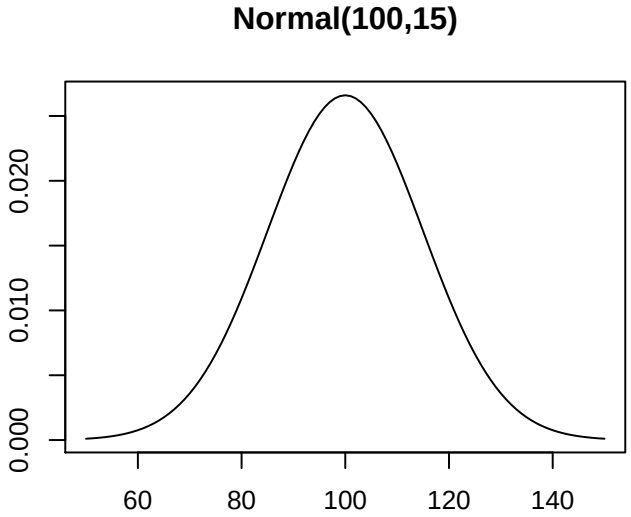
- There are families of well-known probability distributions that are commonly used in statistical analyses
- Examples
  - Binomial: # of successes in n trials
    (e.g., test n bags of contraband and record no. with drugs)
  - Poisson: count # of events
    (e.g., number of consecutive matching stria)
  - normal: bell-shaped curve
    (e.g., measure of weight of packages of drugs found on suspect)
  - log normal: logarithm of observations follow a normal distribution
    (e.g., measure of concentration of chemical in glass)
- We will see that the normal distribution plays a large role in statistical inference

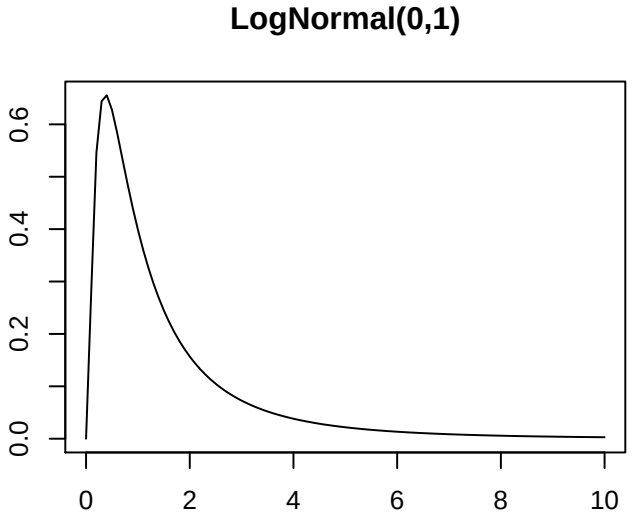
Probability to Statistical Inference

Probability distributions - Examples



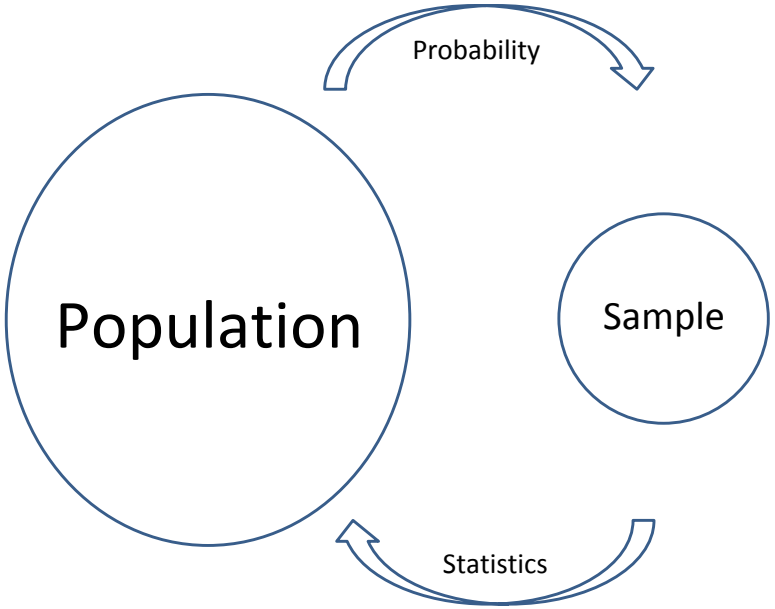






Statistical Inference

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 to carry out statistical inference
  1. build/assume model for population
  2. assess model by comparing sample to what is expected under model
  3. refine model; go back to step 2

Statistical Inference

Background

- Definition - a **parameter** is a numerical characteristic of the population, e.g., population mean, proportion of size 9 shoes
- Statistical methods are usually concerned with learning about population parameters from sample data
- **Key concept:** there is a distinction between the population and a sample
- **Key concept:** the mean of a sample and the mean of a population are different quantities
- We can apply the laws of probability (from earlier in the short course) to draw inferences from a sample
  - observe sample mean
  - if we have a "good" sample, then this should be close to the population mean
  - the laws of probability tells us how close we can expect them to be

Statistical Inference

Background

- Goal: inference about a parameter
- Possible parameters
  - mean concentration of aluminum in population of glass fragments from a given source
  - proportion of bags containing illicit substances
- Different kinds of inferential statements
  - estimate of parameter (point estimate)
  - an interval estimate or range of plausible values for parameter (this provides both a point estimate and a measure of uncertainty associated with the estimate)
  - test a specific hypothesis about the value of a parameter (this can be used for example to tell if two populations have distinguishable means)

Statistical Inference

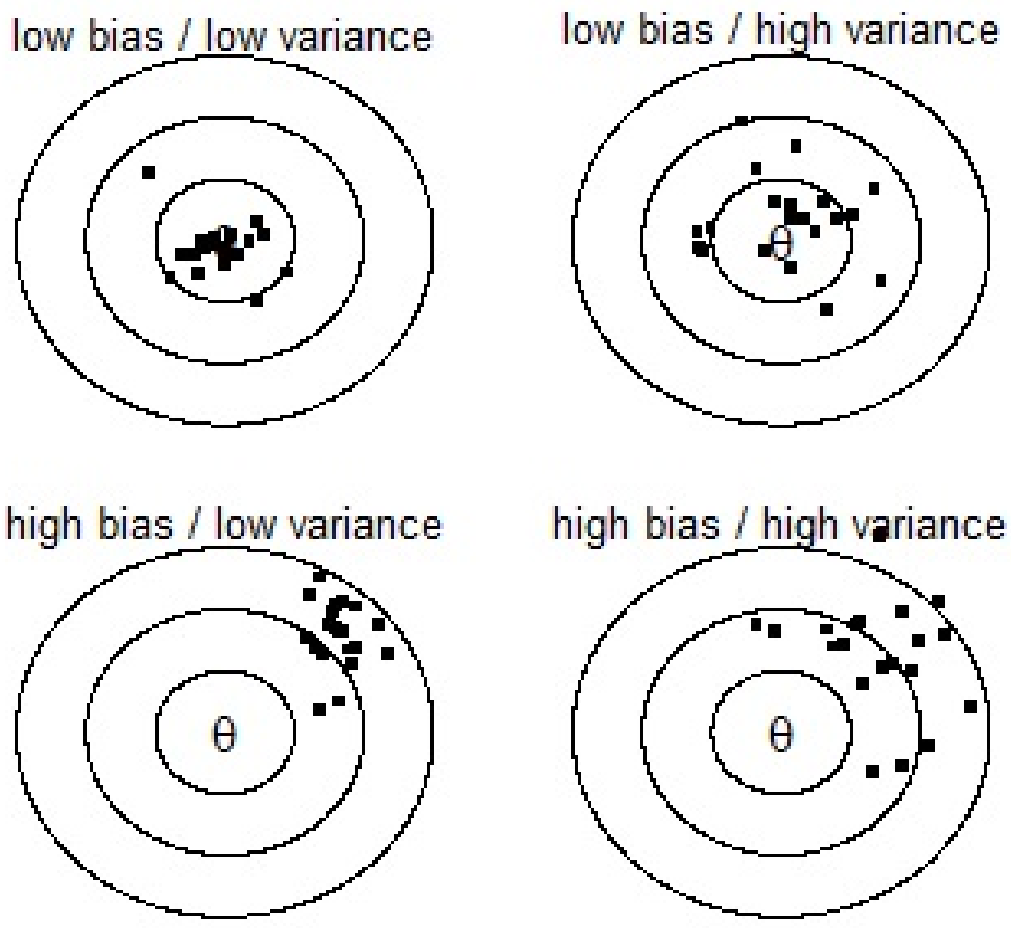
Point estimation

- An estimator is a rule for estimating a population parameter from a sample
- We evaluate estimators by considering certain properties
  - we ask how the estimator performs in repeated samples
  - bias - how close on average to population value
  - variability - how variable is the estimate
- Example - suppose we are interested in estimating the population mean or average
  - the mean of a random sample from the population is one possible estimator (spoiler alert: it is a very good estimator)
  - the median of a random sample is an alternative (less sensitive to wild measurements)
    - if sample=(1,2,3,4,5,6,7,8,90), then mean=14 and the median=5
  - 47 is another possible estimator, i.e., we always estimate 47! (spoiler alert: not very good – unless we are very lucky!)

Statistical Inference

Performance of different estimators for unknown θ

- The figures below are "conceptual" illustrations of bias and variability. The center (θ) is the is the "true" (but unknown) population parameter that we are trying to estimate. The dots represent estimates that we might obtain from different samples.



Statistical Inference

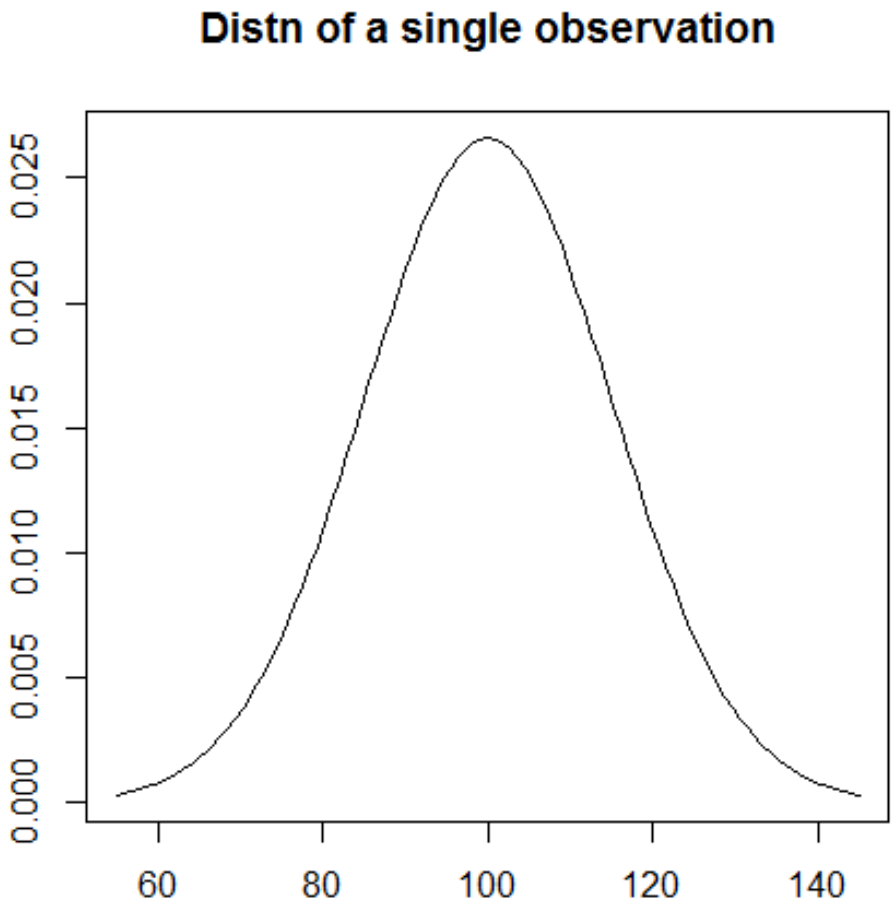
Standard errors

- A limitation of just providing a point estimate is that it doesn't provide any indication of uncertainty
- We can do better than this
- The standard error of an estimator measures the uncertainty in our estimate
  - **Review:** The standard deviation (s.d. or sd) is a measure of the spread (variability) in a sample or in a population (describes uncertainty about a single observation)
  - When we look at a summary statistic (mean, median, percentile) it is also a random quantity (would give diff't value in diff't samples)
  - The standard error is how we measure the variability of an estimator

Statistical Inference

Standard errors

- Consider a normally distributed population with mean 100 and s.d. 15
- This distribution describes IQ scores in the general population
  - expect 68% of observations to be between 85 and 115
  - expect 95% of observations to be between 70 and 130



Statistical Inference

Standard errors

- Suppose we take a random sample of 25 people and give them an IQ test
  - We get these values:
    63, 87, 88, 89, 92, 94, 94, 96, 97, 98, 100, 103, 104, 106, 106,
    107, 108, 109, 111, 114, 115, 118, 126, 136, 142.
  - The sample mean is 104.1 and the sample s.d. is 16.4
  - These are close to the population mean (100) and population s.d. (15)
- Now suppose we take another random sample of 25 people
  - We get these values:
    65, 68, 71, 75, 85, 85, 87, 89, 90, 91, 98, 99, 102, 102, 103,
    103, 103, 105, 105, 106, 109, 110, 111, 120, 122.
  - The mean is 96.2 and the s.d. is 15.2
  - These are again close to the population quantities but they are different than for the first sample
- The different summary statistics in the two samples are to be expected. This represents the variability that we expect in different random samples.

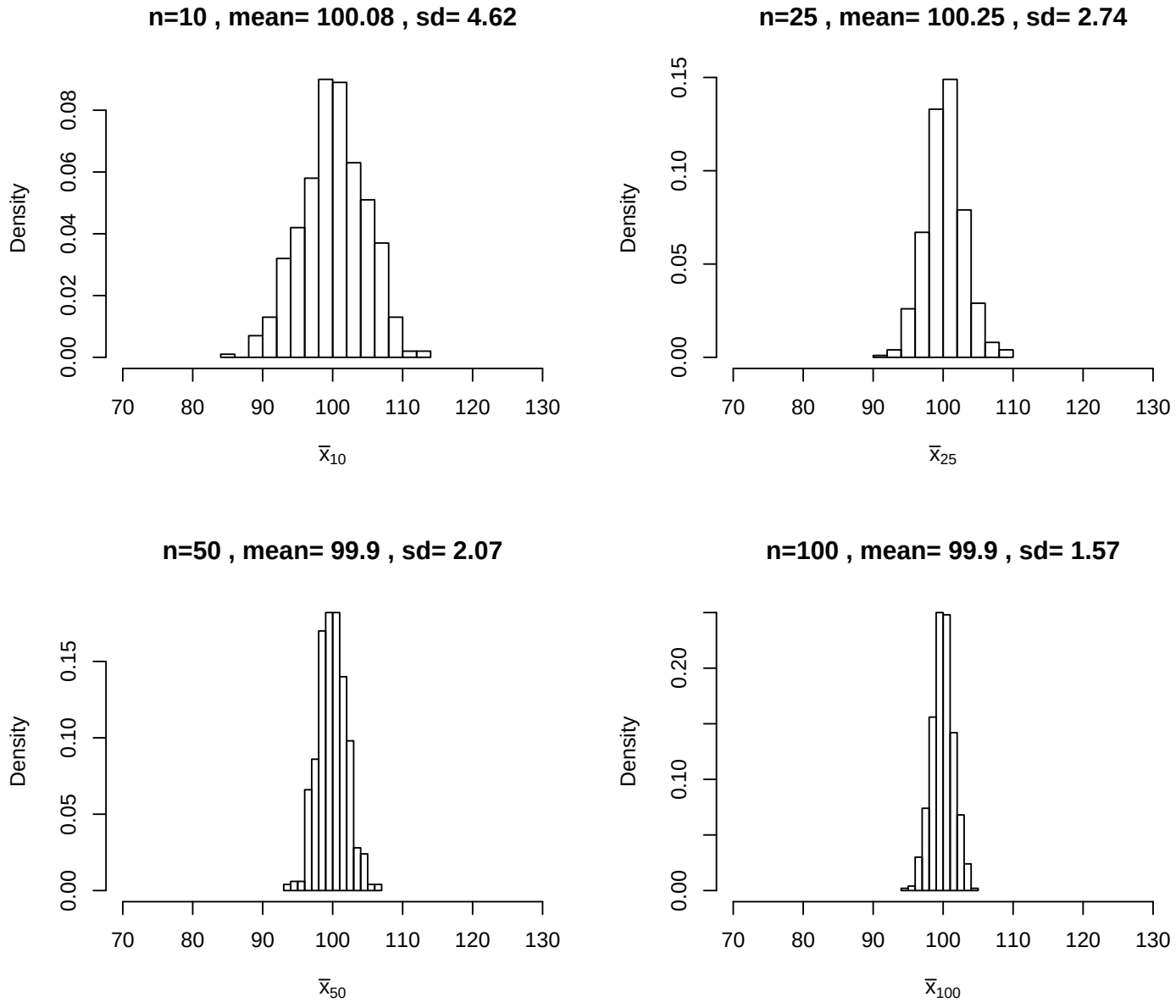
Statistical Inference

Standard errors

- We saw that two different random samples yield two different sample means
- We can study the variability in sample means across many different samples
- This variability is usually measured by the **standard error**
- The standard error is determined by the standard deviation of the individual measurements and the size of the sample
  - standard error $=$ standard deviation / $\sqrt{(\text{sample size})}$
- Note that we can reduce the variability in our sample means by taking bigger samples

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 an approximately 95% confidence interval for a quantity by taking point estimate $\pm$ 2 std errors
- Most common example is trying to estimate the population mean
  - natural point estimate is the sample mean
  - approximate 95% confidence interval is sample mean $\pm$ 2 standard error
  - the "2 standard error" piece is sometimes known as the "margin of error"

Statistical Inference

Interval estimation - example

- Example: 10 glass fragments from crime scene
- Measure concentration of aluminum
- Mean = 0.730, standard devation = 0.040
- Standard error = 0.040 / sqrt(10) = 0.013
- Approximate 95% confidence interval for the mean aluminum concentration in the crime scene window is
  0.73 +/- 2*0.013 = (.704,.756)
- **Interpretation of confidence interval is important:**
  95% **of intervals built in this way will contain the true population parameter**
- Note this type of interval (with higher confidence) is sometimes used in the analysis of glass evidence (ASTM E2927)

Statistical Inference

Interval estimation - important points

- The width of the confidence interval depends on
  - the amount of confidence that we want
    (99% would require a larger margin of error than 95%)
  - the population standard deviation which measures the variability in a single measurement
    (the bigger the s.d., the wider the interval)
  - the number of measurements that we are averaging
    (bigger samples lead to narrower intervals)
  - because of the formula, we would require four times as many samples to cut the width of the interval in half!

Statistical Inference

Hypothesis testing

- Sometimes we wish to formally test a hypothesis about a population parameter
- The hypothesis to be evaluated is known as the null hypothesis and usually refers to an assumption of no difference or no change. Often, we look for evidence against the null hypothesis
- There is an alternative hypothesis that helps us to design the test
- A common scenario is that we want to compare a new medical treatment with the current standard of care (perhaps drugs intended to lower blood pressure)
  - The null hypothesis is that the mean drop in BP is the same for the two drugs ("no change")
  - The alternative hypothesis is that the new drug leads to a bigger mean drop in BP than the current standard of care

Statistical Inference

Hypothesis testing

- Historically, a statistical test would set a decision rule to decide whether to accept or reject the null hypothesis
- If we reject the null hypothesis then we say we have a statistically significant result
- Two types of errors are possible when carrying out a test
  - type I: reject the null hypothesis when it is true (false positive)
  - type II: fail to reject the null when it is false (false negative)
- Type I error often considered more serious: we only want to reject the null hypothesis if there is strong evidence against it
- There is a tradeoff involved between the two types of errors. We can eliminate type I errors by devising a strict test, but then we will make more type II errors.

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
- Our thought process is that if we see a big test statistic (i.e., a big difference) then one of two things has happened. Either we observed a random sample where a big difference occurred by chance or the null hypothesis is not true and that led to the big difference.
- How do we decide?

Statistical Inference

Hypothesis testing

- As mentioned, one approach is to set up a decision rule for the test
- More common now to summarize a statistical test by attaching a probability (known as the p-value) to the test statistic
- Definition: a *p*-**value** gives the probability that we would get data like the data we have observed in the sample (or something even more extreme) **given that the null hypothesis is true**
- Small *p*-values mean unusual data that lead us to question the null hypothesis (since sample data like the observed are unlikely to happen by chance)
- However, the *p*-value only addresses the fit of the data to the null hypothesis. It does not speak to the likelihood of the alternative hypothesis being true

Statistical Inference

Hypothesis testing - comparing two means

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

Statistical Inference

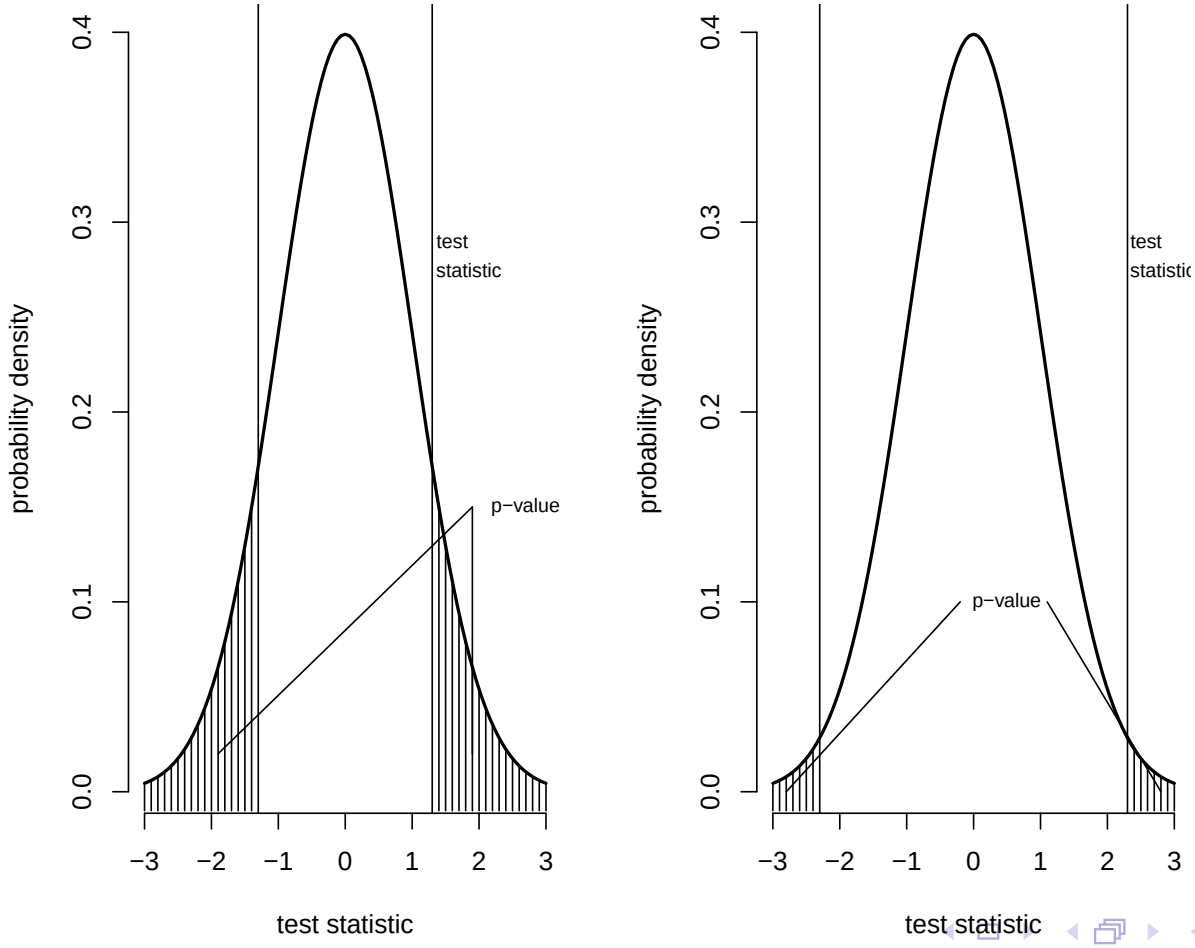
Hypothesis testing - comparing two means

- There is a well established procedure for testing a hypothesis about the equality of means of two normal populations
- Null hypothesis is $H_0 : \mu_{scene} = \mu_{suspect}$
- Alternative hypothesis is $H_a : \mu_{scene} \neq \mu_{suspect}$
- Data Y are measurements from glass fragments at the crime scene
  Data X are measurements from glass fragments found on the suspect
- Test looks at the difference in the two means $(\bar{Y} - \bar{X})$
  (this is called the test statistic)
- Expect this difference to be near zero if the null hypothesis is true
- Reject the null hypothesis if the difference is large compared to the standard error for the difference in two means
- Procedure is known as the t-test and the p-value is easily obtained from a t distribution (software will compute)
- A key statistical theorem guarantees that these procedures work well even if the populations are not normally distributed,
  as long as the sample size is large

Statistical Inference

Two examples of hypothesis testing results

- Figure shows distribution of test statistic we expect to see if the null hypothesis is true (curve) and observed test statistics (lines)
- Left figure: observed test statistic = 1.3, p-value = 0.19
- Right figure: observed test statistic = 2.3, p-value = 0.02



Statistical Inference

Hypothesis testing and forensic science

- The logic of statistical hypothesis tests is sometimes related to concepts in the justice system
  - null hypothesis = innocent, alternative = guilty
  - type I error is to decide guilty when person is innocent
    (Is this a false positive though?)
  - type II error is to decide innocent when person is guilty
- It is not at all clear that statistical hypothesis tests are a good match to analysis of forensic evidence
  - The ASTM procedure is essentially a statistical hypothesis test
  - There are logical problems with this approach
    - ASTM standard indicates that failure to reject the null hypothesis suggests the two populations are *indistinguishable*
    - The statistical test actually says *we can't distinguish the means*
    - These are not exactly the same thing – the former seems more incriminating

Statistical Inference

Hypothesis tests and confidence intervals

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

Statistical Inference

Hypothesis testing - discussion

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

Test yourself

Statistical inference I

- To estimate the amount of narcotics contained in 1000 confiscated bags, a random sample of 50 bags is obtained and analyzed. A 95% interval estimate for the mean weight for the population is obtained by computing the mean of the 50 sample bags and then adding/subtracting 2 standard errors. For each change in the study design, tell whether the interval would get wider or narrow:
  - If a sample of 100 bags was used instead
  - If a 99% interval estimate was used instead
  - The population actually included 10,000 confiscated bags

Test yourself

Statistical inference I - answer

- To estimate the amount of narcotics contained in 1000 confiscated bags, a sample of 50 bags is obtained and analyzed. An 95% interval estimate for the mean weight for the population is obtained by computing the mean of the 50 sample bags and then adding/subtracting 2 standard errors. For each change in the study design, tell whether the interval would get wider or narrow:
  - If a sample of 100 bags was used instead –
    NARROWER, a larger sample reduces the standard error
  - If a 99% interval estimate was used instead –
    WIDER, a wider interval is required to be more confident
  - The population actually included 10,000 confiscated bags –
    IT WOULD STAY THE SAME (trick question!)
    Our inference depends on the sample size but not the population size (as long as the population is large and the sample is random)

Test yourself

Statistical inference II

- In the first stage of a forensic examination of glass evidence, the mean aluminum concentration in the crime scene glass sample and the mean aluminum concentration in glass fragments found on the suspect are compared. A statistical test is used to test the hypothesis that the populations from which the two samples come have the same mean. The p-value for the test turns out to be .23. Which of the following statements are true?
  - We would be likely to reject the null hypothesis of equal means and declare the samples distinguishable.
  - The high p-value means these data could have occurred by chance if the samples came from the same source so we do not reject the null hypothesis.
  - These samples can't be distinguished based on these data
  - The samples came from the same window

Test yourself

Statistical inference II - answer

- In the first stage of a forensic examination of glass evidence, the mean aluminum concentration in the crime scene glass sample and the mean aluminum concentration in glass fragments found on the suspect are compared. A statistical test is used to test the hypothesis that the populations from which the two samples come have the same mean. The p-value for the test turns out to be .23. Which of the following statements are true?
  - We would be likely to reject the null hypothesis of equal means and declare the samples distinguishable. - NOT TRUE
  - The high p-value means these data could have occurred by chance if the samples came from the same source so we do not reject the null hypothesis. - TRUE
  - These samples can't be distinguished based on these data - KIND OF TRUE (the population means can't be distinguished)
  - The samples came from the same window. - NOT TRUE, WE CAN'T BE SURE OF THIS

Statistical Inference

Some key takeaways for forensic practitioners

- Statistical inference uses sample data to draw conclusions about a population
- Point estimation, interval estimation, hypothesis tests are main tools
- Critical that procedures account for variation that could be observed due to chance
- Intervals and tests play a significant role in analyses of some evidence types
- Statistical hypothesis tests can be useful but ...
  - it is difficult to interpret p-values
  - limitations in the standard approach (assumes null hypothesis is true until proven otherwise)

The Forensic Examination

- There are a range of questions that arise in forensic examinations - source conclusions, timing of events, substance ID, cause/effect
- Focus today on source conclusions
  - topics addressed (e.g., need to address uncertainty, logic of the likelihood ratio) apply more broadly
  - Evidence E are items/objects found at crime scene and on suspect (or measurements of items)
    - occasionally write E_c(crime scene), E_s(suspect)
    - may be other information available, I
      (e.g., evidence substrate)
  - Two hypotheses
    - H_s - items from crime scene and suspect have the same (common) source
      (or suspect is source of crime scene item)
    - H_d - diffrent source / no common source
  - Goal: assessment of evidence
    - do items appear to have a common source
    - how unusual is it to find observed evidence / observed agreement by chance

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 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

The Two-Stage Approach

- One common statistical approach solves the forensic problem in two stages
- Stage 1 (Similarity)
  - determine if the crime scene and suspect objects agree on one or more characteristics of interest (typically using a hypothesis/significance test)
  - two samples may be described as "indistinguishable", "can't be distinguished", "match"
- Stage 2 (Identification)
  - assess the significance of this agreement by finding the likelihood of such agreement occurring by chance
- Has a long history (Parker and Holford papers in the 1960s)
- Also known as the comparison/significance approach
- Used in assessment of trace evidence (e.g., glass)
- Conceptually many other disciplines appear to act in this way

The Two-Stage Approach

- Stage 1 - Similarity
  - Determining agreement is straightforward for discrete data like blood type or DNA alleles
  - Statistical significance tests (or other procedures) can be used for continuous data like trace element concentrations in glass
  - We will see (as noted earlier) that there are conceptual problems with this approach
  - Note also that there is a loss of information in summarizing the evidence by a binary decision
    - "Can't be distinguished" might mean an exact match of measurements
    - "Can't be distinguished" might mean difference between samples just misses being statistically significant

The Two-Stage Approach

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

The Two-Stage Approach

- Example: Two glass samples (data from Curran et al. 1997)
- Five measurements of aluminum concentration in crime scene sample

$$.751, .659, .746, .772, .722$$

- Five measurements of aluminum concentration in recovered sample

$$.752, .739, .695, .741, .715$$

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

The Two-Stage Approach

- Example: Two glass samples (data from Curran et al. 1997)
- Five measurements of magnesium concentration in crime scene sample

$$.267, .227, .220, .262, .258$$

- Five measurements of magnesium concentration in recovered sample

$$.117, .090, .113, .117, .109$$

- Control: mean = .247, std.err.=$.0216/\sqrt{5} = .0097$
- Sample: mean = .109, std.err.=$.0112/\sqrt{5} = .0050$
- Test statistic = $\frac{.247-.109}{\sqrt{.0097^2+.0050^2}} = \frac{.138}{.0109} \approx 12.7$
- p-value = .000001 ..... clear evidence to reject hypothesis of equal means
- We would say these two samples are "distinguishable"
- In fact, these are from two different bottles (one brown, one colorless)

The Two-Stage Approach

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

The Two-Stage Approach

- There are a number of technical statistical issues associated with the use of these procedures
  - the formal test procedures (t-test, Hotelling's test) require assumptions about the probability distribution of the data
  - univariate procedures are repeated on multiple elements and the existence of multiple comparisons should be accounted for
  - univariate procedures ignore information in the correlation of elements
  - multivariate procedures (like Hotelling's test) require large samples
- But we do not focus on these
- **The more important concerns are conceptual**

The Two-Stage Approach

- Stage 1 - Concern about the role of the null hypothesis
  - Significance tests do not treat the two hypotheses (equal means, unequal means) symmetrically
  - The null hypothesis (equal means) is assumed true unless the data suggest rejecting this hypothesis
    - In this setting a Type I error is a false exclusion (not a "false positive" as typically assumed)
    - In this setting a Type II error is a false inclusion (not a "false negative" as typically assumed)
  - It seems we have the wrong null hypothesis, i.e., the wrong "starting point". In forensics, the null should perhaps be that the samples are distinguishable.
  - I will avoid using Type I and Type II error terminology (common in statistics textbooks) in our discussion

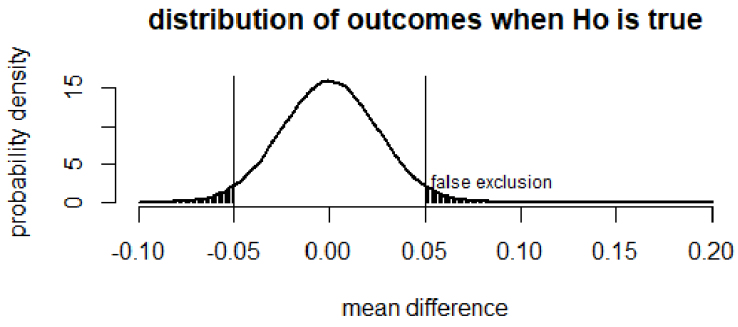
The Two-Stage Approach

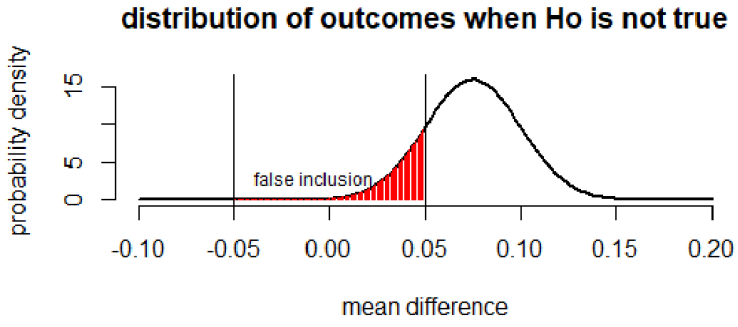
- Stage 1 - Concern about using a binary decision
  - A binary decision (to reject the null hypothesis or not) requires the selection of a cutoff or threshold (e.g., .05 p-value or 4-sigma interval)
  - Choice of threshold impacts the error rates associated with the test
    - a low "threshold" makes it easy to reject ... risks a false exclusion error which rejects a true match and potentially fails to include important evidence
    - a high "threshold" makes it easy to accept the null ... risks a false inclusion error which declares non-matching populations as indistinguishable and could thus incriminate incorrectly
  - It is also the case that inferior measurement protocols (i.e., those with greater variability) will make it easier to accept the null hypothesis

The Two-Stage Approach - thresholds / error rates

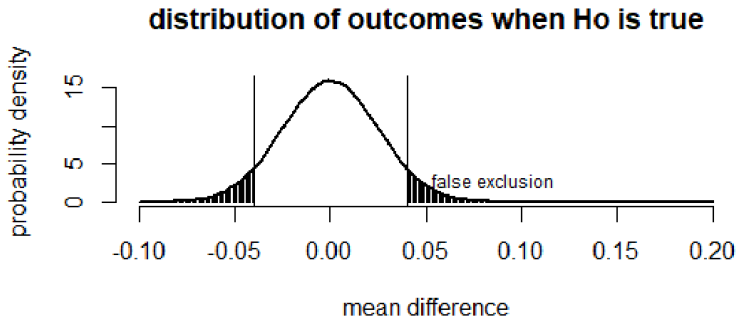
(Black=false exclusion; Red=false inclusion)

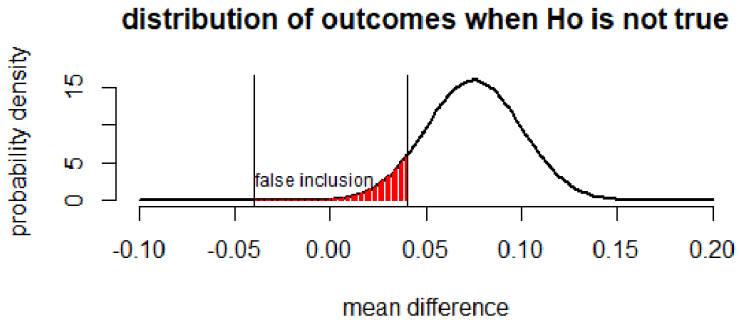
Initial threshold →
low false exclusion rate; higher false inclusion rate





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





The Two-Stage Approach

- Stages 1 and 2 - Concern about separation of the match/non-match decision from the assessment of the probative value
  - Separation into two stages where the first stage may end the analysis is not optimal
  - Much recent attention on this issue in the statistics community
    - failing to find a significant difference is not the same as finding that the null hypothesis is true
    - finding a statistically significant difference may not be practically important
    - the major U.S. statistical association recently recommended not using the term "statistically significant"
  - This issue is not unique to forensics and can also be found in medical research and other disciplines

The Two-Stage Approach

- There are approaches that can address these concerns
  - Equivalence testing instead of significance testing
    (changes the null hypothesis and addresses the first concern)
    - Requires us to specify a "practically" important difference Δ

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

    - The null hypothesis (now assuming distinguishable items) is assumed true until proven otherwise
  - Bayesian approach and the likelihood ratio address the other concerns
    (avoids binary decision, avoids separation of match/significance)

The Two-Stage Approach

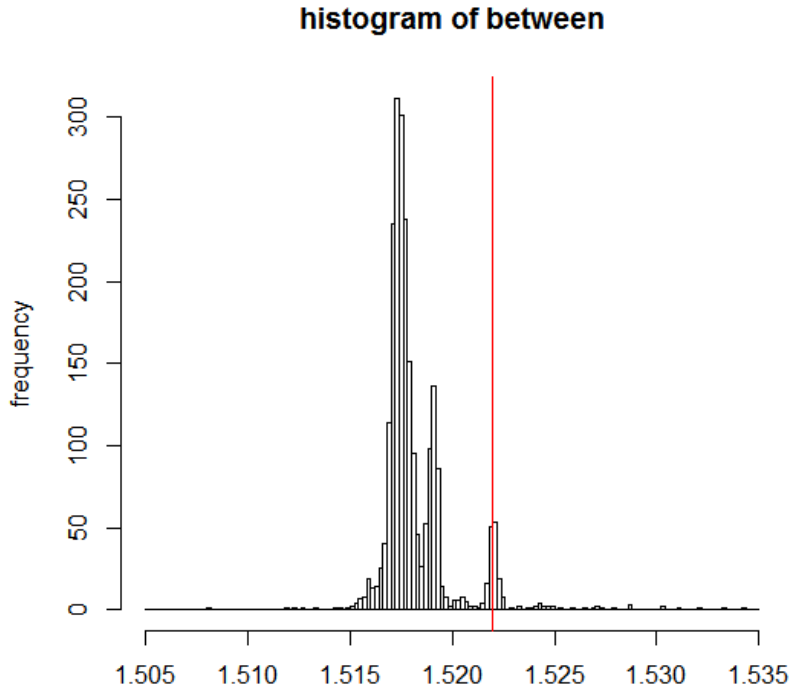
- Now return to the usual significance testing approach and assume we have found a “match” or found the glass to be ”indistinguishable” (i.e., could not reject the null hypothesis)
- Stage 2 - Identification
  - Assess the probability that two samples from different sources would ”match” or be found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

The Two-Stage Approach

- Stage 2 ... the idea for discrete data (e.g., blood type, DNA)
  - Want to find the probability of a match by chance
  - Several important considerations
    - usually crime-scene centered: material from scene is considered fixed and want likelihood that individual would have similar material
    - depends on relevant "information"
      (suspect is male, suspect is of Asian descent, etc.)
    - where do data come from (population records, convenience sample)
  - **Example** - Suppose blood found at the crime scene is type A and the suspect is type A. 42% of US population is blood type A so the matching blood types are not very compelling evidence
  - **Example** - Same situation but with blood type AB. Only 4% of US population is type AB so matching blood types is more informative in this case

The Two-Stage Approach

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



The Two-Stage Approach

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

The Two-Stage Approach

- More realistic to assume many types of glass (i.e., not just three)
- Stage 2 coincidence probability will depend on the distribution of the types of glass in the population
- Probability of a coincidental match is high (i.e, the evidence is weak) when:
  - small difference between control sample and population of randomly chosen sources (i.e., control sample is "ordinary")
  - large amount of variability among the fragments in an individual source
  - large amount of heterogeneity among the potential sources in the population

Test yourself

Stage 1

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

Test yourself

Stage 1 - answer

- Applying the two-stage approach requires that we select a cutoff at the first stage to determine whether two samples are indistinguishable. Which of the following are true?
  - A high cutoff is best because it will make it difficult to eliminate a suspect - **This is not a helpful statement**
  - A low cutoff is best because it gives the benefit of the doubt to the suspect - **This is not a helpful statement**
  - We should choose a cutoff so that there no type I or type II errors - **This is impossible!**
  - The statistician should get to choose the cutoff - **This has to be a societal choice.**
  - Figuring out where to put the cutoff is a hard problem - **This is the best answer.**

Test yourself

Stage 2

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

Test yourself

Stage 2 - answer

- Which of the following statements about state 2 would you support?
  - Stage 2 is very important because we need to know how unusual it is to find indistinguishable samples – YES
  - Stage 2 is not very important because once you have found a match, that is all you need to know – NO
  - Stage 2 is difficult because the relevant population will vary from case to case – YES
  - Stage 2 is not necessary; it should be left to the jury – NO
  - I'm getting tired of this type of question – MAYBE

Statistical Inference and The Two-Stage Approach

Summary

- Statistical inference uses sample data to draw conclusions about a population
- Point estimation, interval estimation, hypothesis tests are main tools
- Critical that procedures account for variation that could be observed due to chance
- Statistical hypothesis tests can be useful but difficult to interpret at times
- Two-stage approach to forensic inference
  - First stage determines if the known and unknown samples appear to "match" or "be indistinguishable"
    - Relies on statistical tests (or intervals)
    - important to recognize the asymmetry in testing a null hypothesis
    - important to design a procedure with appropriate error rates
  - Second stage attempts to quantify the probability of a coincidental match
    - requires careful consideration of the relevant population
    - can be challenging to compute (no standard procedure)
    - this step is important but unfortunately sometimes omitted