Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

October/November 2021

Outline

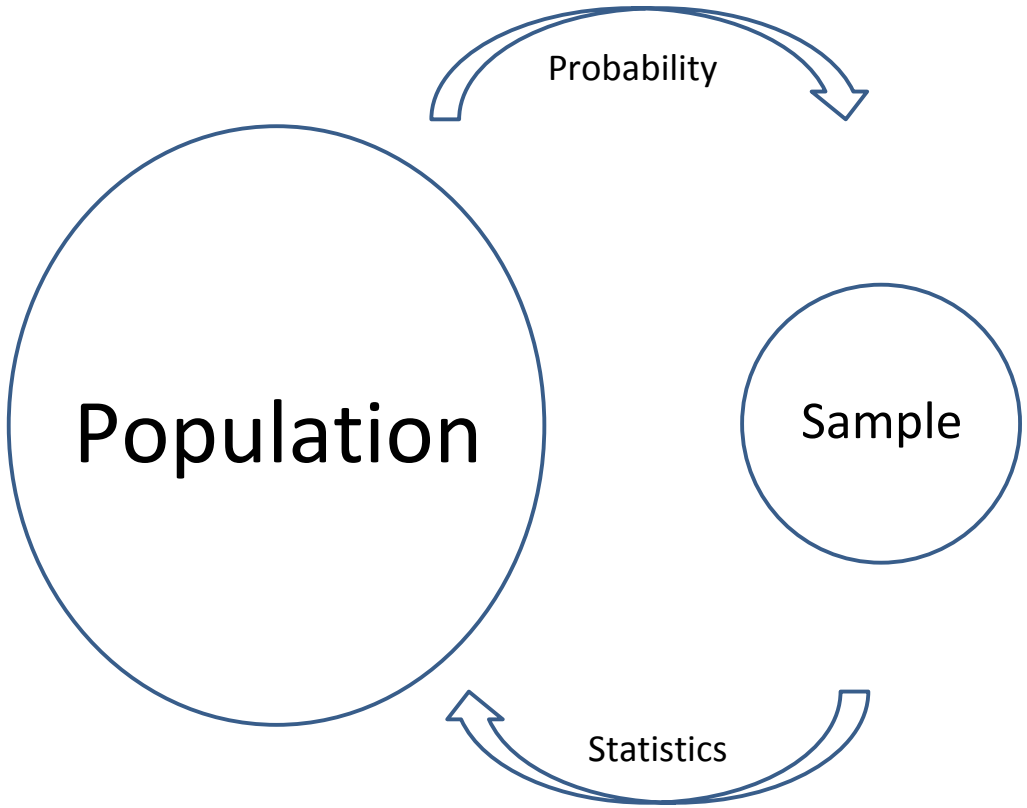
- Part 1 - Probability Concepts and Their Relevance to Forensic Science
  - review of probability concepts
  - conditional probability and independence
  - Bayes' Theorem and likelihood ratio
- **Part 2 - Sampling, Statistics and Forensics**
  - **collecting data**
  - **variability, reliability and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

Learning Objectives for Part 2

- Understand statistical concepts associated with collecting data (sampling and study design)
- Understand how probability and sampling can be used as a basic for inference about the population
- Understand principles of point estimation and interval estimation
- Understand how statistical tests work and their limitations

Probability and Statistical Inference

Recall "The Big Picture"



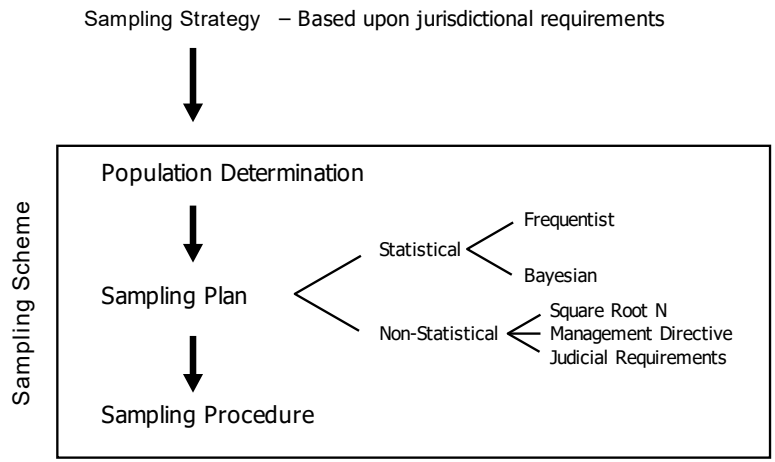
- Population = universe of objects of interest
  Sample = objects available for study
- Probability: population $\rightarrow$ sample (deductive)
- Statistics: sample $\rightarrow$ population (inductive)

Probability to Statistical Inference
Collecting data

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

Motivation - ASTM 2548-16

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



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

Motivation - ASTM 2548-16

- ASTM 2548-16: Standard Guide for Sampling Seized Drugs for Qualitative and Quantitative Analysis
  - Section 4.2.1. Sampling may be statistical or non-statistical
    - 4.2.1.1. In many cases, a non-statistical approach may suffice. The sampling plan shall provide an adequate basis for answering questions of applicable law. For example, Is there a drug present in the population?
    - 4.2.1.2. If an inference about the whole population is to be drawn from a sample, then the plan shall be either statistically based or have an appropriate statistical analysis completed and limits of the inference shall be documented.

Probability to Statistical Inference

Collecting data

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

Probability to Statistical Inference
Collecting data

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

Probability to Statistical Inference

Collecting data

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

Probability to Statistical Inference
Collecting data

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

Probability to Statistical Inference
Collecting data

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

Probability to Statistical Inference

Collecting data

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

Probability to Statistical Inference

Challenges of Causal Inference

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

Probability to Statistical Inference

Challenges of Causal Inference

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

Probability to Statistical Inference

Data collection and Study Design

- Study design has a large impact on the validity and relevance of results
- Key study design principles
  - compare treatments to a control (e.g., current practice)
  - randomly assign treatments to units
  - make sure sample size is large enough to draw reliable conclusions
  - make environment as realistic as possible
  - use blinding where possible to avoid bias
- Principles of good experimental design are relevant to forensic science
  - can use these ideas in evaluating process improvements in the lab
  - for black box studies these suggest integrating test cases with actual casework
  - a key issue in PCAST report (and the DOJ response)

Test yourself

Data collection and study design

- Black box study of packing tape comparisons – A sample of 50 volunteer forensic examiners who make packing tape comparisions are used in a black box study. Each is given 10 pairs of questioned/known pairs and asked to assess whether the questioned tape came from the same roll as the known sample.
  - Give one benefit of using volunteers in the study.
  - Give one disadvantage of using volunteers in the study.
- Some of the examiners do not complete all 10 assigned pairs.
  - The incomplete data might be a concern. Why?

Test yourself

Data collection and study design - answer

- Answer not provided in this file

Probability to Statistical Inference

Measurement, variability and uncertainty

- Once data have been collected that are relevant to the scientific question of interest, the focus shifts to measurement and analysis
- Motivation: ASTM 2927-16 Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using LA ICP-MS for Forensic Comparisons
  - Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition).**If the samples are distinguishable in any of these observed and measured properties, it may be concluded that they did not originate from the same source of broken glass. If the samples are indistinguishable in all of these observed and measured properties, the possibility that they originated from the same source of glass may not be eliminated.** The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass."

Probability to Statistical Inference

Measurement, variability and uncertainty

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

Probability to Statistical Inference

Measurement, variability and uncertainty

- ISO 1725: 7.6.1 **Laboratories shall identify the contributions to measurement uncertainty.** When evaluating measurement uncertainty, all contributions that are of significance, including those arising from sampling, shall be taken into account using appropriate methods of analysis.
- Key point: any measurement process involves some degree of uncertainty
- If you measure the same item multiple times you will not get exactly the same answer (e.g., Bush vs Gore recount)
- This reflects natural variability in the measurement process, environmental factors or other contributors
  (these are often referred to as "noise")
- A measure of the resulting uncertainty should be provided to the user

Probability to Statistical Inference

Measurement, variability and uncertainty

- Scientists focused on physical measurements often use uncertainty to refer to the intrinsic uncertainy in a measurement
  - a thermometer may only be accurate to within 0.1 degrees
- Statisticians tend to use uncertainty more broadly to address all kinds of things that we don't know
- Uncertainty is usually addressed with probability, a probability distribution, or some summary of a probability distribution
  - e.g., probability of rain tomorrow is 60%
  - e.g., weights on this scale are normally distributed with standard deviation 0.1 kg
  - e.g., measurement is accurate to ± 0.5 in

Probability to Statistical Inference

Measurement, variability and uncertainty

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

Probability to Statistical Inference

Measurement, variability and uncertainty

- Variability associated with a set of measurements is often described with quantities like the standard deviation, interquartile range or range
  - standard deviation = a measure of "typical" deviation from the mean (formally it is the square root of the average squared deviation from the mean)
  - range = maximum value - minimum value
  - interquartile range = 75%ile of the set - 25%ile of the set

Probability to Statistical Inference
Reliability

- Variability is related to the concept of reliability. Reliability plays a large role in ongoing discussions about forensic science
- Reliability refers to the consistency of a measurement or a measurement protocol, i.e., will we get the same answer if a process is repeated
- Variability and reliability are related concepts but not the same
  - Example - suppose we have an imprecise scale so each measurement is associated with considerable uncertainty (variability is large)
  - We may be able to get a reliable measurement by averaging a number of readings from the scale
  - So a single scale reading is not reliable but the average of a number of readings may be reliable

Probability to Statistical Inference
Reliability

- There are several aspects of reliability
  - **repeatability** refers to whether a measurement or decision would be the same in two instances using the same item and the same examiner. It is an intra-examiner assessment.
  - **reproducibility** refers to whether a measurement or decision would be the same in two instances using the same item and different examiners. It is an inter-examiner assessment.
- High reliability is required to have a valid/accurate procedure
- But high reliability itself does not guarantee a valid/accurate procedure

Probability to Statistical Inference
Reliability

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

Probability to Statistical Inference

Reliability in forensics - handwriting complexity

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

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

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

Test yourself

Reliability

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

Test yourself

Reliability - answer

- Answer not provided in this file

Probability to Statistical Inference

Summarizing data

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

A	B	AB	O
.42	.10	.04	.44

- For discrete data (e.g., CMS) we may summarize with a table or numerical summaries (mean, standard deviation)
  - Example: CMS measures from an automatic algorithm applied to non-matching bullet lands (Chu et al., For. Sci. Int., 2013)

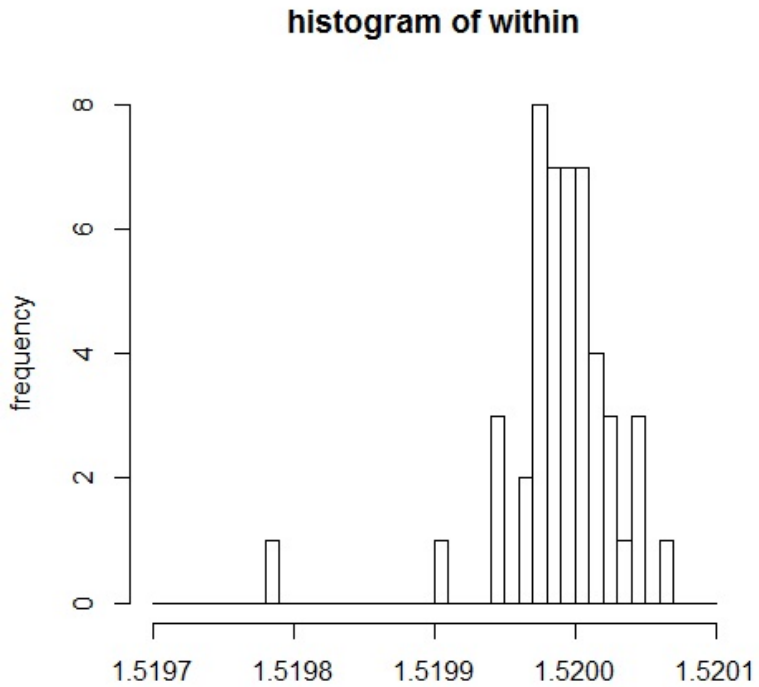
	0	1	2	3	4	5	Total
Count	3887	8219	782	70	2	0	12960
Proportion	.2999	.6342	.0603	.0054	.0002	.0000	1.0000

  - Can also compute mean (0.77) and standard deviation (0.58)

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



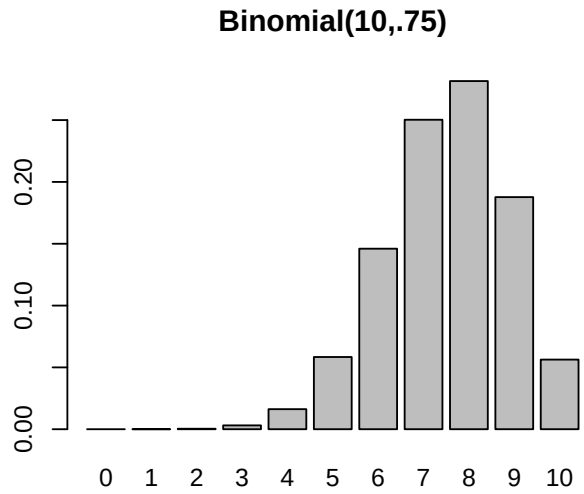
Probability to Statistical Inference

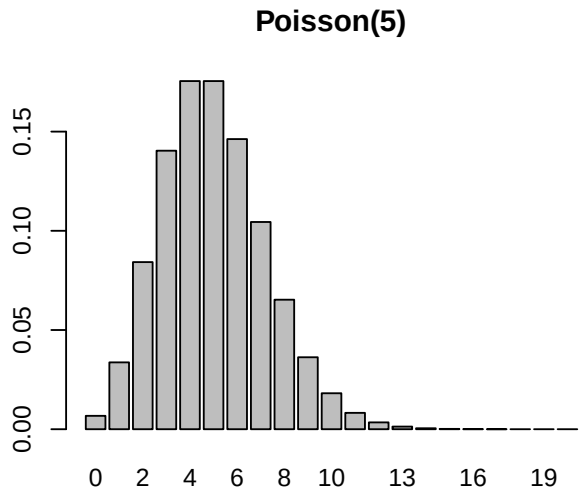
Probability distributions

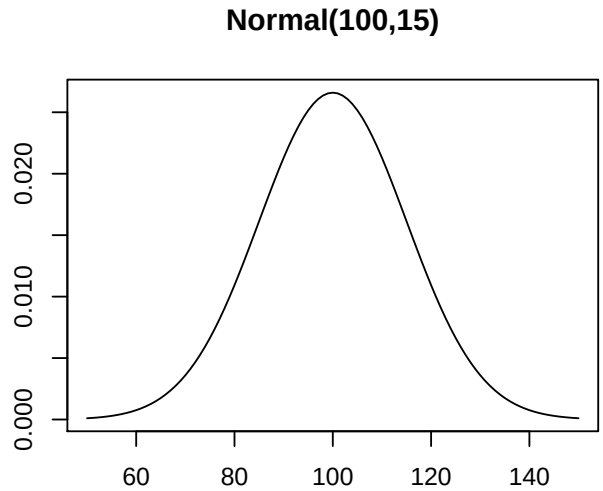
- Suppose we are to collect data on some characteristic for a sample of individuals or objects (weight, trace element concentration)
- The probability distribution is used to describe possible values and how likely each value is to occur
- Knowing about probability distributions is relevant for understanding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

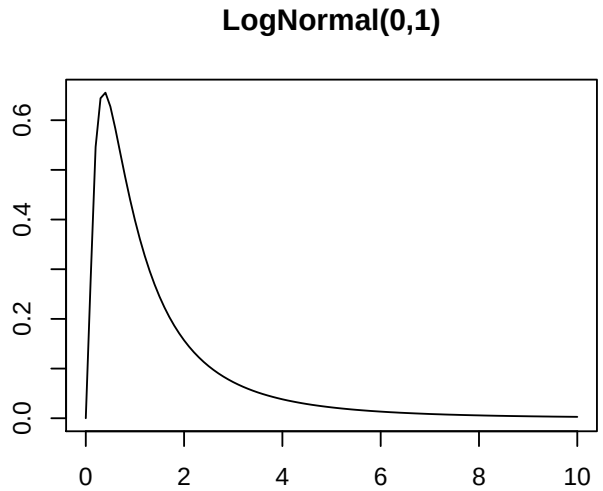
Probability to Statistical Inference

Probability distributions



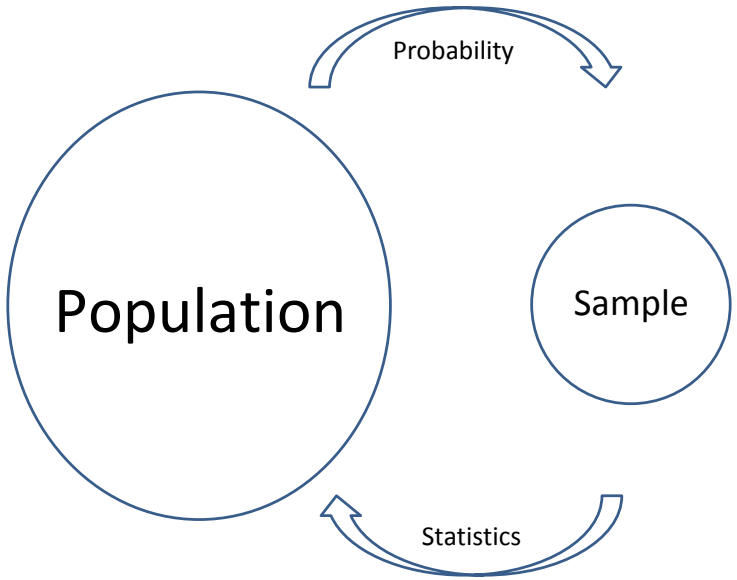






Statistical Inference

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

Motivation - ASTM 2927-16

- ASTM 2927-16: 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 2927-16

- 11. Calculation and Interpretation of Results

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

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

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

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

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

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

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

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

- This is a statistical inference procedure!

Statistical Inference
Background

- Definition - a **parameter** is a numerical characteristic of the population, e.g., a population mean
- Statistical methods are usually concerned with learning about population parameters from sample data
- **Note:** 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 workshop) 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

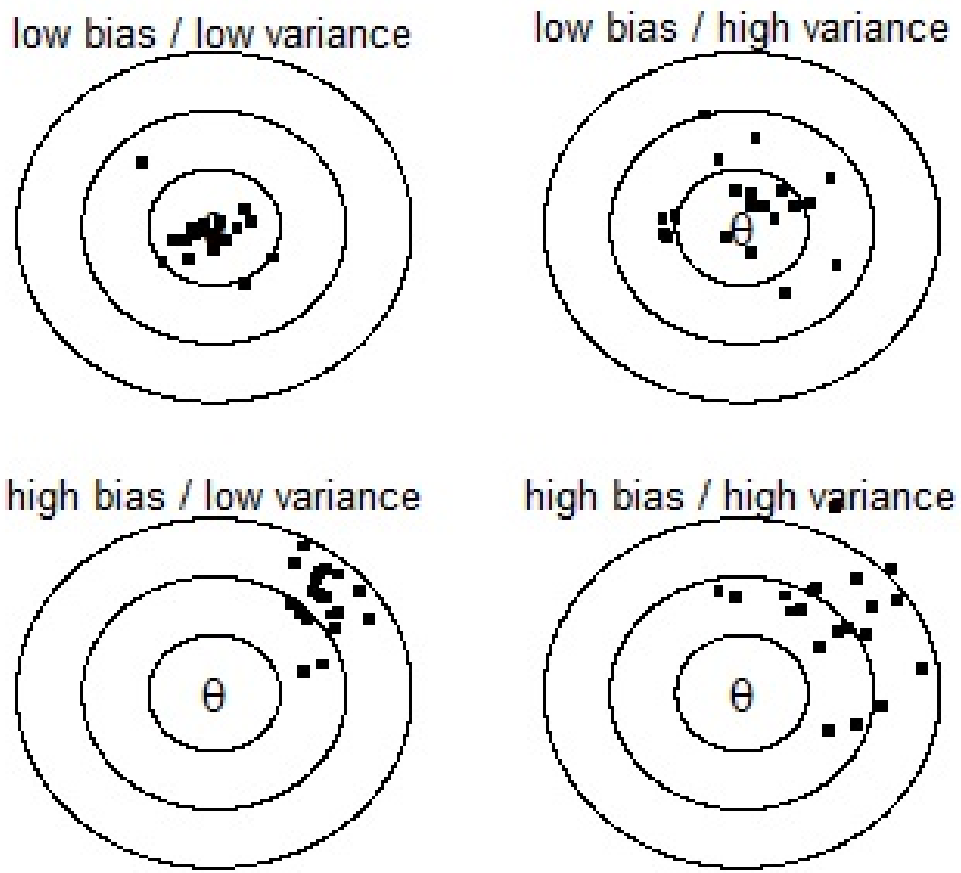
Point estimation

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

Statistical Inference

Performance of different estimators for unknown 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

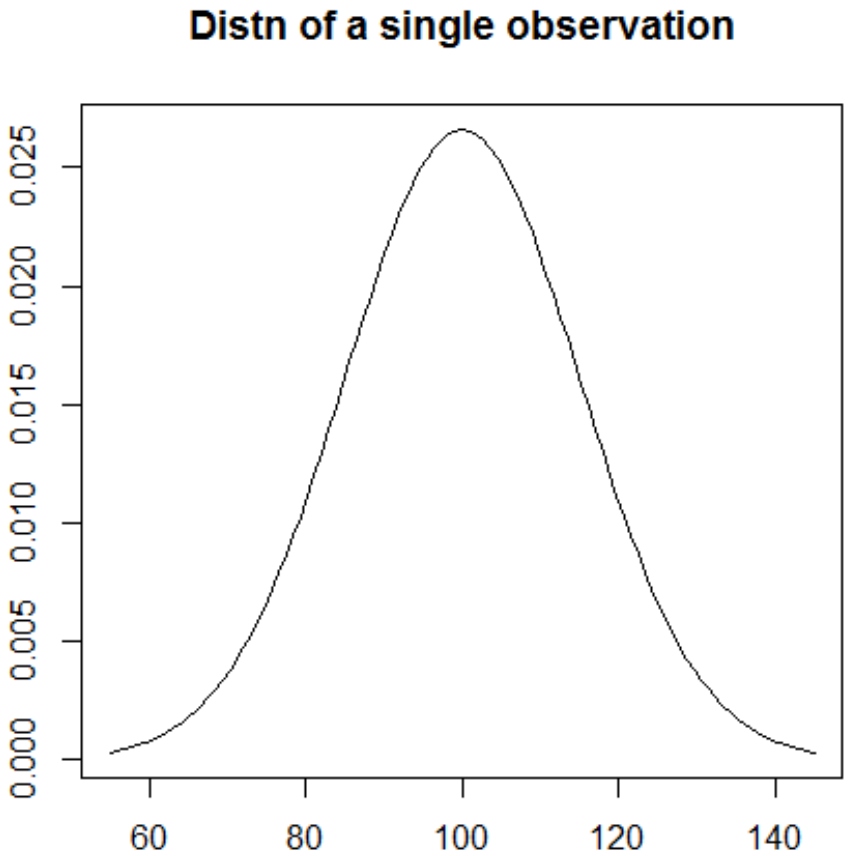
Standard errors

- 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
  - The standard deviation is a measure of the spread (variability) in a sample or in a population (describes uncertainty about a single observation)
  - When we look at a summary statistic (mean, median, percentile) it is also a random quantity (would give 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

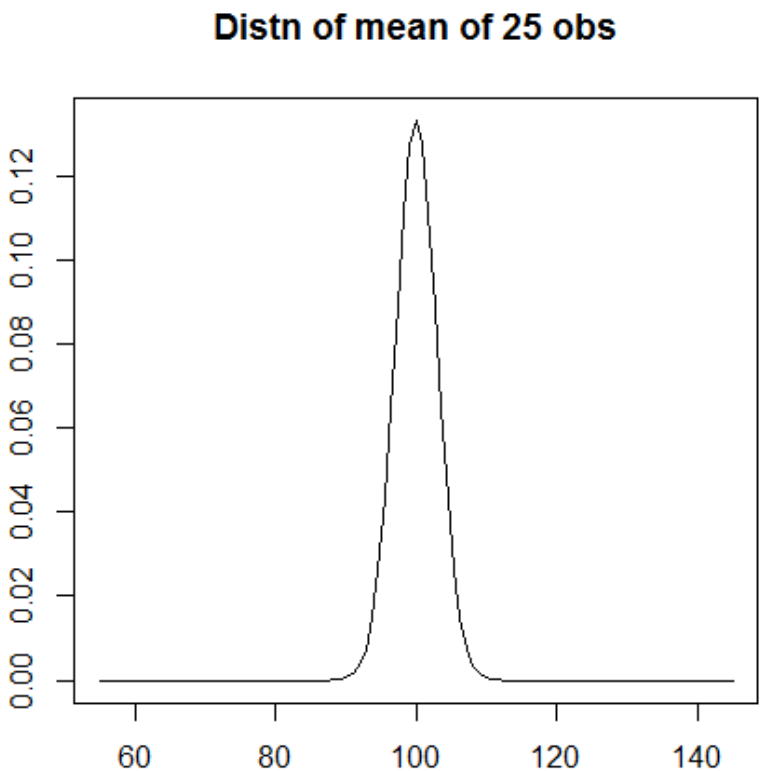
Standard errors

- 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 mean is 104.1 and the s.d. is 16.4
- Now suppose we take another random sample of 5 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 differences are natural. They just represent the variability that we might expect in different samples.
- We can study this variability.
- We can also take steps to reduce this variability (e.g., use bigger samples)

Statistical Inference

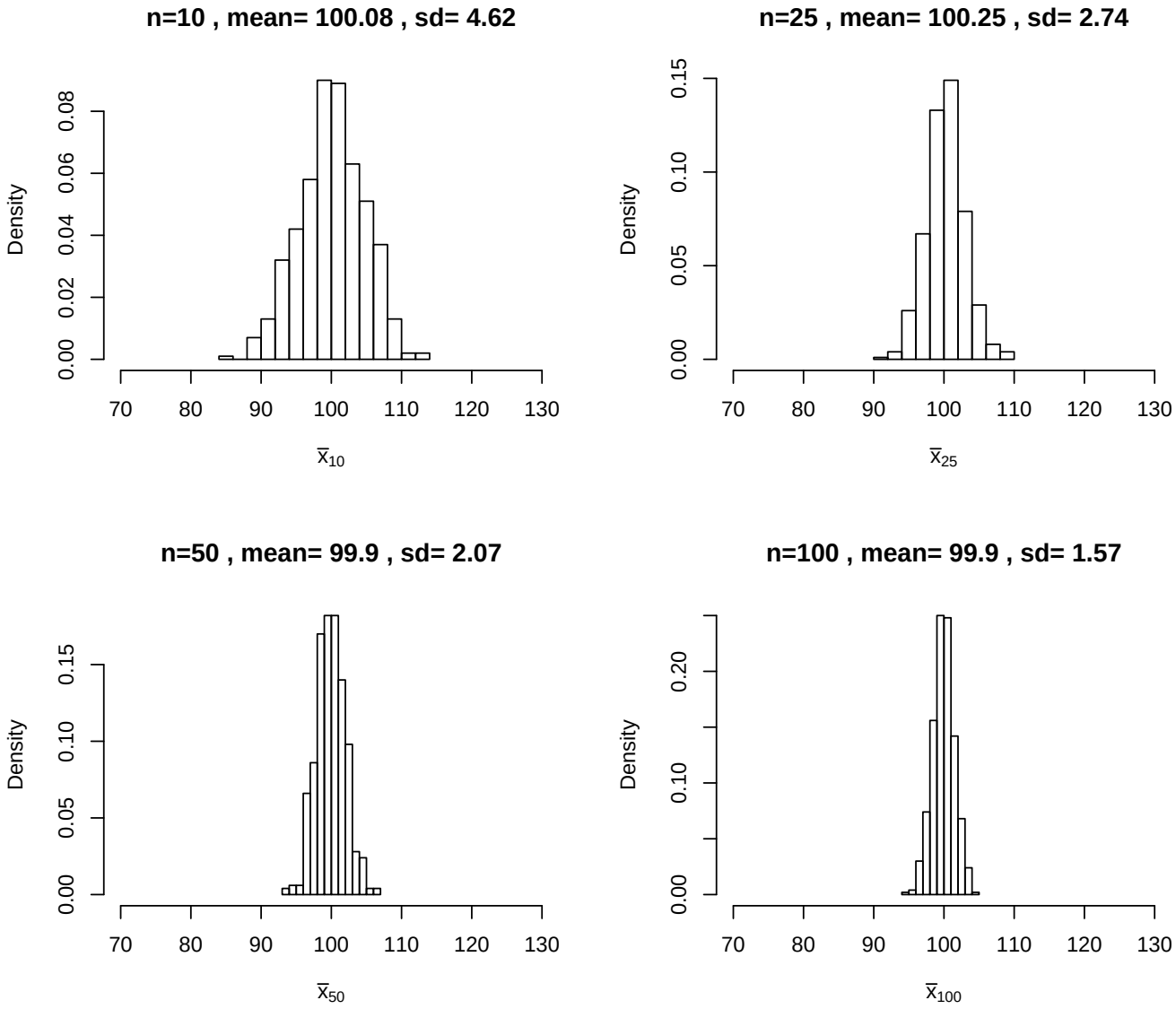
Standard errors

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



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

Statistical Inference

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
    (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. 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)
- In that case the null hypothesis is that the mean drop in BP is the same for the two drugs ("no change")
- The alternative hypothesis would be that the new drug leads to a bigger mean drop in BP than the current standard of care

Statistical Inference

Hypothesis testing

- A statistical test is usually summarized by a p-value measuring the strength of the statistical evidence against the null hypothesis (more on this later)
- Historically it has been common to use a threshold (say $p < .05$) to decide whether to accept or reject the null hypothesis
  - If we reject the null hypothesis then we say we have a statistically significant result
  - This approach has some problems and is currently a topic of much discussion
  - Because it is still popular in some quarters we spend some time on it now

Statistical Inference

Hypothesis testing

- Two types of errors 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 if 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
- Can be interpreted as the number of standard errors the sample estimate is from the hypothesized value under the null hypothesis
- 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

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

Statistical Inference

Hypothesis testing - comparing two means

- In practice, we are often interested in comparing two samples (or more precisely two populations)
- Assume random samples from each of the two populations are available
- Test for equivalence of parameters of the two populations
- Forensic example
  - suppose we have broken glass at a crime scene and glass fragments on the suspect
  - define 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

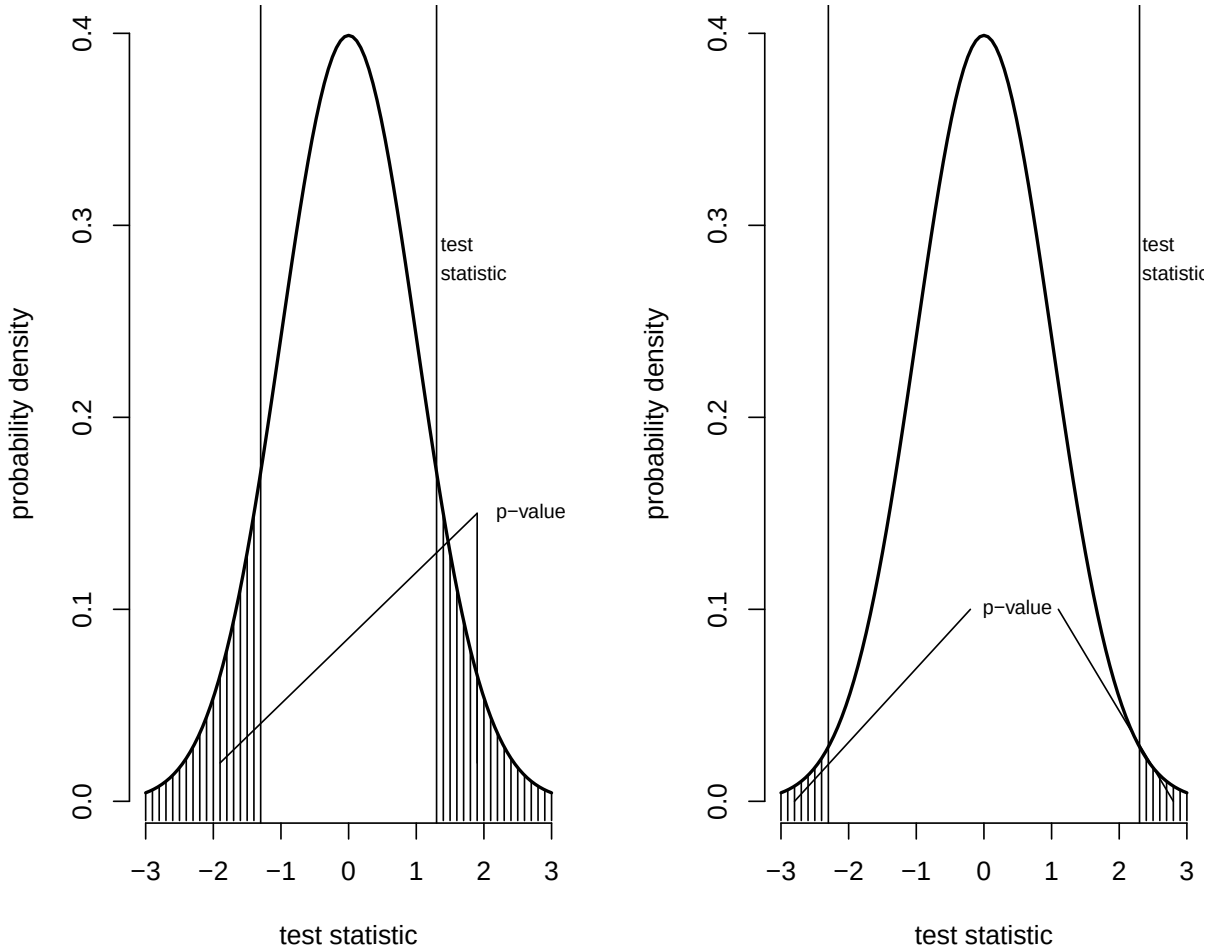
Hypothesis testing - comparing two means

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

Statistical Inference

Two examples of hypothesis testing results

- Left figure: observed test statistic = 1.3, p-value = 0.19
- Right figure: observed test statistic = 2.3, p-value = 0.02



Statistical Inference

Comparing two means - example

- Suppose 9 glass fragments are taken from glass at the scene (Y) and 10 fragments are found on the suspect (X).
  - $\bar{X} = 5.3, s.d. = 0.9, SE(\bar{X}) = 0.9/\sqrt{10} = .28$
  - $\bar{Y} = 5.9, s.d. = 0.85, SE(\bar{Y}) = 0.85/\sqrt{9} = .28$
  - observed difference is $\bar{Y} - \bar{X} = 0.6$
  - standard error for this difference is $SE(\text{diff } \bar{X} - \bar{Y}) = \sqrt{.28^2 + .28^2} = 0.4$
  - test statistic is $0.6/0.4 = 1.5$ which yields a p-value of 0.15
  - Do not reject the hypothesis that the two glass populations agree
- Interpretation is a key issue (can't reject the hypothesis of equal means and the possibility of a common source ... but this doesn't mean it is definitely true)

Statistical Inference

Hypothesis testing and forensic science

- Statistical hypothesis tests can be related to key 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?)
  - type II error is to decide innocent when person is guilty
- But it is not clear that statistical hypothesis tests are a good match to analysis of forensic evidence
  - The ASTM procedure is essentially a statistical hypothesis test
  - But there are logical problems
    - The null hypothesis (equal means) is incriminating which seems counter to the usual null hypothesis (no effect) logic
    - Accepting the null hypothesis says the two populations are indistinguishable but statistical test says we can't distinguish the means

Statistical Inference

Hypothesis tests and confidence intervals

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

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

- Answer not provided in this file

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

- Answer not provided in this file

Statistical Preliminaries

Some key takeaways for forensic practitioners

- 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
  - Intervals and tests play a significant role in analyses of some evidence types (discussed more later)
  - The likelihood ratio approach builds on the ideas in this section