Statistical Thinking for Forensic Practitioners

Course Outline

Part 1: Introduction

- General structure of a scientific method:
  1. Question of Interest
  2. Research
  3. Hypothesis
  4. Experiment
  5. Analysis & Interpretation
  6. Conclusion
- Uncertainty and error are inherent to science
- Share methods, data, code, results through publications; enabling replication and reproducibility
- Hypothesis = testable proposition, law = explanation of phenomenon that is abundantly supported by evidence
- Research methods:
  - Experimentation: explore relationships of two or more variables controlled by investigator
  - Description: systematic observation of some phenomenon
  - Comparison: observational study to determine association between variables
  - Modeling: Mimic nature and produce observations & predictions
- Variability is inherent to all scientific measurements
  - "Error" does not mean "mistake" in a scientific context. No measurement will exactly capture the desired true value.
- Daubert
  - Judge is obligated to screen testimony offered by experts.
  - Testimony must be "tied sufficiently" to the facts of the case
  - "Grounded in the methods and procedures of science."
    - Procedures have been tested, peer-reviewed, and published with understood error rates.
    - Generally accepted by the scientific community
- Population of interest.
  - Can be difficult (yet necessary) to define to evaluate the weight of evidence

  - How unlikely is a particular piece of evidence in the relevant population? Questions of rarity require data collected from relevant population (e.g., genotypes).

- Sampling
  - A random sample enables us to make statements about the population even though we've only observed a subset.

- Unethical practices:
  - Data snooping = formulating hypotheses after looking through data
  - p-hacking = test until you find something significant
  - Report only "good" results
  - Carry out scientifically "iffy" study
  - Overstate significance of findings

Part 2: Probability

- For an event E, E^c is its complement (the event "not E") and

$$P(E^c) = 1 - P(E)$$

- Odds ratio:
  - Odds in favor of E is

$$O_f = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

  - Odds against E is

$$O_a = \frac{P(E^c)}{P(E)} = \frac{1 - P(E)}{P(E)}$$

  - Given odds against E, O_a,

$$P(E) = \frac{1}{O_a + 1}$$

- Sample space is the list (or set) of all possible outcomes.
- Conditional Probability:
  - Definition of the conditional probability of A given B:

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

  - Interpret conditioning as shrinking the population to which we refer
    - If A = "suspect is guilty" and B = "suspect is bald," then, e.g., $P(A|B)$ is the probability that suspect is guilty given they are bald. Thus, we are shrinking the population of interest by not including non-bald individuals.
  - Above definition is equivalent to (by algebra) $P(A \text{ and } B) = P(B|A)P(A) = P(A|B)P(B)$.
  - A and B are independent if $P(A \text{ and } B) = P(A)P(B)$. Equivalently, if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.
  - The "Law of Total Probability:"

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

- Bayes' Theorem:
  - The conditional probability of A given B is equal to

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

  - $P(A)$ captures the "prior" information (i.e., prevalence) about the event A. Consider example:
    - Suppose G = "gunshot residue is present on suspect" and T = "diagnostic test for gunshot residue is positive." We want to know $P(G|T)$.
    - Assume $P(G) = .05$ (prior/prevalence) and $P(T|G) = .98$ (sensitivity/true positive rate) while $P(\text{not } T|\text{not } G) = .96$ (specificity/true negative rate).

* Then

$$P(G|T) = \frac{P(T|G)P(G)}{P(T)} = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|G^c)P(G^c)} = \frac{.98 * .05}{.98 * .05 + (1 - .96)(1 - .05)} = .56.$$

* Takeaway: Prevalence, $P(G)$, has a major impact on probability a positive test is a true positive.
* $P(G)$ would need to be determined based on prevalence of gunpowder in relevant population - requires previously collected (i.e., prior) data.

– In courtroom, commonly consider events E = evidence, H_s = "same source" proposition, and H_d = "different source" proposition.

* Bayes Theorem gives structure to learn about $P(H_s|E)$ based on $P(E|H_s)$ and $P(H_s)$ (or H_d).
* "Odds Form" of Bayes' Theorem:

$$\frac{P(H_s|E)}{P(H_d|E)} = \frac{P(E|H_s)P(H_s)}{P(E|H_d)P(H_d)}$$

  · Pits the competing hypotheses against each other.
  · Larger values provide more evidence towards the same-source hypothesis
* Important difference between the posterior odds, $\frac{P(H_s|E)}{P(H_d|E)}$, and likelihood ratio, $\frac{P(E|H_s)}{P(E|H_d)}$.
  · Prior odds $\frac{P(H_s)}{P(H_d)} \neq 1$ under presumption of innocence (in general, H_d given greater weight *a priori*).

Part 3: Data Collection

- Simple Random Sampling:
  - Every sample of size n drawn from the population has the same probability of selection
  - Appropriate if the population is relatively "homogeneous" (e.g., demographics are approximately evenly represented)
  - If population is of size N, then there are N^n total possible samples.
  - If sampling with replacement, then each sample has a $\frac{1}{N^n}$ probability of being selected.
  - If sampling without replacement, then there are

    $$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

    possible samples of size n.
- Stratified Sampling:
  - Used when population contains different types of items (e.g., shoe brands) and some types are more abundant than others (e.g., more Nike than New Balance).
  - Collect a random sample from each group (or stratum) of the population
  - Items no longer have equal probability of being selected
- Cluster Sampling:
  - Also used when population is heterogeneous.
  - Involves separating population into clusters, randomly selecting a cluster, and sampling all items within that cluster. Examples include:
    - Surveying all individuals living on a randomly selected city block
    - Randomly select a shipping container and search entire container for contraband
- A sample is representative if its composition mirrors that of the population from which it is drawn.
  - May want a non-representative sample if desire is to learn more about a minority group in the population (which is why stratified sampling is used)
- Non-random sampling:
  - Random sampling may not be practical, ethical, etc.
  - If a sample is non-random, we may not know the size of the population from which it is taken or the probability that an element was selected (making it difficult to assume an underlying probability distribution and make population inferences)
  - Convenience sampling involves selecting elements of the population that are easily available. This can be mathematically proven to result in biased results.
  - Snowball/network sampling involves getting selected members of a population to recruit contacts to continue building the sample
  - Purposive sampling involves selecting elements based on some criterion
  - You must know how a sample was collected before deciding whether the data are relevant to your question.
- Possible sampling issues:
  - Incomplete coverage: sample represents only a portion of the population
  - Self-selection bias: participants self-select (there may be some underlying commonality between self-selecting participants that make then non-representative of the overall population)

  - Non-response bias: participants decide not to respond to (or lie about) some or all of a survey
    - E.g., may refuse to answer a question regarding their habits/behaviors if they are embarrassed/ashamed. This would be an example of a non-ignorable non-response, which can severely bias results

- Estimating population size using "marked recapture"
  1. Collect a random of items, some of which are of interest (e.g., unregistered firearms)
  2. Mark items of interest - say M are marked
  3. Collect a second sample of size C and observe R items that were marked from original sample
  4. Estimated number of items of interest in population is $N = \frac{M \times C}{R}$. Derived from proportion $\frac{M}{N} = \frac{R}{C}$ ("M is to N as R is to C").

- Observational Studies:
  - Select a subset of a population and researcher observes a set of attributes
  - Used if experimentation is unethical (e.g., unethical to "assign" (i.e., force) a group of patients to smoke and others to not to observe effects of smoking)
  - Can establish an association between two variables (e.g., smoking and lung cancer), but cannot draw a causal relationship

- Randomized controlled studies:
  - Researcher randomly assigns participants to a treatment
  - Randomization ensures differences between participants get evened out across different treatments
  - By using randomization, the only difference between participants in different treatment groups is the treatment itself.
  - If a measured outcome is different for the two treatment groups, we can say that the treatment must have been the cause of the difference since all other differences were evened out.

- Blinding:
  - Blinded experiment means participant is unaware of the treatment group to which they are assigned (e.g., to avoid placebo effect)
  - Double-blinded experiments means both participant and researcher are unaware of treatment group assignments (e.g., to avoid treatment bias from the researcher)

- Databases:
  - Smaller data sets may be used for testing new methods, etc.
  - Conclusions from applying a method to a small data set should not, in general, be extrapolated to a larger population
  - Reference data bases are larger and are built to include all possible variants of an item of interest (e.g., all shoe brands/models sold in the US)
  - Reference data bases allow us to discuss the *probative* value of a piece of evidence
    - E.g., suppose markings on two Nike Air Max 270s are found to be indistinguishable. Before making a conclusion about whether this means the two shoes form a matching pair, we need to know how likely it is for *any* two Nike Air Max 270s to have indistinguishable markings. If it is relatively common for two Nike Air Max 270s to have indistinguishable markings, then the evidence does not have a high *probative* value.
  - To construct a reference data base, the probability of selecting each sample unit should be known ahead of time (so that population frequencies can be estimated).

Part 4: Types of Data

- Qualitative/Categorical data:
  - Observations can be placed into distinct categories
  - Nominal categorical variables have no meaningful order (classified into named categories)
    - E.g., blood type (A, B, AB, O) have no meaningful order
  - Ordinal categorical variables may have meaningful order
    - E.g., grades received in a course (A+, A, A-, B+, etc.) have a meaningful order
  - Summarizing categorical data:
    - Counts and/or proportions
    - Pie or bar chart (showing counts/proportions)
- Quantitative data:
  - Observations take on numeric values and arithmetic operations (adding, averaging, etc.) are meaningful.
  - Continuous data take on any numeric value in a range (e.g., glass refractive index can be any non-negative number)
  - Discrete data take on one of a set of values (e.g., number of consecutively matching striae between two land engraved areas can be 0, 1, 2, etc.)
  - Summarizing quantitative data:
    - Bar chart (for discrete data) or histogram (for continuous data)
    - Describe the "distribution" of quantitative data using:
      - Modality: uniform, unimodal, bimodal, multimodal
      - Shape: symmetric, left-skewed, right-skewed (any of these 3 with outliers)
      - Center: mean, median, or mode
      - Location: 5 number summary (minimum, 1st quartile, median, 3rd quartile, maximum)
      - Spread: Range, Interquartile Range (IQR), Variance/Standard Deviation
    - Mean and Standard Deviation/Variance are sensitive to skewed distributions/outliers
    - Rules of thumb:
      - Mean and Standard Deviation/Variance typically used for symmetric data without outliers
      - Median and IQR more appropriate for skewed data/data with outliers
    - Boxplots visualize 5 number summary
    - Time series visualize data obtained over time (e.g., temperature)
    - Scatterplots visualize relationship between 2 variables. Described by:
      - Direction: positive/negative/no association (as one variable increases, how does other variable change?)
      - Form: linear/non-linear/no association
      - Strength: none/weak/moderate/strong
      - Outliers: "extreme" data (typically in the y-direction)
- Data visualizations can easily be misinterpreted (or intentionally created to deceive)

Part 5: Probability Models and Uncertainty

- We can express our uncertainty about an event (including uncertainty in measurement, prediction, etc.) using probability.
- Variability refers to variation observed in repeated measurements.
  - E.g., measuring a single glass shard for refractive index 3 times and obtaining 3 different values
  - Variability is inherent in even the most robust measuring procedures
  - We aim for variability between measurements to be small (the more expensive the machine, the less variable measurements should be in theory), but it can never be removed outright
  - Probability can be used to mathematically account for variability (e.g., "there is an 80% chance the true refractive index of this shard of glass is between x and y")
  - Often quantified via measures of spread (variance/standard deviation, IQR, range)
  - We can characterize variability in terms of:
    - Reliability: how consistently a method measures some information
    - Repeatability: how consistently a method measures a quantity under the same environmental conditions (same operator, measuring device, etc.)
    - Reproducibility: how consistently a method measures a quantity across different environmental conditions (different operator, etc.)
  - Reliability consists of a method's Repeatability and Reproducibility
  - Practical considerations:
    - How reliable are measurement tools when used by same/different operators?
    - How consistent are evaluations of the same evidence/object by the same/multiple examiners?
- Probability distributions:
  - We use probability to express our uncertainty in the outcome of an event.
    - Probability models provide us with mathematical tools for expressing an assumed structure on a random outcome (e.g., tomorrow's temperature in Houston will be between 0° and 100° Fahrenheit with 99% probability).
    - We often assume that a random outcome behaves according to a particular probability distribution. Certain probability distributions are used so often that they are given names.
  - Probability distributions consist of two parts: a range/set of possible values and a description of how likely any one of these values is to occur.
    - This "description" often takes the form of a mathematical function. For example, $P(\text{coin flip is heads}) = 0.5 = P(\text{coin flip is tails})$ is called a "probability mass function."
  - Probability distributions are often characterized by a single (or set of) *parameter(s)* that affect how the probability is distributed among the range/set of possible values.
    - Parameters are numerical characteristics of the population
    - E.g., suppose we wanted to know whether a coin were fair. Then the parameter of interest might be the probability that the coin lands on heads. If the coin were fair, then this probability should be 0.5.
  - The *sample space* (denoted Ω or sometimes S) is the set of all possible values. The notation $E \in \Omega$ means that the event E is an element of the sample space.
  - Any probability distribution, denote it by P, must satisfy:
    1. $0 \leq P(E) \leq 1$ for any $E \in \Omega$. Recall that $P(E)$ means the "probability that event E occurs."
    2. $\sum_{A \in \Omega} P(A) = 1$.
- Common discrete probability distributions:

- Binomial distribution
  - Often used to describe the number of "successes," X say, out of n independent, binary trials.
    - E.g., a single coin flip would mean $n = 1$. A "success" is often (arbitrarily) chosen to be a heads.
    - E.g., If $n = 10$ bags of suspected contraband are tested, a "success" might be the number of bags found to contain contraband.
  - The binomial distribution is most often characterized by two numbers: the number of trials, n, and the probability of success in a single trial, p.
  - Notation used to follow represent that a random variable, X say, follows a binomial distribution is $X \sim Binomial(n, p)$.
  - If $X \sim Binomial(n, p)$, then the probability mass function is given by

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, 2, ..., n \\ 0 & \text{otherwise.} \end{cases}$$

- Hypergeometric distribution
  - Often used to describe the number of successes out of n trials without replacement our of a population of N objects, of which K objects have a feature of interest.
    - E.g., A jar contains N marbles; K of which are black and $N - K$ are white. We are interested in black marbles. If n marbles are drawn, then X is the number of black marbles.
  - If $X \sim Hypergeometric(N, K, n)$, then the probability mass function is given by

$$P(X = k) = \begin{cases} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} & \text{if } k = 0, 1, ..., K \\ 0 & \text{otherwise.} \end{cases}$$

- Poisson distributions
  - Often used to describe the number of events, X, occurring in an interval of time/space.
    - E.g., number of consecutive matching striae on two land engraved areas.
  - Parameter λ is the average number of events in the defined interval (of space/time)
  - If $X \sim Poisson(\lambda)$, then the probability mass function is given by

$$P(X = k) = \begin{cases} \frac{\lambda^k \exp(-\lambda)}{k!} & \text{if } k = 0, 1, 2, ... \\ 0 & \text{otherwise.} \end{cases}$$

- Common continuous probability distributions:
  - Normal distribution
    - Commonly used to model data that are symmetric and unimodal. Also describes the probabilistic behavior of a mean as the sample size increases to infinity (a result called the Central Limit Theorem).
      - Heights, weights, blood pressure
    - Parameter μ is the mean (average) of the distribution. Parameter σ^2 is the variance of the distribution.
    - If $X \sim Normal(\mu, \sigma^2)$, then the probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

    for any $x \in \mathbb{R}$.
  - Log-Normal distribution
    - Often used to model skewed data (specifically, right-skewed, non-negative data). The natural logarithm of Log-Normal-distributed observations follow a normal distribution.
      - E.g., elemental concentration of Aluminium (or really any element) might be modeled using a Log-Normal distribution.

* Due to the relationship with the normal distribution, this distribution is also parameterized by μ, σ^2, although their interpretations are not the same as with the normal distribution.
* If $X \sim Log - Normal(\mu, \sigma^2)$, then the probability density function is given by

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\ln(x) - \mu)^2\right) & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- More on parameters:
  - We collect data (a sample) of size n, call it $(x_1, x_2, ..., x_n)$, to learn about the parameters of interest.
    * E.g., we might flip a coin a large number of times and record the outcome of each flip (so perhaps x_1 = heads, x_2 = heads, x_3 = tails, etc.). We can use this information to decide whether the coin is fair.
  - To understand how an observation governed by a particular probability distribution behaves, it is often useful to know:
    1. The center: the average value that we can anticipate the observation to take on. This is described by the *expected value* (equivalently, the mean or the expectation) of the distribution.
    2. The spread: how variable the observation will be around the expected value. This is described by the *variance* of the distribution.
  - In most situations, we do not know the true value of the expectation, the variance, or of some other parameter of interest.
    * E.g., it's difficult to know whether a random coin is *truly* fair. In this case the true probability of heads, p say, is unknown.
  - The *expected value* of a random variable X, denoted $E(X)$ is defined by

$$E(X) = \begin{cases} \sum_{k \in \Omega} kP(X = k) & \text{if } X \text{ is discrete} \\ \int_\Omega xf(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

    * Interpret this as the weighted average of all possible values of X where the weights are the respective probabilities.
    * The expected values for the distributions described above are:

Distribution	$E(X)$
$Binomial(n, p)$	$n * p$
$Hypergeometric(N, K, n)$	$n * \frac{K}{N}$
$Poisson(\lambda)$	λ
$Normal(\mu, \sigma^2)$	μ
$Log - Normal(\mu, \sigma^2)$	$\exp\left(\mu + \frac{1}{2}\sigma^2\right)$

Table 1: Expected Values of Common Distributions

  - The *variance* of a random variable X, denoted $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E\left\{[X - E(X)]^2\right\} = \begin{cases} \sum_{k \in \Omega}[k - E(X)]^2 P(X = k) & \text{if } X \text{ is discrete} \\ \int_\Omega [x - E(X)]^2 f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

    * Interpret this as the average squared distance from the mean
    * It is a fact that $\text{Var}(aX) = a^2\text{Var}(X)$ for $a \in \mathbb{R}$.
    * The variances for the distributions described above are:

Distribution	Var(X)
$Binomial(n,p)$	$n*p*(1-p)$
$Hypergeometric(N,K,n)$	$n*\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$
$Poisson(\lambda)$	λ
$Normal(\mu,\sigma^2)$	σ^2
$Log-Normal(\mu,\sigma^2)$	$\exp\left(2\mu+\sigma^2\right)\left(\exp(\sigma^2)-1\right)$

Table 2: Variances of Common Distributions

- Accuracy vs. Precision
  - Accuracy: whether observed values are, on average, close to the true parameter value
    - Observed values that are systematically far from the true value are called *biased*.
    - Bias is often inherent to a non-representative sample. May need to change sampling scheme used.
    - Biased estimates may also occur due to an incorrectly specified probability distribution. Consider alternative probability distributions.
  - Precision: how closely distributed values are to one another.
    - Highly precise values means that they have low variance.
    - To increase precision (specifically of parameter estimates), either take a larger sample or use a more precise measurement method/machine.
  - The *covariance* between two random variables X and Y is defined as
$$Cov(X,Y) = E\left\{[X-E(X)][Y-E(Y)]\right\}.$$
    - Measures the linear association between variables X and Y.
    - Positive/negative/0 covariance indicates positive/negative/no linear association between X and Y, respectively.
    - It is a fact that
      1. $\text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2abCov(X,Y)$ for $a,b \in \mathbb{R}$
      2. $\text{Cov}(X,Y) = 0$ if X and Y are independent.
    - E.g., Cadmium may occur in higher elemental concentrations along with Aluminium.
  - The *correlation* between two variables X and Y is defined as
$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} \tag{1}$$
    - Interpreted similar to covariance, but bounded between -1 and 1.
    - Correlation close to 1/-1 means that two variables have a strong positive/negative association, respectively (think in terms of scatterplots)
    - It is a fact that $Corr(X,Y) = 0$ if X and Y are independent.
- Propagation of Error
  - For two variables with non-zero covariance, recall that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2Cov(X,Y)$.
    - If $Cov(X,Y) > 0$, then the combined variance of $X+Y$ is actually larger than the individual (marginal) variances of X and Y.
    - In practical terms: adding two measurements together that each have an associated degree of uncertainty may inflate the overall uncertainty if these two measurements are positively associated.
  - If dividing two random variables, then
$$\text{Var}\left(\frac{X}{Y}\right) = \left(\frac{E(X)}{E(Y)}\right)^2\left(\frac{\text{Var}(X)}{E(X)} + \frac{\text{Var}(Y)}{E(Y)}\right)$$
  - Slide 50 contains approximate variances for other transformations.

Part 6: Inference

- Statistical methods involve learning about population parameters based on collected samples
  - E.g., calculate a sample mean, $\bar{x}$. If our sample is "good," then we would expect $\bar{x}$ to be close to the population mean, μ. The laws of probability (and our assumed probability distribution) tell us how close we can expect $\bar{x}$ to, in fact, be from μ.
- Common population parameters of interest and their associated estimators include:

Pop. Parameter	Sample Estimator	Formula
Mean μ	$\bar{x}$	$\left(\sum_{i=1}^{n} x_i\right)/n$
Variance σ^2	S^2	$\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)/(n-1)$
Standard deviation	S	$\sqrt{S^2}$
Proportion π	p	Proportion of "successes" in sample

  - Alternatively, we may denote the sample estimators using "hats" - e.g., $\hat{\mu}, \hat{\sigma}^2, \hat{\pi}$, etc.
- Point Estimators
  - An *estimator* is a rule for estimating a population parameter. An *estimate* is the resulting value.
  - A *sampling distribution* is the probability distribution governing the behavior of a sample statistic.
    - E.g., for a sample $(x_1, x_2, ..., x_n)$ with expectation μ and variance σ^2, the sample mean $\bar{x}$ also has expectation μ and variance $\frac{\sigma^2}{n}$.
    - If the sample is drawn from a $Normal(\mu, \sigma^2)$, then $\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$.
  - The *standard error* (SE) of an estimator is its estimated standard deviation.
    - Standard error quantifies the uncertainty about a point estimate.
    - E.g., the standard error of the sample mean is

$$SE = \sqrt{\frac{S^2}{n}}$$

      where S^2 is the sample variance.
  - The *Central Limit Theorem* says that for a sufficiently large sample drawn from a distribution with mean μ and variance σ^2, the distribution of $\bar{X}$ will be normal with mean μ and variance $\frac{\sigma^2}{n}$ even if the distribution from which the sample was drawn is not normal.
    - The CLT says that the sample mean is *unbiased*, meaning $E(\bar{X}) = \mu$.
    - It also says that the sample mean is *consistent*, meaning as n increases, the sampling distribution of $\bar{X}$ gets more concentrated around μ. Equivalently, $\frac{\sigma^2}{n}$ shrinks.
- Interval Estimators
  - Confidence intervals provide a range of plausible values for the parameter of interest.
  - Confidence intervals are all of the general form:

$$\text{(point estimate)} \pm \text{(critical calue)} * SE(\text{point estimate})$$

    - The quantity to the right of the $\pm$ sign is called the *margin of error* (ME)
    - All else staying the same, a higher confidence level results in a wider confidence interval. Wider intervals are less precise.
    - The confidence interval is dependent on the sample used to construct it. Since the sample is random, a confidence interval is random too.

* The critical value is dependent on 2 factors: the desired confidence level and whether the population variance is known.
  - Confidence level will be of the form $(1-\alpha)*100\%$ for some α (e.g., 95% confidence means $\alpha = .025$).
  - If the test concerns population proportions or if the population variance is explicitly given, then assume it is known. In this case, use a standard normal $z_{1-\alpha/2}$ quantile as a critical value (`NORM.S.INV` in Excel).
  - In most problems involving population means, the population variance is unknown. In this case, use a $t_{1-\alpha/2,d.f.}$ critical value, which is the $(1-\alpha/2)$-th quantile of a t-distribution with $d.f.$ degrees of freedom (e.g., $d.f. = n-1$ for a test involving a single population mean). This can be calculated using the `T.INV` function in Excel.

– A $(1-\alpha)*100\%$ confidence interval for the population mean, μ, is of the form

$$\bar{x} \pm (\text{critical value}) * SE(\bar{x})$$

  * If σ^2 is assumed known, then $SE(\bar{x}) = \frac{\sigma^2}{n}$. Otherwise, $SE(\bar{x}) = \frac{S^2}{n}$.

– A $(1-\alpha)*100\%$ confidence interval for the difference between population means, $\mu_x - \mu_y$ say, is

$$(\bar{x} - \bar{y}) \pm (\text{critical value}) * SE(\bar{x} - \bar{y})$$

  * If population variances are assumed known, then $SE(\bar{x}-\bar{y}) = \sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}$. Otherwise, $SE(\bar{x}-\bar{y}) = \sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}$ where n_1, n_2 are the two samples' sizes.
  * Degrees of freedom here are $n_1 + n_2 - 2$.

– A $(1-\alpha)*100\%$ confidence interval for a population proportion π is of the form

$$\hat{p} \pm (\text{critical value}) * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the proportion of successes. If Y is the total number of successes in the sample, then $\hat{p} = \frac{Y}{n}$.

– A $(1-\alpha)*100\%$ confidence interval under a hypothesized population proportion, e.g., if testing $H_0 : \pi = \pi_0$, is of the form

$$\hat{p} \pm (\text{critical value}) * \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

where $\hat{p}$ is the proportion of successes. If Y is the total number of successes in the sample, then $\hat{p} = \frac{Y}{n}$.

- Hypothesis Testing procedure
  1. Formulate 2 hypotheses
     * Null hypothesis, H_0, represents a proposition we wish to test the validity of.
       - E.g., $H_0 : \mu = \mu_0$ where μ_0 is a specific value.
     * The alternative hypothesis, H_a or H_1, challenges the null. We determine to what degree the alternative is supported over the null using a collected sample.
       - E.g., $H_a : \mu \neq \mu_0$ or $H_a : \mu < \mu_0$ or $H_a : \mu > \mu_0$ depending on the goals of the study
  2. Collect data and calculate relevant statistic
     * E.g., the sample mean and its standard error if testing the population mean.
  3. Calculate the "distance" between the sample statistic and the hypothesized parameter value
     * The "distance" is commonly quantified using a test statistic of the following form:

$$\frac{\text{point estimate} - \text{null hypothesized value}}{\text{SE of point estimate}}.$$

* For testing a hypothesized population mean value, $H_0 : \mu = \mu_0$ say:

$$\frac{\bar{x} - \mu_0}{SE_{\bar{x}}}.$$

  · If σ^2 is assumed known, then $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Otherwise, $SE_{\bar{x}} = \frac{S}{\sqrt{n}}$.

* For testing equality between two population means, $\mu_x = \mu_y$ say (with unknown population variances):

$$\frac{\bar{x} - \bar{y} - 0}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}}$$

  where $(\bar{x}, \bar{y})$, (S_x^2, S_y^2), and (n_1, n_2) are the sample means, variances, and sizes, respectively, of the two samples. In this case, degrees of freedom are $n_1 + n_2 - 1$.

* For testing a hypothesized population proportion value, $H_0 : \pi = \pi_0$ say:

$$\frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

* For testing equality between two population proportions, $H_0 : \pi_1 = \pi_2$ say:

$$\frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  where $\hat{p}_1 = \frac{Y_1}{n_1}, \hat{p}_2 = \frac{Y_2}{n_2}$, and $\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$.

4. Decide between 2 hypotheses:
   * Select a confidence level $(1 - \alpha)$
   * Determine the *decision threshold* (*critical value*) of the test.
     · If σ^2 is assumed known or test concerns a population proportion, then Excel function `NORM.S.INV` can calculate critical value.
     · If σ^2 is assumed unknown, then Excel function `T.INV` calculates a one-sided critical value and `T.INV.2T` a two-sided critical value.
   * Compute p-value.
     · Quantifies the probability of getting data like the observed data (or something more extreme) if the null hypothesis is true.
     · If σ^2 is assumed known or if test concerns a population proportion, then `NORM.S.DIST` can be used to calculate p-value.
     · If σ^2 is unknown, then `T.DIST` or `T.DIST.2T` can be used to calculate p-value (depending on sidedness of H_a).
   * If p-value $\leq \alpha$, then reject H_0 in favor of H_a.
5. Interpret results in the context of the original research question.

– Type I error: reject the null hypothesis when it is true (a false positive)
  * Often considered serious in legal contexts (e.g., convicting a non-guilty person). Thus, minimizing chances of a Type I error is often prioritized over minimizing chances of a Type II error.
  * Probability of a Type I error is denoted α. Associated confidence interval has a $(1 - \alpha) * 100\%$ confidence level.

– Type II error: fail to reject the null when it is false (a false negative)
  * Small Type II error corresponds to correctly rejecting H_0 if its false.
  * Probability of Type II error is denoted β. The *power* of a hypothesis test is $1 - \beta$.

- Any α-level two-sided hypothesis test can be performed by determining whether the associated $(1-\alpha)*100\%$ confidence interval contains the hypothesized value.
  - E.g., if $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$, then construct the confidence interval $\bar{x} \pm$(critical value)$*SE(\bar{x})$ and determine whether it contains μ_0. If it does, then fail to reject null.
- Rejecting the null does not mean results are practically significant.
  - The "effect size," d, can be a better measure of practical significance:

$$d = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(n_1-1)S_x^2+(n_2-1)S_y^2}{n_1+n_2-2}}}.$$

  - d is called Cohen's d statistic. E.g., values around 0.8 are considered large.
- Performing multiple tests without correcting erodes the collective (family-wise) confidence level.
  - E.g., if 5 tests are performed at the $\alpha = .05$ level, then the family-wise confidence level is actually $.95^5 = .77$.
  - Bonferroni correction can be used to divide the α significance level up across the different tests. Run each of k tests at a significance level of $\frac{\alpha}{k}$.
  - Controlling for the False Discovery Rate (FDR) is another method

- Equivalence Testing
  - Standard hypothesis testing procedures "favor" the null, which commonly dictates that a population parameter is equal to some value (e.g., $H_0 : \mu = \mu_0$).
  - We may want the null to be that a population parameter is different from some value and collect evidence to show that they are actually equal. This is called performing an *equivalence test*.
  - Two One-Sided Tests (TOST) is one way of performing an equivalence test. Consider an example of showing that two population means μ_x, μ_y are equal for illustration:
    - Determine threshold Δ_1, Δ_2 within which it would be appropriate to call the two means equal (e.g., $\pm.5\%$).
    - Two null hypotheses to test are $H_{0_1} : \mu_x - \mu_y \leq \Delta_L$ and $H_{0_2} : \mu_x - \mu_y \geq \Delta_U$. If both of these nulls are rejected, then there's evidence of equality.
    - Statistics are (assuming unknown population variances)

$$\frac{\bar{x} - \bar{y} - \Delta_L}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}} \quad \text{and} \quad \frac{\bar{x} - \bar{y} - \Delta_U}{\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}}$$

    - Compare these statistics to their respective $\pm t_{1-\alpha/2, n_1+n_2-2}$ critical values. If both tests reject, then conclude that equivalence is supported.
    - Could construct a confidence interval for $\mu_x - \mu_y$ as well

- Sample Size Calculation
  - We may want to know before performing a study how large the sample size needs to be to accomplish the study's goals.
  - Suppose a population proportion π is to be estimated. We want a large enough sample to achieve a margin of error $ME \leq m$ with a confidence level of $(1-\alpha)*100\%$. Then the estimated sample size n is

$$n = \left(\frac{z_{1-\alpha/2}}{m}\right)^2 \hat{p}(1-\hat{p})$$

    where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th standard normal quantile (e.g., 1.96 for 95% confidence).

- Suppose we were instead interested in estimating a population mean, μ. We often want to be within some proportion of the true value (called the *relative margin of error* (RME)). Suppose we want our estimate to be within $q\%$ of the true value at a $(1-\alpha)*100\%$ confidence level. Then the sample size calculation would be

$$n = \left(\frac{\sigma^2}{\mu}\right)^2 \left(\frac{z_{1-\alpha/2}}{q}\right)^2.$$

- Slides 41-46 contain information for determining a sample size based on the Hypergeometric distribution.
- If we want to detect an effect of size d with power $1-\beta$ and confidence $1-\alpha$, then a rough sample size estimate is

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2}$$

Part 7: Regression & ANOVA

- Covariance and Correlation
  - Sample covariance between two samples, $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ and $\boldsymbol{y} = (y_1, y_2, ..., y_n)$ say, is

$$Cov(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

  - Sample correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$Corr(\boldsymbol{x}, \boldsymbol{y}) = r_{x,y} = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{S_x S_y}$$

  - A test for significant correlation between two variables, $\boldsymbol{x}, \boldsymbol{y}$ say:
    1. Hypotheses are $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$ where ρ is the true correlation between $\boldsymbol{x}, \boldsymbol{y}$
    2. Test statistic is

$$t = r_{x,y}\sqrt{\frac{n-2}{1-r_{x,y}^2}}$$

    3. Compare against a t distribution with $n-2$ degrees of freedom. So critical values are $\pm t_{1-\alpha/2,n-2}$.

- "All models are wrong, but some are useful." (Box, 1985)
  - Probability/statistical models tell us how a random quantity should behave if our model assumptions are correct.
  - Hypothesis testing provides a formal procedure by which we can determine the validity of an assumed probability/statistical model
    - "If the null were true, then our data should look like [...]. Our observed data look like [...], which [supports/contradicts] the null hypothesis."
  - It is common to assume that one variable might be able to "predict" another if they are associated (e.g., elemental concentration of Cadmium can predict concentration of Aluminium)
  - *Response/dependent variable*: variable of interest in a study - depends on the value of associated factors
  - *Independent/predictor variable:* variable that is viewed as not depending on the value of other variables, but can affect a change on the response.

- Simple Linear Regression (SLR)

– Assume that the average response Y, given the value of the predictor $X = x$, is

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

  * β_0 is the *intercept* parameter, which is equal to $E(Y|X = 0)$. Rarely of interest in problems.
  * β_1 is the *slope* parameter, which is the change in $E(Y|X = x)$ when x increases by 1 unit. This is often the parameter of interest (positive/negative/0 slope means positive/negative/no association).

– In practice, our observed data will not exactly satisfy the assumed linear relationship above (i.e., response values will not lie on a perfect line). As such, we also assume that our observations, y_i for $i = 1, ..., n$, have been "corrupted" by some random noise:

$$y_i = E(y_i|X_i = x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is a random value (the *error*) added to the linear function $\beta_0 + \beta_1 x_i$.

  * In simple linear regression, we assume $\epsilon_i \sim N(0, \sigma^2)$ for some unknown $\sigma^2 > 0$ and ϵ_is are independent of each other, $i = 1, ..., n$. This means that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_is are independent.
  * We can think of the model as making the following assumption: "In reality, an observation y_i follow a perfect linear relationship with the predictor x_i. The only reason why our observed values *don't* all lie on a straight line is because some random values, ϵ_i, "pushed" them off of the line."
  * As you might expect, the simple linear relationship is often a naive assumption (the relationship may in fact be more complicated). However, such a simple model may still be useful even if it is approximately correct.
  * It can be a fool's errand to try to find the "perfect" model. Remember: "all models are wrong, but some are useful." We often settle for "good" models if they tell us useful information.

– Note: this is an *assumed* relationship between the response and predictor variables. In real problems, we often need to verify that this assumption is appropriate (and search of other models if it is deemed extremely inappropriate).

  * The error term ϵ_i indicates how wrong our model specification is.
  * Unfortunately, $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$ meaning, because β_0, β_1 are unknown parameters, the true value of ϵ_i is also unknown.
  * If we obtain estimates of β_0, β_1, call them b_0, b_1, then we can estimate the error terms by $e_i = y_i - b_0 - b_1 x_i$. The e_i values are called *residuals*.

– An often-used criterion for estimating β_0, β_1 is to find b_0, b_1 such that

$$\sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

is minimized. This is called the *least-squares* criterion. The b_0, b_1 that satisfy this criterion are called the *least-squares (LS) estimators*.

  * For simple linear regression, the LS estimators are

$$b_0 = \bar{y} - b_1 \bar{x}$$
$$b_1 = r_{x,y} \frac{S_y}{S_x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

– The estimate for $\beta_0 + \beta_1 x_i$ is $\hat{y}_i = b_0 + b_1 x_i$. The only other parameter is the error term variance, σ^2. This is estimated using the *Mean Square Error*:

$$\hat{\sigma}^2 = S_e^2 = MSE = \frac{SSE}{n - 2}$$

where SSE is the *sum of squared errors* defined as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2.$$

- b_1 is calculated using the y_i, which are assumed random with distribution $Normal(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, ..., n$. As such, b_1 is in fact also random with distribution $Normal(\beta_1, \sigma^2_{b_1})$.
  - $\sigma^2_{b_1} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ is estimated by
$$\hat{\sigma}^2_{b_1} = S^2_{b_1} = \frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
  - A $(1-\alpha) * 100\%$ confidence interval for β_1 is given by
$$b_1 \pm t_{1-\alpha/2, n-2} S_{b_1}.$$
  - A test statistic for testing $H_0 : \beta_1 = \beta_1^*$ is
$$t = \frac{b_1 - \beta_1^*}{S_{b_1}}$$
    which is compared to a t distribution with $n-2$ degrees of freedom.
- Prediction in SLR
  - Prediction refers specifically to anticipating what a new value might be based on the currently fitted model (i.e., applying the model to a new situation on which it was not fitted).
  - Prediction is different from estimation in that we need to incorporate uncertainty about the fitted model, instead of solely the assumed variability of responses.
  - The standard error for a *predicted mean response* $\hat{y}_{x^*} = b_0 + b_1 x^*$ based on a new predictor value $X = x^*$ is
$$SE_{\hat{y}_{x^*}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2 (x^* - \bar{x})^2}{(n-1) S_x^2}}.$$
    Since σ^2 is most likely unknown, we can instead replace it with the MSE.
  - A $(1-\alpha) * 100\%$ CI for $E(Y|X = x^*)$ is then
$$\hat{y}_{x^*} \pm t_{1-\alpha/2, n-2} SE_{\hat{y}_{x^*}}$$
  - Instead of predicting the *mean* response for a new predictor value x^*, we may be interested in predicting the value of a *single* response, y_{n+1} say, given x_{n+1}. This introduces even more variability since, under the assumed model, observations vary randomly around the mean response.
  - The standard error of a predicted single response y_{n+1} given predictor x_{n+1} is
$$SE_{y_{n+1}} = \sqrt{MSE} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1) S_x^2}}.$$
- It is not recommended to predict a response associated with a predictor value that lies outside of the range of predictor values on which a regression model is fitted. This is called *extrapolation*. Predicting response within the original range of predictor values is called *interpolation* and is generally more appropriate.

- Regression with categorical predictors
  - A response variable may depend on a categorical predictor (e.g., heights of males vs. females)
  - We can use *dummy* (or *indicator*) *variables* to represent the values of categorical predictors in a model.
  - Consider an example in which height y_i is response variable. The sternum height is a continuous predictor, call it x_i, and the sex ("Male" or "Female") is a categorical variable.
    - We need to "binarize" the sex variable by arbitrarily assigning each category to 0 or 1. Say 1 corresponds to "Male" and 0 to "Female."

* The dummy variable d_i will represent this binarization:

$$d_i = \begin{cases} 1 & \text{if subject } i \text{ is Male} \\ 0 & \text{if subject } i \text{ is Female.} \end{cases}$$

* The regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i.$$

  · Note

$$\begin{aligned} y_i|\text{subject } i \text{ is male} &= \beta_0 + \beta_1 x_i + \beta_2(1) + \epsilon_i = (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i \\ y_i|\text{subject } i \text{ is female} &= \beta_0 + \beta_1 x_i + \beta_2(0) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i. \end{aligned}$$

  So adding the dummy variable d_i allows for the intercepts between the "Male" and "Female" models to differ.

* We may desire a model that allows for both the intercept and slope to differ between these two models. Said another way, we may *assume* that the effect of sternum length on an individual's height depends on whether that individual is Male or Female. This type of an effect is referred to as an *interaction* between sternum length and sex.
* To represent an interaction, we can introduce the *interaction variable* z_i where

$$z_i = d_i * x_i = \begin{cases} x_i & \text{if subject } i \text{ is male} \\ 0 & \text{if subject } i \text{ if female.} \end{cases}$$

* The model with included z_i interaction is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 z_i + \epsilon_i.$$

  · Now note that

$$\begin{aligned} y_i|\text{subject } i \text{ is male} &= \beta_0 + \beta_1 x_i + \beta_2(1) + \beta_3(x_i) + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \epsilon_i \\ y_i|\text{subject } i \text{ is female} &= \beta_0 + \beta_1 x_i + \beta_2(0) + \beta_3(0) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i \end{aligned}$$

  meaning the interaction has the desired effect of allowing the slope on x_i to differ.

- Quadratic Regression
  - The relationship between a response and predictor may be non-linear (e.g., curved).
  - A *quadratic* relationship between variables y and x is of the form

$$y = a + bx + cx^2.$$

    The shape of a quadratic relationship when plotted is called a *parabola*.
  - A quadratic regression model is of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

    where ϵ_i are independent with distribution $Normal(0, \sigma^2)$, $i = 1, ..., n$.
  - Estimation for $\beta_0, \beta_1, \beta_2$ is commonly done using least-squares. However, the solutions can't be expressed as "nicely" as in SLR, so we rely on statistical packages to estimate (e.g., Analysis TookPak in Excel).
  - After obtaining estimates b_0, b_1, b_2, the estimator for σ^2 is

$$MSE = \frac{1}{n-3} \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_i - b_2 x_i^2\right)^2$$

    with associated degrees of freedom $n - 3$.

- Model diagnostics
  - Residual diagnostics
    * We often want to know how poor a model fit is to determine if we should search for better models.
    * The residuals, $e_i = y_i - \hat{y}_i$ are an estimate of the error terms and can indicate how well a model fits.
    * In particular, we can use the e_i values to test 2 assumptions made about the error terms:
      1. The expectation of the error terms is 0
      2. The variance of the error terms is a constant value, σ^2.
    * It is common to visualize the residuals e_i as a scatter plot against the predictor values x_i. This is called a *residual plot*.
      · For the model assumptions to be met, the residual plot should show a random scatter about the horizontal 0 line.
      · If a curved pattern is visible in the residuals, then the 0 expectation assumption is not met. This may require adding additional terms to the model.
      · If the concentration of residuals changes with the predictor values, then the constant variance assumption is not met. This may require changing the assumed variance term σ^2.
  - The *coefficient of determination*, $R^2 = r^2_{x,y}$, is the proportion of variability in the response that can be explained by the predictor variable.
  - Influential points
    * An *outlier* is an extreme value in the y (response) direction
    * A *leverage point* is an extreme value in the x (predictor) direction.
    * A point that is both an outlier and a leverage point is called an *influential point* since it can have a large influence on the slope/intercept of the least-squares fitted regression model.
    * One can detect such points using a scatter plot.

- Classification and Logistic Regression
  - We may be interested in predicting a categorical variable given other information (e.g., can we predict brand of shoe if we're told the number of triangles on its sole?)
  - Prediction of a categorical variable is commonly referred to as *classification* (as opposed to regression, which is a prediction of a continuous variable).
  - Binary classification is important for forensics (match or non-match, same source or different source, etc.)
  - Logistic regression provides a framework for estimating the probability that a categorical response, y_i, takes on one of two values given a predictor x_i.
  - Consider example of predicting someone's sex, y_i, based on hand breadth, x_i.
    * Binarize y_i arbitrarily. Say 1 corresponds to "female" and 0 to "male."
    * Logistic regression model assumes $y_i|X_i = x_i$ are independent with distribution $Binomial(1, \pi_i)$.
    * In this example, $\pi_i = Pr(y_i = 1|x_i) = Pr(\text{subject } i \text{ is female}|x_i)$.
    * The relationship between π_i and x_i is assumed to be

$$\pi_i = f(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

      where β_0, β_1 are unknown. $f(x)$ is known as the *logistic function*.
    * Estimation of β_0, β_1 is complicated. We let software packages do it for us.
    * Upon obtaining b_0, b_1, the estimated probability that the ith subject if female is $\hat{\pi}_i = f(b_0 + b_1 x_i)$.
    * It seems reasonable to classify subject i as female if $\hat{\pi}_i > .5$ and male otherwise.
      · This is called formulating a classification *decision rule*.
      · Simplistic decision rules, while easy to apply, are bound to make errors. One should always be cautious about the results of an automatic classification algorithm.

- Given a $(1-\alpha)*100\%$ confidence interval for $\beta_0+\beta_1 x_i$, $[L,U]$ say, we can calculate a confidence interval for π_i by considering

$$\left[\frac{\exp(L)}{1+\exp(L)}, \frac{\exp(U)}{1+\exp(U)}\right].$$

  * Two-sided α-level hypothesis tests for, say, $H_0: \pi_i = \pi_0$ can be performed by determining whether the above $(1-\alpha)*100\%$ confidence interval for π_i contains π_0.

Part 8: Assessing Evidence

- Source conclusions are an important aspect of forensic examination
  - E.g., evidence collected from suspect, E_s matches evidence found at crime scene, E_c - the two pieces of evidence have the same source
- Consider the same source hypothesis, S, and the different source hypothesis, S^c.
- Logic of forensic examination:
  1. Examine the evidence E_s, E_c to identify similarities/differences.
  2. Assess the observed similarities and differences to see if they are expected (or likely) under the same source hypothesis
  3. Assess the observed similarities and differences to see if they are expected (or likely) under the different source hypothesis
     - This requires assessing whether matching features would be unusual if the suspect is not the source.
- Expert Assessment
  - Expert evaluates evidence based on experience, training, use of accepted methods in field
  - Assessment reflects examiner's expert opinion
  - Conclusions are typically categorical in nature (e.g., identification, inconclusive, exclusion)
  - Conclusions should refer to the evidence, not the the hypotheses themselves
    * Odds form of Bayes' rule is

$$\frac{Pr(S|E)}{Pr(S^c|E)} = \frac{Pr(E|S)}{Pr(E|S^c)}\frac{Pr(S)}{Pr(S^c)}$$

    * A statement like "Based on the evidence, the author of the known samples wrote the questioned document" is a statement about the left hand side.
    * To make a statement about the left hand side requires inserting one's pre-evidence opinion about the same source proposition (the prior odds)
      · The expert opinion should not come with pre-evidence "strings" attached.
      · The expert is there to educate the jury about how likely the evidence is under the two hypotheses.
  - Important to understand the reliability and accuracy of an approach to forensic analysis to satisfy Daubert
- Two-Stage Approach
  1. Stage 1 (Similarity)
     * Determine if the crime scene and suspect objects agree on one or more characteristics
       · Can use hypothesis tests to assess strength of evidence towards same source hypothesis
     * Conclusion is that two samples "are indistinguishable" or "match."
     * Steps of Implementation (based on hypothesis test):

(a) Characterize each object by mean value (e.g., mean trace elemental concentration in population of glass fragments)
(b) Obtain sample values from crime scene object
(c) Obtain sample values from suspect's object
(d) Use sample values to test hypothesis that two samples have the same population mean (i.e., same source). Can use t-test or equivalence test to do so.
(e) Summarize test results using p-value, probability of data like like the observed data, assuming null hypothesis is true
(f) Reach a conclusion (based on an α-level like .05 or .01): small p-value indicates strong evidence towards alternative hypothesis
(g) Otherwise, can't reject the null hypothesis.

2. Stage 2 (Identification)
   * Assess the significance of the agreement by finding the likelihood of such agreement occurring by chance
     · E.g., if blood types are found to be indistinguishable, how likely is it that two random individuals' blood types (say, of a particular ethnic descent, sex, etc.) are indistinguishable?
   * Important to consider that "someone else" could have left evidence (even if E_s, E_c are deemed "indistinguishable").
     · Difficult to pin-down what "someone else" refers to. Requires defining a *relevant population*.
     · For example, suppose E_s, E_c are elemental concentrations from a glass fragment found on suspect and found at the crime scene, respectively. Suppose they are deemed indistinguishable in step 1.
     · Need to consider, for example, a large sample of glass fragments (of the same type, manufacturer, etc.) to determine how likely it is that two non-match glass fragments (from two different panes) have indistinguishable elemental concentrations.
     · Amassing such a large data set can be expensive, time-intensive, etc.
   * Probability of a coincidental match is high when
     · When the known sample is "ordinary" relative to the relevant population
     · Large amount of heterogeneity among the potential random sources in the population (e.g., elemental concentrations of float glass vary wildly when produced by manufacturer A)
     · There is a large amount of variability within a single source (e.g., elemental concentration highly dependent on where a fragment is sampled in a pane of glass)

– Problems with Two-Stage Approach
   * The null hypothesis (often that pieces of evidence have equal population means) is assumed true unless there's strong evidence suggesting otherwise.
     · Equal mean null hypothesis means that the two pieces of evidence are of same source - which is against the suspect. This doesn't align well with the presumption of innocence principle.
     · Equivalence test can be used as an alternative.
   * Hypothesis tests result in binary decision (same source or different source) and require setting an α level.
     · Difficult to know what the "best" α level is.
     · Small α means we'll reject more often - may fail to include important evidence. Large α means we fail to reject more often - may mistakenly incriminate a non-guilty person.
     · Absence of precision in data favors the "same source" hypothesis (samples look more alike due to imprecise measurements)
   * Separation into two steps is not optimal
     · Step 2 (quantifying probative value/practical significance of evidence) only occurs if results are found to statistically significant. But statistical significance is not the same as practical significance.

· E.g., suppose hypothesis test finds that two blood samples are of same type. This doesn't have a high probative value of evidence as many people share blood types.
· The Bayesian approach via likelihood ratios addresses this issue.

- Likelihood Ratio Approach
  - Likelihood ratio is

$$\frac{Pr(E|S)}{Pr(E|S^c)}$$

  - The numerator assumes common source, S, and asks about the likelihood of the evidence in that case
    * Similar to finding a p-value, but doesn't require a binary decision at the end
  - The denominator assumes different source, S^c, and asks about likelihood of evidence in that case
    * Analogous to finding the coincidental match probability.
  - In a sense, the likelihood ratio combines both stages of the Two-Stage Approach into one value.
  - An LR-based conclusion: "The evidence is [LR] times more likely if the objects have the same source than if the objects have different sources."
  - The *Prosecutor's Fallacy* corresponds to mistaking $P(S^c|E)$ for $P(E|S^c)$.
    * Evidence is unlikely under S^c is interpreted as saying that S^c is unlikely.
  - Problems with the Likelihood Ratio Approach
    * What do we mean by "evidence?"
      · Data can be very high dimensional (e.g., images)
      · Unclear what feature(s) (say, of an image) should be treated as evidence.
      · E is often taken to be the observed similarities and differences between two pieces of evidence.
    * Requires information about...
      · variation expected in repeated measurement of the same source to inform $P(E|S)$.
      · variation expected in measurements across different items in the population (i.e., the "coincidental match" probability) to inform $P(E|S^c)$.
      · how identifiable features are formed on the evidence (e.g., manufacturing. distribution, wear pattern information to understand how Randomly Acquired Characteristics appear on shoe prints).
      · One approach is to assign probability distributions to these features. This requires some subjectivity though (similar to choosing any model).
    * Some advocate for mapping the value of a LR to a "verbal equivalent" conclusion. One example of a table from ENFSI is:

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

* Oftentimes, the features to which we assign a probability distribution are "similarity scores" measuring the similarity between two pieces of evidence.
    - We can consider the likelihood ratio of these scores under the competing hypotheses. These are called *score-based likelihood ratios* (SLRs).
    - Say D is the score between two pieces of evidence. The SLR would be $\frac{Pr(D|S)}{Pr(D|S^c)}$.
    - Calculating these probabilities requires that we understand the distribution of D under both S and S^c.
    - Learning about $Pr(D|S)$ is relatively straightforward, assuming we have a sufficiently large number of matching pieces of evidence (i.e., we can understand how D behaves among actually matching sources).
    - Learning about $Pr(D|S^c)$ is harder. It requires we specify to which population we are referring when we talk about "different source" (e.g., do we include only shoes of the same model, of the same brand, of the same year, etc.). Is there even a *single* non-match/different source score distribution, or should we consider multiple across different definitions of "different source?"

Part 9: Reporting Testimony

- When giving expert testimony, it is important to understand how conclusions are interpreted by a juror. For example, is there a way of phrasing the statement that is best understood by jurors?
- Types of conclusions include:

  1. Likelihood Ratios: "Evidence is [LR] times more likely if the same source hypothesis is true than if the different source hypothesis is true."
     - Examiners may provide their conclusions in terms of an LR that reflects their beliefs about the relevant likelihoods in light of their training and experience (e.g., 1/10000). Sometimes criticized for being a number constructed "from nowhere;" that is, without a rigorous mathematical calculation.
  2. Categorical conclusion: "Two pieces of evidence are indistinguishable."
     - Criticized for implying absolute certainty (uncertainty is inherent to conclusions)
     - Doesn't indicate the probative value of evidence (e.g., blood types may be indistinguishable)
  3. Random Match Probability: "The probability that a randomly chosen [member of the population] would match this evidence is 1 in [some value]."
     - If the matching features are certain to be observed under the same source hypothesis (the numerator of the LR is $Pr(E|S) = 1$), then the RMP and LR convey the same information (e.g., "1 in 10" if LR = 1/10).
     - If $Pr(E|S) < 1$, then the RMP and LR do not convey the same information. In this case, LRs are preferred as RMPs provide an incomplete and potentially misleading account of the strength of evidence
  4. Strength of Support (SOS) statements: "The forensic findings provide [some level of] support] for the same source proposition relative to the different source proposition."
     - Intended to be a "verbal equivalent" to an LR (see table in previous section)
  5. Source Probability (SP) statements: "There is a [number between 1-100]% chance that the two pieces of evidence have a common source."
     - Criticized for being a statement regarding the validity of the same source hypothesis given the data (i.e., about $\frac{Pr(S|E)}{Pr(S^c|E)}$), when, by Bayes' Rule, this logically must imply that the forensic examiner is inserting their own prior belief about the validity of the two hypotheses (see discussion in previous section).
     - Forensic scientists cannot logically draw conclusions about source probabilities based on forensic science alone.

- See slides 26-34 for a discussion of the Thompson et al. (2018) study *Perceived strength of forensic scientists' reporting statements about source conclusion.*