Statistical Thinking for Forensic Practitioners

Important Results & Procedures

This document is intended to be a summary of the important results (formulas, theorems, etc.) and "rote" procedures (e.g., hypothesis tests) covered in the course. Use this as a "quick" reference for homework assignments, etc.

Part 2: Probability

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

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

Odds Ratio

- Odds in favor of E is

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

- Odds against E is

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

- Given odds against E, O_a,

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

Conditional Probability

- Definition of the conditional probability of A given B:

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

- Above definition is equivalent to (by algebra) $Pr(A \text{ and } B) = Pr(B|A)Pr(A) = Pr(A|B)Pr(B)$.
- A and B are independent if $Pr(A \text{ and } B) = Pr(A)Pr(B)$.
  - Equivalently, if $Pr(A|B) = Pr(A)$ and $Pr(B|A) = Pr(B)$.
- The "Law of Total Probability:"

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

Bayes' Theorem

- The conditional probability of A given B is equal to

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

- In courtroom, commonly consider events E = evidence, H_s = "same source" proposition, and H_d = "different source" proposition.
- "Odds Form" of Bayes' Theorem:

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

Part 3: Data Collection

- Simple Random Sampling:
  - Every sample of size n drawn from the population of size N has the same probability of selection.
  - If sampling with replacement, then each sample has a $\frac{1}{N^n}$ probability of being selected where N^n is the total number of possible samples.
  - If sampling without replacement, then there are

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

possible samples of size n. 1 over this quantity gives the probability that any one sample is selected.

Part 5: Probability Models and Uncertainty

Common discrete probability distributions

- Binomial distribution
  - Often used to describe the number of "successes," X say, out of n independent, binary trials.
  - If $X \sim Binomial(n, p)$, then the probability mass function is given by

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

- 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.
  - If $X \sim Hypergeometric(N, K, n)$, then the probability mass function is given by

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

- Poisson distributions
  - Often used to describe the number of events, X, occurring in an interval of time/space.
  - 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

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

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).
  - 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.
  - 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

- The *expected value* of a random variable X, denoted $E(X)$ is defined by

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

  - 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 Var(X), is defined by

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

  - 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
$Log - Normal(\mu, \sigma^2)$	$\exp\left(2\mu + \sigma^2\right)(\exp(\sigma^2) - 1)$

Table 2: Variances of Common Distributions

Covariance and Correlation

- The *covariance* between two random variables X and Y is defined as

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}.$$

  - 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.

Part 6: Inference

Point Estimation

- Common population parameters of interest and their associated estimators include:

Pop. Parameter	Sample Estimator	Formula
Mean μ	$\bar{x}$	$\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$
Variance σ^2	S^2	$\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)$
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.

- 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 as n grows.

Interval Estimators

- Confidence intervals are all of the general form:

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

- 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.

- Table summarizing standard errors under various situations:

		Standard Error	
Parameter of Interest	Pt. Estimator	Pop. Var. Known	Pop. Var. Unknown
μ	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$ $d.f. = n-1$
$\mu_1 - \mu_2$	$\bar{x} - \bar{y}$	$\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}$	$\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}$ $d.f. = n_1 + n_2 - 1$
π (A pop. proportion)	$\hat{p} = \frac{Y}{n}$ $Y = \#$ successes	$\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ (E.g., $H_0 : \pi = \pi_0$)	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Hypothesis Testing

1. Formulate 2 hypotheses, H_0 and H_a.
2. Collect data and calculate relevant statistic
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:

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

   - Table summarizing test statistics in various situations:

Null Hypothesis	Test Statistic	Standard Error: Pop. Var. Known	Standard Error: Pop. Var. Unknown
$H_0 : \mu = \mu_0$	$\frac{\bar{x}-\mu_0}{SE_{\bar{x}}}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{S}{\sqrt{n}}$ $d.f. = n - 1$
$H_0 : \mu_1 - \mu_2 = d$ ($d = 0$ often)	$\frac{(\bar{x}-\bar{y})-d}{SE_{\bar{x}-\bar{y}}}$	$\sqrt{\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}}$	$\sqrt{\frac{S_x^2}{n_1} + \frac{S_y^2}{n_2}}$ $d.f. = n_1 + n_2 - 1$
$H_0 : \pi = \pi_0$ (A pop. proportion)	$\frac{\hat{p}-\pi_0}{SE_{\hat{p}}}$	$\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$	$\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ (Note π_0 used here)
$H_0 : \pi_1 - \pi_2 = d$ ($d = 0$ often)	$\frac{(\hat{p}_1-\hat{p}_2)-d}{SE_{\hat{p}_1-\hat{p}_2}}$	$\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $\hat{p} = \frac{Y_1+Y_2}{n_1+n_2} = \frac{n_1 * \hat{p}_1 + n_2 * \hat{p}_2}{n_1+n_2}$	$\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

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 the test concerns a (single) 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.

- 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

- Two One-Sided Tests (TOST) is one way of performing an equivalence test. Consider an example of showing that two population means μ_1, μ_2 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 $H_{0_1} : \mu_1 - \mu_2 \leq \Delta_L$ and $H_{0_2} : \mu_1 - \mu_2 \geq \Delta_U$. 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 $\mu_1 - \mu_2$ 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

Testing for significant 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 $\pm t_{1-\alpha/2, n-2}$.

Simple Linear Regression (SLR)

- SLR model assumes for each $i = 1, ..., n$:

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

  where $\epsilon_i \sim Normal(0, \sigma^2)$ are independent.

- For simple linear regression, the least-squares estimators are

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

- $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$ is called a *residual.*
- The estimator for σ^2, the *Mean Square Error*, is:

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

  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 random with distribution $Normal(\beta_1, \sigma_{b_1}^2)$.
  - $\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is estimated by

$$\hat{\sigma}_{b_1}^2 = S_{b_1}^2 = \frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 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
  - 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

$$\hat{y}_{x^*} \pm t_{1-\alpha/2,n-2} SE_{\hat{y}_{x^*}}$$

  - The standard error of a predicted single response y_{n+1} given predictor x_{n+1} is

$$SE_{y_{n+1}} = \sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{(n-1)S_x^2}}.$$

Regression with categorical predictors

Regression with categorical predictors

- 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 expessed as

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

    * Note

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

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

  - 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

- 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 $Normal(0, \sigma^2)$, $i = 1, ..., n$.

- 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 $n - 3$.

Classification & Logistic Regression

- 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*.
  - Upon obtaining b_0, b_1, the estimated probability that the ith subject if female is $\hat{\pi}_i = f(b_0 + b_1 x_i)$.

- 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

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?

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.
- 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."
- 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** ..."