STAT 509 Final Exam Formula Sheet
Instructor: Shiwen Shen
Exam Day: June 23

- Construct confidence interval: **point estimate $\pm$ margin of error**:
  - Population proportion p: $\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - One population mean μ (σ known): $\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
  - One population mean μ (σ unknown): $\bar{x} \pm t_{\frac{\alpha}{2},\text{df}} \frac{s}{\sqrt{n}}$ where $\text{df} = n-1$
  - $\mu_1 - \mu_2$ with known σ_1^2 and σ_2^2:
  $$(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
  - $\mu_1 - \mu_2$ with unknown σ_1^2 and σ_2^2, assuming $\sigma_1^2 = \sigma_2^2$:
  $$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$
  where
  $$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$
  - $\mu_1 - \mu_2$ with unknown σ_1^2 and σ_2^2, assuming $\sigma_1^2 \neq \sigma_2^2$:
  $$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$
  where
  $$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$
  - Matched Pairs:
  $$\overline{y}_D \pm t_{n-1,\alpha/2} \frac{s_D}{\sqrt{n}}$$
  - $p_1 - p_2$:
  $$\hat{p_1} - \hat{p_2} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- 4-step procedure to construct hypothesis testing:

  1. State H_0 and H_a
  2. Calculate test statistic:
     - Population proportion: $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
     - Population mean (σ known): $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
     - population mean (σ unknown): $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
     - $\mu_1 - \mu_2$ with known σ_1^2 and σ_2^2:

       $$z_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

     - $\mu_1 - \mu_2$ with unknown σ_1^2 and σ_2^2, assuming $\sigma_1^2 = \sigma_2^2$:

       $$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

     - $\mu_1 - \mu_2$ with unknown σ_1^2 and σ_2^2, assuming $\sigma_1^2 \neq \sigma_2^2$:

       $$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

     - Matched Pairs:

       $$t_0 = \frac{\bar{y}_D}{s_D/\sqrt{n}}$$

     - $p_1 - p_2$:

       $$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

       where

       $$\hat{p}_0 = \frac{Y_1 + Y_2}{n_1 + n_2}$$

  3. Calculate p-value

     $$\begin{aligned} H_a: p > p_0(\mu > \mu_0): &\quad P(Z > z_0) \text{ or } P(t > t_0) \\ H_a: p < p_0(\mu < \mu_0): &\quad P(Z < z_0) \text{ or } P(t < t_0) \\ H_a: p \neq p_0(\mu \neq \mu_0): &\quad 2P(Z < -|z_0|) \text{ or } 2P(t < -|t_0|) \end{aligned}$$

  4. Make decision and state conclusion:
     - p-value $\leq \alpha \Longrightarrow$ Reject H_0
     - p-value $> \alpha \Longrightarrow$ Fail to reject H_0

- Least Squares Estimators for simple linear regression:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}} \left[= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \right]$$

- Residual for the i^{th} observation: $e_i = y_i - \hat{y}_i$.
- $SSTO = \sum_{i=1}^n (y_i - \overline{y})^2$
- $SSR = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$
- $SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $MSE = \hat{\sigma}^2 = \frac{SSE}{n-2}$ (Simple Linear Regression)
- Coefficient of Determination: $r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO}$
- Box-Cox transformation:

$$\text{BoxCox}(Y) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}$$

- Logistic regression: $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$ and $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$
- Odds: $\frac{p}{1-p}$
- Odds ratio: $e^{\hat{\beta}_1}$
- Multiple linear regression model: $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, ..., n.$
- Matrix Representation: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

- Least Squares Estimator: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$
- Prediction function:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{H}\mathbf{Y}$$

  where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is hat-matrix.
- Matrix A is symmetric if $A = A^T$; matrix A is idempotent if $AA = A$.
- $SSE = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$
- $MSE = \hat{\sigma}^2 = \frac{SSE}{n-p-1}$ (Multiple Linear Regression)