FORMULA SHEET

Chapter 1

NONE

Chapter 2

Probability:

- Any Two Events:
  - $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$
  - $P(E_1|E_2) = \frac{P(E_1 \text{ and } E_2)}{P(E_2)}, \quad P(E_2) > 0$
  - $P(E_1 \text{ and } E_2) = P(E_1)P(E_2|E_1) = P(E_2)P(E_1|E_2)$
- Mutually Exclusive Events:
  - $P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$
- Independent Events:
  - $P(E_1|E_2) = P(E_1), \quad P(E_2)$
  - $P(E_2|E_1) = P(E_2), \quad P(E_1)$
  - $P(E_1 \text{ and } E_2) = P(E_1)P(E_2)$

Chapter 3

Measures of Center/Location:

- Sample Mean:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Median: the center value that divides the numerically ordered data collection in two halves.

- Percentiles: The p^{th} percentile in a collection of ordered data is a value that divides the data set into two parts. The lower segment contains at least $p\%$ and the upper segment contains at least $(100 - p)\%$ of the data.

- Quartiles: Quartiles are a special case of the percentiles where the first quartile, Q_1, has $p = 25$ and the third quartile, Q_3, has $p = 75$.

Measures of Spread:

- Range: The range of the data is the difference between the maximum and minimum value in the data set.

- Interquartile Range: The interquartile range of the data is the difference between the third and first quartile.
$$\text{IQR} = Q_3 - Q_1$$

- Sample Variance:
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Sample Standard Deviation:
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Measures of Shape:

- Sample Skewness:
$$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$
- Sample Kurtosis:
$$g_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4\right)$$
- Sample Excess Kurtosis:
$$g_2^* = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4\right) - 3$$

Normal Distribution

- Density Curve:
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$
- Standardized Value:
$$z = \frac{x-\mu}{\sigma}$$

Chapter 4

Sampling Distributions

- Sample Means:

$$\bar{x} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

- Sample Proportions:

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} \sim \mathrm{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - \pi}{\left(\sqrt{\frac{\pi(1-\pi)}{n}}\right)}$$

Chapter 5

Confidence Intervals

- General Form:
  - Point Estimate $\pm$ Margin of Error
  - Margin of Error = Critical Value $\times$ Standard Error

- Sample Means:
  - Confidence Interval:
  $$\bar{x} \pm (t^*)\frac{s}{\sqrt{n}}$$
  - Sample Size:
  $$n = \frac{z^2\hat{\sigma}^2}{e^2}$$

- Sample Proportions:
  - Confidence Interval:
  $$\hat{p} \pm (z^*)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
  - Sample Size:
  $$n = \frac{z^2\hat{\pi}(1-\hat{\pi})}{e^2}$$

Chapter 6

Hypothesis Testing

- General Form of Test Statistic:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Null Value}}{\text{Standard Error}}$$

- Sample Means:
  - Test Statistic:

$$t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$$

- Sample Proportions:
  - Test Statistic:

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

Chapter 7

Correlation

- Sample Correlation Coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Hypothesis Test for Correlation Coefficient:
  - Hypotheses:

$$H_0 : \rho_1 = 0, \qquad H_a : \rho_1 \neq 0$$

  - Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$d.f. = n - 2$$

Simple Linear Regression

- Population Simple Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

- Sample Simple Linear Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

- Residual:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

  - Sample Slope Coefficient Calculation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = r \cdot \frac{s_y}{s_x}$$

– Sample Intercept Coefficient Calculation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Sum of Squares Error (Residuals):

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Sum of Squares Regression:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Total Sum of Squares:

$$\text{TSS} = \text{SSE} + \text{SSR} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Coefficient of Determination:

$$r^2 = R^2 = \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- Inference for Regression:

$$\begin{aligned} \text{Test Statistic} &= \frac{\text{Statistic} - \text{Null Value}}{\text{Standard Error}} \\ t &= \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}}, \qquad d.f. = n - 2 \end{aligned}$$

  – Estimate for $\sigma_{\hat{\beta}_1}$:

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

  – Estimate for σ_ε:

$$s_\varepsilon = \sqrt{\frac{SSE}{n - k - 1}}$$

- Confidence Interval for Slope:

$$\hat{\beta}_1 \pm t^* \cdot s_{\hat{\beta}_1}$$

- Confidence Interval for the Mean Value of y for $x = x_p$:

$$\hat{y} \pm (t_{\alpha/2}) s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$d.f. = n - 2$$

- Prediction Interval for an Individual y for $x = x_p$:

$$\hat{y} \pm (t_{\alpha/2}) s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

$$d.f. = n - 2$$

Multiple Linear Regression

- Population Multiple Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

- Sample Multiple Linear Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i}$$

- R^2 Value:

$$R^2 = \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}$$

- Adjusted R^2 Value:

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

Inference for Multiple Regression

- Sum of Squares Error (Residuals):

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Sum of Squares Regression:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Total Sum of Squares:

$$\text{TSS} = \text{SSE} + \text{SSR} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Mean Square Regression:

$$\text{MSR} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{k} = \frac{\text{SSR}}{k}$$

- Mean Square Error:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = \frac{\text{SSE}}{n-k-1}$$

- F-Test Statistic:

$$F = \frac{\left(\frac{\text{SSR}}{k}\right)}{\left(\frac{\text{SSE}}{n-k-1}\right)} = \frac{\text{MSR}}{\text{MSE}}$$

- Test Statistic:

$$\begin{aligned} \text{Test Statistic} &= \frac{\text{Statistic} - \text{Null Value}}{\text{Standard Error}} \\ t &= \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}}, \qquad d.f. = n-k-1 \end{aligned}$$

- Standard Error:

$$s_{\hat{\beta}_j} = \frac{s_\varepsilon}{(1-R_j^2)\sqrt{\sum_{i=1}^{n}(x_{j,i} - \bar{x}_j)^2}}, \quad s_\varepsilon = \sqrt{\frac{\text{SSE}}{n-k-1}} = \sqrt{\text{MSE}}$$

Polynomial Regression Model

- Sample Polynomial Linear Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{1,i}^2 + \ldots + \hat{\beta}_k x_{1,i}^k$$

Interaction Terms

- Sample Model with Interaction Between Two Variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \hat{\beta}_3 x_{1,i} x_{2,i}$$

Multicollinearity

- Variance Inflation Factor:

$$VIF = \frac{1}{1 - R_j^2}$$

Residual Analysis

- Regression Residual:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- Properties of Regression Residuals:

  1. Mean of Residuals:

  $$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

  2. Standard Deviation of Residuals:

  $$s = \sqrt{\frac{\sum \hat{\varepsilon}_i^2}{n - (k+1)}} = \sqrt{\frac{SSE}{n - (k+1)}}$$

- Box-Cox Transformation:

$$y' = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

- Durbin-Watson d statistic:

$$d = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

- Outlier & Influential Observations:

  - Standardized Residual:

  $$\hat{\varepsilon}^* = \frac{y_i - \hat{y}_i}{s}$$

– Studentized Residual:

$$\hat{\varepsilon}^{**} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_i}}$$

– Influential Point Cut-off:

$$h_i > \frac{2(k+1)}{n}$$

– Cook's Distance:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE} \left(\frac{h_i}{(1 - h_i)^2} \right)$$

  * Influential if $D_i > \frac{4}{n}$

– DFFITS:

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s_{(i)}\sqrt{h_i}}$$

  * Influential if $|DFFITS| > 2\sqrt{\frac{k+1}{n}}$

– DFBETA:

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{\hat{\beta}_j}}$$

  * Influential if $|DFBETA| > \frac{2}{\sqrt{n}}$

Chapter 8

Two Population Means

- Equal Variances:
  - Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{1,0} - \mu_{2,0})}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

$$d.f. = n_1 + n_2 - 2$$

  - Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Unequal Variances:
  - Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{1,0} - \mu_{2,0})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$d.f. = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}\right)}$$

  - Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two Population Variances

- Test Statistic:

$$F = \frac{s_i^2}{s_j^2}$$

$$\text{numerator } d.f. = n_i - 1, \text{ denominator } d.f. = n_j - 1$$

Paired Differences

- Sample Differences:

$$\bar{d} = \frac{1}{n_d} \sum_{i=1}^{n_d} d_i, \quad d_i = x_{1,i} - x_{2,i}$$

- Test Statistic:

$$t = \frac{\bar{d} - \mu_{d,0}}{\left(\frac{s_d}{\sqrt{n_d}}\right)}$$

$$s_d = \sqrt{\frac{1}{n_d - 1} \sum_{i=1}^{n_d} \left(d_i - \bar{d}\right)^2}$$

$$d.f. = n_d - 1$$

- Confidence Interval:

$$\bar{d} \pm t^* \cdot \frac{s_d}{\sqrt{n_d}}$$

Two Population Proportions

- Test Statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (\pi_{1,0} - \pi_{2,0})}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

- Confidence Interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

One-Way ANOVA

- Sum of Squares Between:

$$SSB = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2$$

- Sum of Squares Within:

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

- Total Sum of Squares:

$$TSS = SSB + SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2$$

- Mean Sum of Squares Between:

$$MSB = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 = \frac{SSB}{k-1}$$

- Mean Sum of Squares Within:

$$MSW = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2 = \frac{SSW}{N-k}$$

- Hypothesis Statement:

$$\begin{aligned} \mathrm{H}_0 &: \ \mu_1 = \mu_2 = \ldots = \mu_k \\ \mathrm{H}_A &: \ \text{At least two means are not equal} \end{aligned}$$

- Test Statistic:

$$F = \frac{MSB}{MSW}, \ \text{numerator } d.f. = k-1, \ \text{denominator } d.f. = N-k$$

- Tukey-Kramer Critical Range:

$$\text{Critical Range (Margin of Error)} = q_\alpha \cdot \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

ANOVA with Randomized Block

- Sum of Squares Blocks:

$$SSBL = \sum_{i=1}^{b} k\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

- Total Sum of Squares:

$$TSS = SSBL + SSB + SSW$$

- Mean Sum of Squares Blocks:

$$MSB = \frac{1}{b-1}\sum_{i=1}^{b} k\left(\bar{x}_j - \bar{\bar{x}}\right)^2 = \frac{SSBL}{b-1}$$

- Mean Sum of Squares Between:

$$MSB = \frac{SSB}{k-1}$$

- Mean Sum of Squares Within:

$$MSW = \frac{SSW}{(k-1)(b-1)}$$

- Hypothesis Statement:

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2 = \ldots = \mu_k \\ \mathrm{H}_A &: \text{At least two means are not equal} \end{aligned}$$

- Test Statistic:
  $F = \dfrac{MSB}{MSW}$, numerator $d.f. = k{-}1$, denominator $d.f. = (k{-}1)(b{-}1)$

- Hypothesis Statement:

$$\begin{aligned} \mathrm{H}_0 &: \mu_{b_1} = \mu_{b_2} = \ldots = \mu_{b_b} \\ \mathrm{H}_A &: \text{At least two block means are not equal} \end{aligned}$$

- Test Statistic:
  $F = \dfrac{MSBL}{MSW}$, numerator $d.f. = b{-}1$, denominator $d.f. = (k{-}1)(b{-}1)$

- Fisher's Least Squares Difference:

$$\mathrm{LSD} = t^* \cdot \sqrt{MSW} \cdot \sqrt{\frac{2}{b}}$$

Chapter 9

Categorical Data Analysis

- Pearson χ^2 Test of Association:

$$Q_P = \sum \left[\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \right]$$

- Likelihood Ratio χ^2 Test of Association:

$$Q_{LR} = 2 \cdot \sum_i \sum_j O_{i,j} \log \left(\frac{O_{i,j}}{E_{i,j}} \right)$$

- Mantel-Haenszel χ^2 Test of Association:

$$Q_{MH} = (n-1)r^2$$

- Cramer's V Statistic:

$$V = \sqrt{\frac{Q_P/n}{\min(R-1, C-1)}}$$

- Odds:

$$odds(A) = \frac{P(A)}{1 - P(A)}$$

Chapter 10

Clustering

- Euclidean Distance:

$$d_{i,j}^{(E)} = \sqrt{(x_{1,i} - x_{1,j})^2 + (x_{2,i} - x_{2,j})^2 + \cdots + (x_{k,i} - x_{k,j})^2}$$

- Correlation-Based Distance:

$$d_{i,j}^{(C)} = 1 - r_{i,j}^2$$

$$r_{i,j} = \frac{\sum_{m=1}^{k} (x_{i,m} - \bar{x}_m)(x_{j,m} - \bar{x}_m)}{\sqrt{\sum_{m=1}^{k} (x_{i,m} - \bar{x}_m)^2 \sum_{m=1}^{k} (x_{j,m} - \bar{x}_m)^2}}$$

- Mahalanobis Distance:

$$d_{i,j}^{(M)} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \hat{\mathbf{\Sigma}}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

- Single Linkage:

$$\min d(A_i, B_j)$$

- Complete Linkage:

$$\max d(A_i, B_j)$$

- Average Linkage:

$$\frac{\sum d(A_i, B_j)}{nm}$$

- Centroid Distance:

$$d(\bar{x}_A, \bar{x}_B)$$