**Numerical Summaries :**

*For a random sample,* $(y_1,...,y_n)$, *of size* $n$

- Sample Mean : $\hat{\mu}_Y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Sample Variance: $s_Y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}{n-1}$
- Sample Standard Deviation $= \sqrt{\text{Sample Variance}} = s_Y$

**Sampling Distribution of $\bar{Y}$ :**

- The mean and standard deviation of the sample mean, $\bar{Y}$, based on a random sample of size $n$, from a population with mean $\mu_Y$ and standard deviation $\sigma_Y$ are
  1. $mean(\bar{Y}) = \mu_{\bar{Y}} = \mu_Y$ (*unbiased estimator*)
  2. $S.D.(\bar{Y}) = \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$.
- If $Y \sim N(\mu_Y, \sigma_Y)$ then $\bar{Y} \sim N\left(\mu_{\bar{Y}} = \mu_Y,\ \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}\right)$.
- Central Limit Theorem:
  *If* $Y \sim ?(\mu_Y, \sigma_Y)$ then for large n,
  $$\bar{Y} \approx N\left(\mu_{\bar{Y}} = \mu_Y,\ \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}\right).$$

**General T - tools for "Parameter" :**

- For $H_0$ : Parameter $= H_0$ value :
  $$\text{Test-Statistic} = t_0^* = t = \frac{Estimate - H_0 \text{ value}}{SE(Estimate)}$$
- A $100(1-\alpha)\%$ Confidence Interval:
  $$Estimate \pm C.V. \times SE(Estimate)$$
  $$\Rightarrow C.V. = t^*_{\alpha/2,df}$$

**One Population Mean, $\mu$ :**

- $Estimate = \bar{y},\ SE(Estimate) = \frac{s}{\sqrt{n}}$
- $df = n-1$
- For $H_0 : \mu_Y = \mu_0$ :
  $$t = \frac{\bar{y}-\mu_0}{s\ /\sqrt{n}} = \frac{\bar{y}-\mu_0}{SE(\bar{y})}$$
- A $100(1-\alpha)\%$ C.I. for $\mu$:
  $$\bar{y} \pm t^*_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

**Two Means - Paired Samples, $\mu_d = \mu_1 - \mu_2$ :**

- $Estimate = \bar{d},\ SE(Estimate) = SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
- $s_d$ = the sample standard deviation of the differences
  $$= \sqrt{\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} d_i\right)^2}{n-1}}$$
- For $H_0 : \mu_d = d_0$
  $$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{\bar{d}}{SE(\bar{d})} \text{ OR } t = \frac{\bar{d}-\Delta_0}{s_d/\sqrt{n}} \quad df = n-1$$
- A $100(1-\alpha)\%$ C.I. for $\mu_d$ :
  $$\bar{d} \pm t^*_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right)$$

**Two Means - Independent Samples, $\mu_1 - \mu_2$ :**

- Assume that $\sigma_1^2 = \sigma_2^2$.
- $Estimate = \bar{y}_1 - \bar{y}_2$,
  $$SE(Estimate) = SE(\bar{y}_1 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$
  where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
- For $H_0 : \mu_1 - \mu_2 = 0$
  $$t = \frac{\bar{y}_1-\bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{y}_1-\bar{y}_2}{SE(\bar{y}_1-\bar{y}_2)} \quad df = n_1+n_2-2$$
- A $(1-\alpha)100\%$ C.I. for $\mu_1 - \mu_2$ :
  $$(\bar{y}_1-\bar{y}_2) \pm t^*_{\alpha/2,n_1+n_2-2}\left(s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

**Natural Log Transformations:**

- $\overline{LnY}_1 - \overline{LnY}_2$ estimates
  $$\ln[Median(Y_1)] - \ln[Median(Y_2)]$$
  $$= \ln\left[\frac{Median(Y_1)}{Median(Y_2)}\right]$$
- And $e^{\overline{LnY}_1 - \overline{LnY}_2}$ estimates $\left[\frac{Median(Y_1)}{Median(Y_2)}\right]$

## ANOVA for Several Means, $\mu_1, \mu_2, ..., \mu_k$ :

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ $(1 - Mean\ Model)$

$H_a$ : Not all means are equal $(k - Mean\ Model)$

- Test-Statistic

$$F = \frac{MS_{Treatment}}{MS_{Error}} = \frac{SS_{Treatment}/(k-1)}{SS_{Error}/(n-k)}$$

- $df = (k-1, n-k)$  $F \sim F^{k-1}_{n-k}$
- Sum of squares:
  - $SS_{Treatment} = SS_{\underline{B}etween} = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2 = \sum n_j(\bar{y}_j - \bar{\bar{y}})^2$
  - $SS_{Error} = SS_{\underline{W}ithin} = \sum\sum(y_{ij} - \bar{y}_j)^2$
  - $SS_{Total} = \sum\sum(y_{ij} - \bar{\bar{y}})^2 = SS_{Treatment} + SS_{Error}$

**Extra-Sum-of-Squares F-test:**

$H_0$: (r)reduced model

$H_a$: (f)full model

- Extra SS = $SS_E$(reduced) - $SS_E$(full)

  Extra $df$ = $df_E$ (reduced) – $df_E$ (full)

- $F = \dfrac{(Extra\ SS)/(Extra\ df)}{SS_E(\text{Full})/df_E(Full)}$

  $= \dfrac{[SS_E(reduced) - SS_E(full)]/[df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$

- $df = [Extra\ df, df_E(Full)] = [Extra\ df, n-k]$
- $F \sim F^{df(r)-df(f)}_{df(f)} = F^{extra\ \text{df}}_{df(f)}$

**Multiple-Comparisons:**

- Pairwise comparisons ($m$):

$m = \dfrac{k(k-1)}{2}$, where $k$ = number of means

- **Tukey** Multiple Comparisons:

  Confidence interval for the difference, $\boldsymbol{\mu_i - \mu_j}$

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_\alpha}{\sqrt{2}} \times \sqrt{MS_E}\sqrt{(1/n_i) + (1/n_j)}$$

$df = (k, n-k)$

- **Bonferroni's Method**

  Individual comparison-wise error rate ($\alpha_I$) based on the experiment-wise (or family-wise) error rate ($\alpha_F$):

$$\alpha_I = \frac{\alpha_F}{m} \qquad df = n - k$$

$$ME_{ij} = t_{\alpha_I/2, n-k} \times \sqrt{MS_E}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$\mu_i - \mu_j \neq 0$, **if** $\left|\bar{y}_i - \bar{y}_j\right| \geq ME_{ij}$

**Linear Combinations of Group Means:**

- *Parameter*: $\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_k\mu_k$
- *Estimate:* $\hat{\gamma} = C_1\bar{y}_1 + C_2\bar{y}_2 + \cdots + C_k\bar{y}_k$
- $SE(\hat{\gamma}) = s_p\sqrt{\dfrac{C_1^2}{n_1} + \dfrac{C_2^2}{n_2} + ... + \dfrac{C_k^2}{n_k}}$

  where $s_p = \sqrt{MS_E} = \sqrt{\dfrac{(n_1 - 1)s_1^2 + ... + (n_k - 1)s_k^2}{n - k}}$

- $t = \dfrac{\hat{\gamma} - 0}{SE(\hat{\gamma})}$, $df = n - k$
- A (1 – $\alpha$)100% CI for $\gamma$ :
- $\hat{\gamma} \pm t^*_{\alpha/2, n-k} \times SE(\hat{\gamma})$

**Kruskal-Wallis Test:**

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

$H_a : \mu_1, \mu_2, ..., \mu_k$ *(Not all equal)*

Test Statistic = $H = \dfrac{12}{n(n+1)}\displaystyle\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(n+1)$

Where n = total number of observations

$n_1, n_2, ..., n_k$ denote sample sizes of $k$ samples

$R_1, R_2, ..., R_k$ denote the sums of the ranks

Critical value of H is $\chi^2_\alpha$ with $df = k - 1$

**Simple Linear Regression (SLR):**

Model: $\mu(Y\mid X)=\beta_0+\beta_1 X$

Estimated model:

$$\hat{y}=\hat{\mu}(Y\mid X)=\hat{\beta}_0+\hat{\beta}_1 x$$

Slope is: $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$

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

Standard error of the model ($s_e$):

$$s_e=\hat{\sigma}=\sqrt{\frac{\sum e^2}{n-2}}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SS_{ERROR}}{n-2}}=\sqrt{MS_{ERROR}}$$

**Inferences in SLR (*t*-Procedures):**

Standard error of the Slope:

$$SE(\hat{\beta}_1)=\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Regression t-statistic:

$$t=\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}}=\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \qquad df=n-2$$

Confidence interval for the slope:

$$\hat{\beta}_1\pm t_{\alpha/2,n-2}\times SE(\hat{\beta}_1)$$

Confidence interval for the mean response of *y* for a given *x*:

$$\hat{y}_p=\hat{\beta}_0+\hat{\beta}_1 x_p$$

$$\hat{y}_p\pm t_{\alpha/2,n-2}\times\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

Where $S_{xx}=(n-1)s_x^2$

Prediction Interval for all single observation responses of *y* for a given *x*:

$$\hat{y}_p\pm t_{\alpha/2,n-2}\times\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{S_{xx}}}$$

**Regression Identity in SLR:**

$$SS_{TOTAL}=S_{yy}=SS_{REGR}+SS_{ERROR}$$

**ANOVA *F*-test for significance of slope in SLR:**

$$F=\frac{MS_{REGR}}{MS_{ERROR}}=\frac{SS_{REGR}/(2-1)}{SS_{ERROR}/(n-2)}$$

Where $df=(1,n-2)$ or $F^1_{n-2}$

**Coefficient of Determination ($R^2$) in SLR and MLR:**

$$R^2=\frac{SS_{REGR}}{SS_{TOTAL}}=1-\frac{SS_{Error}}{SS_{TOTAL}}=\frac{SS_{TOTAL}-SS_{Error}}{SS_{TOTAL}}$$

$$R^2_{adj}=1-\frac{MS_{ERROR}}{MS_{TOTAL}}$$

**Linear Correlation coefficient (r):**

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum(x_i-\bar{x})^2\right]\left[\sum(y_i-\bar{y})^2\right]}}$$

Where $df=n-2$

Also, $r=\sqrt{R^2}$

but may be – or + depending on the relationship

**Interpretation of Model Effects in SLR after Log Transformation:**

"ln" = the natural logarithm

1. $\mu(\ln(Y)\mid X)=\beta_0+\beta_1 X$

$k=$ Final - Initial $\Rightarrow$ Apply: $e^{k\beta_1}$

2. $\mu(Y\mid\ln(X))=\beta_0+\beta_1\ln(X)$

$k=\frac{\text{Final}}{\text{Initial}}$ $\Rightarrow$ Apply: $\beta_1\ln(k)$

3. $\mu(\ln(Y)\mid\ln(X))=\beta_0+\beta_1\ln(X)$

$k=\frac{\text{Final}}{\text{Initial}}$ $\Rightarrow$ Apply: $k^{\beta_1}$

**Multiple Linear Regression (MLR):**

General model:

$$\hat{\mu}(y\mid x)=\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots+\hat{\beta}_kx_k$$

Where:

$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2\ldots\hat{\beta}_k\Rightarrow$ Regression Coefficients

$\hat{\beta}_1,\hat{\beta}_2\ldots\hat{\beta}_k\Rightarrow$ Partial slopes

Standard error of the model:

$$s_e=\hat{\sigma}=\sqrt{MS_{ERROR}}=\sqrt{\frac{SS_{ERROR}}{n-(k+1)}}$$

**ANOVA Test for the Overall MLR Model:**

$$F=\frac{MS_{REGR}}{MS_{ERROR}}=\frac{SS_{REGR}/k}{SS_{ERROR}/(n-(k+1))}$$

$$df=(k,n-(k+1))$$

Where *k* = number of predictor variables

**Inferences for the Usefulness of Single Predictor Variables (Coefficients):**

Multiple regr. t-test for the significance of a slope:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \qquad df = n-(k+1)$$

Confidence interval for the slope:

$$\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} \times SE(\hat{\beta}_i)$$

Confidence interval for the mean response of *y* for given $x_1, x_2, \ldots x_k$:

"Fit" ± Critical value x SE(Fit)

OR $\hat{y}_p \pm t_{\alpha/2,n-(k+1)} \times SE(Fit)$

Prediction Interval for all single observation responses of *y* for given $x_1, x_2, \ldots x_k$:

"Fit" ± Critical value x $\sqrt{MS_{ERROR} + [SE(Fit)]^2}$

OR $\hat{y}_p \pm t_{\alpha/2,n-(k+1)} \times \sqrt{\hat{\sigma}^2 + [SE(Fit)]^2}$

**Inference for a Subset of Predictor Variables:**

Extra-Sum-of-Squares F-Test for selected slopes:

$H_0$: all selected beta's equal zero (reduced model)

$H_a$: not all are equal to zero (full model)

(See Extra-SS F-Test above for formula)

$$df = (\text{Number of selected } \beta_i' s, n-(k+1))$$

**Two-Factor ANOVA (With Interaction) (Non-Additive Model):**

F (Overall model) = F (Corrected model)

$$= \frac{\text{Corrected } SS / \text{Corrected } df}{\text{Error } SS / \text{Error } df}$$

$$= \frac{\text{Corrected } SS / (ab-1)}{\text{Error } SS / (n-ab)} = \frac{\text{Corrected } MS}{\text{MSE}}$$

Where $df = [(ab-1),(n-ab)]$

F-statistic for Factor A:

$$F_A = \frac{SSA/(a-1)}{SSE/(n-ab)} = \frac{MSA}{MSE}$$

F-statistic for Factor B:

$$F_B = \frac{SSB/(b-1)}{SSE/(n-ab)} = \frac{MSB}{MSE}$$

F-statistic for AB Interaction:

$$F_{AB} = \frac{SSAB/(a-1)(b-1)}{SSE/(n-ab)} = \frac{MSAB}{MSE}$$

Where: a = number of levels of Factor A
b = number of levels of Factor B
n = total number of observations
= a x b x (no. of replicates)

**Two-Factor ANOVA (Without Interaction) (Additive Model):**

F (Overall model) = F (Corrected model)

$$= \frac{\text{Corrected } SS / \left[(a-1)+(b-1)\right]}{SSE / \text{Error } df}$$

$$= \frac{\text{Corrected } MS}{MSE}$$

F-statistic for Factor A:

$$F_A = \frac{SSA/(a-1)}{SSE/\text{Error } df} = \frac{MSA}{MSE}$$

F-statistic for Factor B:

$$F_B = \frac{SSB/(b-1)}{SSE/\text{Error } df} = \frac{MSB}{MSE}$$

Where: Error $df = (n-1)-(a-1)-(b-1)$
a = number of levels of Factor A
b = number of levels of Factor B
n = total number of observations
= a x b x (no. of replicates)

F-statistic for AB Interaction can also be performed by comparing the Additive Model with the Non-additive model using an Extra-Sum-of-Squares F-test:

$H_0$ : *additive* model (no interaction)(reduced model)

$$\mu(Y \mid X_1, X_2) = \beta_0 + X_1 + X_2$$

$H_a$ : *non – addtive* model (interaction)(full model)

$$\mu(Y \mid X_1, X_2) = \beta_0 + X_1 + X_2 + (X_1 \times X_2)$$

$$F = \frac{[SS_E(reduced) - SS_E(full)] / [df_E(reduced) - df_E(full)]}{SS_E(full)/df_E(full)}$$

$$df = [Extra\ df, df_E(Full)]$$

**Randomized Block ANOVA**

$$F_{TREATMENT} = \frac{SSTR/(k-1)}{SSE/(k-1)(b-1)} = \frac{MSTR}{MSE}$$

$$df = [(k-1),(k-1)(b-1)]$$

$$F_{Blocks} = \frac{SSBL/(b-1)}{SSE/(k-1)(b-1)} = \frac{MSBL}{MSE}$$

$$df = [(b-1),(k-1)(b-1)]$$