Outline

1. Analysis of Variance
   - Comparing Two Independent Samples
   - Comparing Multiple Independent Samples
## Comparing Two Independent Samples: Normal Case

### Data/Model:

- \( x_1, x_2, \ldots, x_{n_1} \) i.i.d. \( N(\mu_x, \sigma^2) \)
- \( y_1, y_2, \ldots, y_{n_2} \) i.i.d. \( N(\mu_y, \sigma^2) \)

### Regression model specification:

\[
\begin{align*}
\mathbf{y}^* &= \mathbf{X}^* \mathbf{\beta}^* + \mathbf{e}^* \\
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_{n_1} \\
  y_1 \\
  y_2 \\
  \vdots \\
  y_{n_2}
\end{bmatrix} &= \begin{bmatrix}
  1 & 0 \\
  1 & 0 \\
  \vdots & \vdots \\
  1 & 0 \\
  0 & 1 \\
  0 & 1 \\
  \vdots & \vdots \\
  0 & 1
\end{bmatrix}
\begin{bmatrix}
e_1 \\
e_2 \\
\vdots \\
e_{n_1} \\
e_{n_1+1} \\
e_{n_1+2} \\
\vdots \\
e_{n_1+n_2}
\end{bmatrix} \\
\mathbf{\beta}^* &= \begin{bmatrix}
\mu_x \\
\mu_y
\end{bmatrix}
\end{align*}
\]
Comparing Two Independent Normal Samples

- **Least-Squares /ML Estimates of $\beta^*$**
  
  $$\hat{\beta}^* = [(X^*)^T X^*]^{-1}(X^*)^T y^*$$
  
  $$= \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1 \bar{x} \\ n_2 \bar{y} \end{bmatrix} = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}$$

- **Unbiased Estimate of $\sigma^2$**
  
  $$SS_{ERR} = \sum (y_i^* - \hat{y}_i^*)^2 = \sum_1^n (x_i - \bar{x})^2 + \sum_1^n (y_i - \bar{y})^2$$

  $$\sim \sigma^2 \chi^2_{n_1-1} + \sigma^2 \chi^2_{n_2-1} \quad \text{(independent)}$$

  $$\sim \sigma^2 \chi^2_{n_1+n_2-2}$$

  $$\Rightarrow \hat{\sigma}^2 = \frac{SS_{ERR}}{(n_1 + n_2 - 2)} \quad \text{“pooled est“}$$

- **Two-Sample $t$-test of $H_0: \mu_x = \mu_y$**
  
  $$\bar{x} - \bar{y} \sim N(0, 0, \sigma^2[\frac{1}{n_1} + \frac{1}{n_2}])$$

  and $\hat{\sigma}^2$ indep.

  so $t = \frac{(\bar{x} - \bar{y})/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\hat{\sigma}} \sim t$ with $df = (n_1 + n_2 - 2)$
Comparing Two Independent Normal Samples

Regression Model Implementation of Two-Sample \( t \)-Test

Regression model specification:

\[
y^* = X^{**} \beta^{**} + e^*
\]

\[
y^* = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1} \\ y_1 \\ y_2 \\ \vdots \\ y_{n_2} \end{bmatrix} \quad X^{**} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \quad e^* = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n_1} \\ e_{n_1+1} \\ e_{n_1+2} \\ \vdots \\ e_{n_1+n_2} \end{bmatrix} \quad \beta^{**} = \begin{bmatrix} \mu_x \\ \mu_y - \mu_x \end{bmatrix}
\]

Note: \( \hat{\beta}^{**} \) estimates \( \mu_x - \mu_y \) directly
Comparing Two Independent Normal Samples

Example 12.1.A: Kirchhoefer (1979) data

- Measurement of chlorpheniramine maleate in tablets
- Nominal dosage equal to 4mg.
- 7 Labs
- 10 Measurements Per Lab

Two-Lab Comparison: \textit{RProject11_Tables_TwoSampleT.r}

- Two-Sample \textit{t}-Test
  - Custom R function: \textit{fcn.TwoSampleTTest()}
  - Built-in R function: \textit{t.test()}
- Regression model implementation of \textit{t}-Test
  - Built-in R function: \textit{lm()}
  ("t value" for slope in simple linear regression)
1. Analysis of Variance
   - Comparing Two Independent Samples
   - Comparing Multiple Independent Samples
Comparing Multiple Independent Samples: Normal Case

- **Data/Model:**
  
  \[ y_{1,1}, y_{1,2}, \ldots, y_{1,J} \text{ i.i.d. } N(\mu_1, \sigma^2) \]
  
  \[ y_{2,1}, y_{2,2}, \ldots, y_{2,J} \text{ i.i.d. } N(\mu_2, \sigma^2) \]
  
  \[ \vdots \]
  
  \[ y_{I,1}, y_{I,2}, \ldots, y_{I,J} \text{ i.i.d. } N(\mu_I, \sigma^2) \]

- **One-Way ANOVA Model**
  
  \[ y_{i,j} = \mu + \alpha_i + e_{i,j} \]

- \( I \) groups (\( i = 1, 2, \ldots, I \))
- \( J \) independent observations for each group \( i \).
- Re-parametrize sample parameters
  
  \[ \mu = \bar{\mu} = \frac{1}{I} \sum_{i=1}^{I} \mu_i / I \]
  
  \[ \alpha_i = \mu_i - \bar{\mu}, \; i = 1, 2, \ldots, I \; (\text{Constraint: } \alpha_i = 0) \]

- Regression errors/residuals
  
  \[ e_{i,j} \text{ i.i.d. } N(0, \sigma^2). \]
One-Way ANOVA Model

- Least-Squares / ML Estimation of ANOVA Model
  \[
  \hat{\mu} = \bar{y}_. = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} y_{i,j}
  \]
  \[
  \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_. = \frac{1}{J} \sum_{j=1}^{J} y_{i,j}
  \]

- Unbiased Estimation of \( \sigma^2 \)
  - The “Within-Group Sum-of-Squares” for each group \( i \) has distribution
  \[
  \sum_{j=1}^{J} (y_{i,j} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{J-1}
  \]
  - Because the groups are independent the sum has distribution
    \[
    SS_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{i,j} - \bar{y}_{i.})^2 \sim \sigma^2 \chi^2_{df_W}
    \]
    with degree of freedom: \( df_W = I \times (J - 1) \)
  - \( \hat{\sigma}^2 = SS_W / df_W \) is unbiased and independent of \( \hat{\mu} \) and of all \( \hat{\alpha}_i \)
    Note: \( \hat{\sigma}^2 \) is average of I independent within-group estimates

- One-Way ANOVA Null Hypothesis:
  \[
  H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0.
  \]
Under $H_0$:

- The null hypothesis $H_0$ is equivalent to
  \[ H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu \] (any fixed value)

- The group means are i.i.d.
  \[ \bar{y}_{i,:} \sim N(\mu, \sigma^2/J), \; i = 1, \ldots, I \]

- Treating these as a sample of size $I$, their sample variance
  \[ \frac{\sum_{i=1}^{I} (\bar{y}_{i,:} - \bar{y}_{:,:})^2}{I - 1} \]
  has expectation $\sigma^2/J$ so
  \[ \tilde{\sigma}^2 = J \times \frac{\sum_{i=1}^{I} (\bar{y}_{i,:} - \bar{y}_{:,:})^2}{I - 1} \]
  is an unbiased estimate of $\sigma^2$ which is independent of $\hat{\sigma}^2$.

- Under $H_0$ the statistic
  \[ \hat{F} = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \sim F_{df_1,df_2} \]
  an $F$ distribution with degrees of freedom:
  \[ df_1 = (I-1) \] and \[ df_2 = I(J - 1). \]
One-Way Anova

- **Sum-of-Squares Identity**
  \[ SS_{TOT} = \sum_{i} \sum_{j} (y_{i,j} - \bar{y}_{..})^2 \]
  \[ = \sum_{i} \sum_{j} [(y_{i,j} - \bar{y}_{i..}) + (\bar{y}_{i..} - \bar{y}_{..})]^2 \]
  \[ = \sum_{i} \sum_{j} [(y_{i,j} - \bar{y}_{i..})^2 + (\bar{y}_{i..} - \bar{y}_{..})^2] + 0 \]
  \[ = \sum_{i} \sum_{j} (y_{i,j} - \bar{y}_{i..})^2] + [J \sum_{i} (\bar{y}_{i..} - \bar{y}_{..})^2] \]
  \[ = SS_W + SS_B \]

  - \( SS_W = I(J - 1)\hat{\sigma}^2 \) (Within-Group SS)
  - \( SS_B = (I - 1)\tilde{\sigma}^2 \) (Between-Group SS)

- **Independent Mean Squares**
  - \( MS_W = SS_W/df_2 = \hat{\sigma}^2 \) (Within-Group Mean-Square)
  - \( MS_B = SS_B/df_1 = \tilde{\sigma}^2 \) (Between-Group Mean-Square)

- **F Test Statistic for \( H_0 \)**
  - \( \hat{F} = MS_B/MS_W = \tilde{\sigma}^2/\hat{\sigma}^2 \)
  - Under \( H_0 \): \( \hat{F} \sim F_{df_1,df_2} \)
## Analysis of Variance

### One-Way ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>$df_B = I - 1$</td>
<td>$SS_B$</td>
<td>$MS_B = SS_B / df_B$</td>
<td>$F = \frac{MS_B}{MS_W}$</td>
</tr>
<tr>
<td>Within Groups</td>
<td>$df_W = I(J - 1)$</td>
<td>$SS_B$</td>
<td>$MS_B = SS_B / df_W$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1 = IJ - 1$</td>
<td>$SS_{TOT}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 12.1.A: Kirchhoefer (1979) data
- Measurement of chlorpheniramine maleate in tablets
- Nominal dosage equal to 4mg.
- 7 Labs
- 10 Measurements Per Lab

R Script: *RProject11_Tables_OneWayAnova.r*

- Built-in R function: `aov()`
  - Define *factor* variable in *R* to distinguish groups/Labs
  - Summary table: “Analysis of Variance” with $F$-statistic
  - Display tables of means with R function `model.tables()`
  - Validation of standard error for difference of means

- Multiple Comparisons: simultaneous confidence intervals
  - R function: `TukeyHSD()`
    (Tukey’s “Honest significant Difference”)


R Script: *RProject11_Tablets_OneWayAnova.r*

- Using linear regression to implement ANOVA F Test
- Residual standard error from `lm()` equals `sigma` from `aov()` equals root Mean Sq residuals for both.
- $F$ statistics, $p$-values are identical.