Analysis of Variance

MIT 18.443

Dr. Kempthorne

Spring 2015

Outline

1. Analysis of Variance
   - Comparing Two Independent Samples
   - Comparing Multiple Independent Samples

Comparing Two Independent Samples: Normal Case

- Data/Model:

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

- Regression model specification:

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}^* + \mathbf{e}^*$$

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

Comparing Two Independent Normal Samples

- Least-Squares /ML Estimates of β^*

$$\begin{aligned}\hat{\beta}^* &= [(\mathbf{X}^*)^T\mathbf{X}^*]^{-1}(\mathbf{X}^*)^T\mathbf{y}^* \\ &= \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} n_1\overline{x} \\ n_2\overline{y} \end{bmatrix} = \begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix} \\ &\sim N_2\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \sigma^2 \begin{bmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{bmatrix} \right)\end{aligned}$$

- Unbiased Estimate of σ^2

$$\begin{aligned} SS_{ERR} &= \sum(y_i^* - \hat{y}_i^*)^2 = \sum_1^n(x_i - \overline{x})^2 + \sum_1^n(y_i - \overline{y})^2 \\ &\sim \sigma^2\chi^2_{n_1-1} + \sigma^2\chi^2_{n_2-1} \ \ (independent) \\ &\sim \sigma^2\chi^2_{n_1+n_2-2} \\ \Longrightarrow \hat{\sigma}^2 &= SS_{ERR}/(n_1+n_2-2) \ \ \text{“pooled est”} \end{aligned}$$

- Two-Sample t-test of $H_0 : \mu_x = \mu_y$

$$\overline{x} - \overline{y} \sim N(0, 0, \sigma^2[\frac{1}{n_1} + \frac{1}{n_2}]) \text{ and } \hat{\sigma}^2 \text{ indep.}$$

so $t = \frac{(\overline{x}-\overline{y})/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\hat{\sigma}} \sim t$ with $df = (n_1 + n_2 - 2)$

Comparing Two Independent Normal Samples

Regression Model Implementation of Two-Sample *t*-Test

Regression model specification:

$$\mathbf{y}^* = \mathbf{X}^{**}\beta^{**} + \mathbf{e}^*$$

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

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: *RProject11_Tablets_TwoSampleT.r*

- Two-Sample t-Test
  - Custom R function: *fcn.TwoSampleTTest*()
  - Built-in R function: *t.test*()
- Regression model implementation of t-Test
  - Built-in R function: *lm*()
    ("t value" for slope in simple linear regression)

Outline

1. Analysis of Variance
   - Comparing Two Independent Samples
   - Comparing Multiple Independent Samples

Comparing Multiple Independent Samples: Normal Case

- Data/Model:

$$
\begin{array}{l}
y_{1,1}, y_{1,2}, \ldots, y_{1,J} \text{ i.i.d. } N(\mu_1, \sigma^2) \\
y_{2,1}, y_{2,2}, \ldots, y_{2,J} \text{ i.i.d. } N(\mu_2, \sigma^2) \\
\vdots \\
y_{\mathrm{I},1}, y_{\mathrm{I},2}, \ldots, y_{\mathrm{I},J} \text{ i.i.d. } N(\mu_{\mathrm{I}}, \sigma^2)
\end{array}
$$

- One-Way ANOVA Model

$$y_{i,j} = \mu + \alpha_i + e_{i,j}$$

  - I groups $(i = 1, 2, \ldots, \mathrm{I})$
  - J independent observations for each group i.
  - Re-parametrize sample parameters

$$
\begin{array}{l}
\mu = \overline{\mu} = \sum_{i=1}^{\mathrm{I}} \mu_i/\mathrm{I} \\
\alpha_i = \mu_i - \overline{\mu},\ i = 1, 2, \ldots, \mathrm{I}\ (\mathit{Constraint}: \sum_i \alpha_i = 0)
\end{array}
$$

  - Regression errors/residuals

$$e_{i,j}\ \mathit{i.i.d.}\ N(0, \sigma^2).$$

One-Way ANOVA Model

- Least-Squares / ML Estimation of ANOVA Model

$$\hat{\mu} = \overline{y_{..}} = \quad {}_{i} \quad {}_{j}\, y_{i,j}/\mathrm{I}J$$

$$\hat{\alpha}_i = \overline{y_{i.}} - \overline{y_{..}} = \quad {}_{j}\, y_{i,j}/J - \overline{y_{..}}$$

- Unbiased Estimation of σ^2
  - The “Within-Group Sum-of-Squares” for each group i has distribution

$$ {}_{j=1}^{J}(y_{i,j} - \overline{y_{i.}})^2 \sim \sigma^2 \chi^2_{J-1}$$

  - Because the groups are independent the sum has distribution

$$SS_W = \quad {}_{i=1}^{\mathrm{I}} \quad {}_{j=1}^{J}(y_{i,j} - \overline{y_{i.}})^2 \sim \sigma^2 \chi^2_{df_W}$$

    with degress of freedom: $df_W = \mathrm{I} \times (J-1)$
  - $\hat{\sigma}^2 = SS_W/df_W$ 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:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_{\mathrm{I}} = 0.$$

One-Way ANOVA: Testing H_0

Under H_0:

- The null hypothesis H_0 is equivalent to
  $$H_0 : \mu_1 = \mu_2 = \cdots = \mu_{\mathrm{I}} \equiv \mu \text{ (any fixed value)}$$
- The group means are i.i.d.
  $$\overline{y_{i,\cdot}} \sim N(\mu, \sigma^2/J),\ i = 1, \ldots, \mathrm{I}$$
- Treating these as a sample of size I, their sample variance
  $$\sum_{i=1}^{\mathrm{I}} (\overline{y_{i,\cdot}} - \overline{y_{\cdot,\cdot}})^2/(\mathrm{I}-1)$$
  has expectation σ^2/J so
  $$\tilde{\sigma}^2 = J \times \sum_{i=1}^{\mathrm{I}} (\overline{y_{i,\cdot}} - \overline{y_{\cdot,\cdot}})^2/(\mathrm{I}-1)$$
  is an unbiased estimate of σ^2 which is independent of $\hat{\sigma}^2$.
- Under H_0 the statistic
  $$\hat{F} = \tilde{\sigma}^2/\hat{\sigma}^2 \sim F_{df_1, df_2}$$
  an F distribution with degrees of freedom:
  $$df_1 = (\mathrm{I}\text{-}1) \text{ and } df_2 = \mathrm{I}(J-1).$$

One-Way Anova

- **Sum-of-Squares Identity**

$$
\begin{aligned}
SS_{TOT} &= \quad {}_{i} \quad {}_{j}(y_{i,j} - \overline{y_{\cdot\cdot}})^2 \\
&= \quad {}_{i} \quad {}_{j}[(y_{i,j} - \overline{y_{i\cdot}}) + (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})]^2 \\
&= \quad {}_{i} \quad {}_{j}[(y_{i,j} - \overline{y_{i\cdot}})^2 + (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2] + 0 \\
&= [\quad {}_{i} \quad {}_{j}(y_{i,j} - \overline{y_{i\cdot}})^2] + [J \quad {}_{i}(\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2] \\
&= SS_W + SS_B
\end{aligned}
$$

  - $SS_W = \mathrm{I}(J-1)\hat{\sigma}^2$ (Within-Group SS)
  - $SS_B = (\mathrm{I}-1)\tilde{\sigma}^2$ (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: $\hat{F} \sim F_{df_1, df_2}$

One-Way ANOVA

ANOVA Table

Source	df	SS	MS	F
Between Groups	$df_B = \mathrm{I} - 1$	SS_B	$MS_B = SS_B/df_B$	$F = \dfrac{MS_B}{MS_W}$
Within Groups	$df_W = \mathrm{I}(J-1)$	SS_B	$MS_B = SS_B/df_W$	
Total	$n - 1 = \mathrm{I}J - 1$	SS_{TOT}		

Comparing Multiple 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

R Script: *RProject11_Tablets_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")

Comparing Multiple Independent Normal Samples

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.

18.443 Statistics for Applications
Spring 2015