Analysis of Variance
Bios 662

Michael G. Hudgens, Ph.D.
mhudgens@bios.unc.edu

http://www.bios.unc.edu/~mhudgens

2008-10-21 13:34
Outline

• Introduction
• Alternative models
• SS decomposition
• Example w/ SAS, R
Analysis of Variance Model

• Ch 10 text (skip 10.3-10.5); Ch 12

• How do we test hypotheses about the mean of more than 2 groups? Analysis of variance (ANOVA) model

• Definition 10.1 An analysis of variance model is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the levels of the variable.

• Categorical predictor variables are also called qualitative factors
Notation

• Let $Y_{ij}$ be the $j^{th}$ observation in the $i^{th}$ group

• $i = 1, \ldots, K$; $j = 1, \ldots, n_i$

• Let $N = \sum_{i=1}^{K} n_i$

• $\bar{Y}_i = \sum_{j} Y_{ij} / n_i$
ANOVA Model and Hypotheses

• Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$

• Want to test

\[ H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \]

versus

\[ H_A : \text{at least one} \neq \]
Two variance estimators

- The pooled estimate of $\sigma^2$ is:

$$s_p^2 = \frac{\sum_{i=1}^{K} (n_i - 1) s_i^2}{\sum_{i=1}^{K} (n_i - 1)}$$

- Under $H_0$, the (weighted) variance of the $\bar{Y}_i$'s will estimate $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i (\bar{Y}_i - \bar{Y})^2}{K - 1}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij}}{N}$$
ANOVA: F test

• It can be shown under $H_0$:

\[(N - K)s_p^2/\sigma^2 \sim \chi^2_{N-K}\]
\[(K - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{K-1}\]

and $s_p^2$ and $\hat{\sigma}^2$ are independent

• Therefore, under $H_0$,

\[F \equiv \frac{\hat{\sigma}^2}{s_p^2} \sim F_{K-1, N-K}\]
ANOVA: F test

• To test $H_0$,

$$C_\alpha = \{ F : F > F_{1-\alpha;K-1,N-K} \}$$

• The test uses $F > F_{1-\alpha;K-1,N-K}$ because under $H_A$,

$$E(\hat{\sigma}^2) > E(s_p^2)$$

• In particular, $E(s_p^2) = \sigma^2$ whereas

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{\sum_i n_i (\mu_i - \mu)^2}{K - 1}$$

where $\mu$ is the overall mean defined in equation (1) below
ANOVA: Example

- Ex: Passive smoking and lung function
- A study was conducted to compare the lung function of groups of smokers and non-smokers. Lung function was measured by forced mid-expiratory flow (FEF)
ANOVA: Example

FEF for males by smoking status

<table>
<thead>
<tr>
<th>Group</th>
<th>$n_i$</th>
<th>Mean (L/sec)</th>
<th>sd (L/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-smokers</td>
<td>200</td>
<td>3.78</td>
<td>0.79</td>
</tr>
<tr>
<td>Passive smokers</td>
<td>200</td>
<td>3.30</td>
<td>0.77</td>
</tr>
<tr>
<td>Noninhalers</td>
<td>50</td>
<td>3.32</td>
<td>0.86</td>
</tr>
<tr>
<td>Light smk.</td>
<td>200</td>
<td>3.23</td>
<td>0.78</td>
</tr>
<tr>
<td>Mod. smk.</td>
<td>200</td>
<td>2.73</td>
<td>0.81</td>
</tr>
<tr>
<td>Heavy smk.</td>
<td>200</td>
<td>2.59</td>
<td>0.82</td>
</tr>
</tbody>
</table>
ANOVA: Example

\[ C_{.05} = \{ F > F_{5,1044,.95} = 2.22 \} \]

\[ s_p^2 = \frac{199(.79)^2 + 199(.77)^2 + \cdots + 199(.82)^2}{1044} = 0.636 \]

\[ \hat{\sigma}^2 = \frac{200(3.78 - 3.14)^2 + \cdots + 200(2.59 - 3.14)^2}{5} = 36.875 \]

- \[ F = 36.875/0.636 = 58.0; \text{ Reject } H_0 \]
Aside: Obtaining Quantiles/CDFs

- In R

  ```
  > qf(.95,5,1044)
  [1] 2.222674
  
  > pf(2.222674,5,1044)
  [1] 0.95
  ```

- In SAS

  ```
  data;
  y = finv(.95,5,1044);
  y1 = quantile('F',.95,5,1044);
  fy = cdf('F',2.22267,5,1044);
  
  proc print;
  Obs    y       y1       fy
  1   2.22267  2.22267  0.95000
  ```
Cell Means Model

- Heretofore, we have looked at ANOVA model

\[ Y_{ij} = \mu_i + \epsilon_{ij} \]

for \( i = 1, 2, \ldots, K \); \( j = 1, 2, \ldots, n_i \) where

\[ \epsilon_{ij} \sim N(0, \sigma^2) \]

for all \( i, j \)
Factor Effects Model

- Alternatively, an equivalent model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

for $$i = 1, 2, \ldots, K; j = 1, 2, \ldots, n_i$$ where

$$\mu = \frac{\sum_{i=1}^{K} n_i \mu_i}{N} \quad (1)$$

$$\alpha_i = \mu_i - \mu$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2)$$ for all $$i, j$$
Factor Effects Model

- Note typo in text page 363
- Constraint

\[ \sum_{i=1}^{K} n_i \alpha_i = 0 \]

- \( \alpha_i \) does not denote type I error
Model Equivalency

- Equivalency of null hypotheses

\[ H_0 : \mu_1 = \cdots = \mu_K \iff H_0 : \alpha_i = 0; i = 1, 2, \ldots, K \]

- \( \alpha_i \) is called the \( i^{th} \) main effect or factor effect

\[
Y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij} \\
= \mu + \alpha_i + \epsilon_{ij} \\
= \text{mean} + i^{th} \text{ main effect} + \text{error}
\]

- Data can be partitioned similarly

\[
Y_{ij} = \bar{Y} + (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \\
= \bar{Y} + a_i + e_{ij}
\]
ANOVA: Sum of Squares

• It can be shown (below)

\[
\sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2
\]

• That is

\[
SST = SSA + SSW
\]

\[
= (K - 1)\hat{\sigma}^2 + (N - K)s_p^2
\]
ANOVA: Sum of Squares

- Expected value of sum of squares

\[ E\left\{ \sum_{i=1}^{K} n_i \left( \overline{Y}_i - \overline{Y} \right)^2 \right\} = \sum_{i=1}^{K} n_i \alpha_i^2 + (K - 1)\sigma^2 \]

\[ E\left\{ \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left( Y_{ij} - \overline{Y}_i \right)^2 \right\} = (N - K)\sigma^2 \]

- Under \( H_0 : \alpha_1 = \cdots = \alpha_K = 0 \),

\[ E\left\{ \sum_{i=1}^{K} n_i \left( \overline{Y}_i - \overline{Y} \right)^2 \right\} = (K - 1)\sigma^2 \]
ANOVA: F test

• Therefore, under $H_A$: at least one $\alpha_i \neq 0$,

$$E(F) > 1$$

• I.e. we reject $H_0$ if $F$ is too large

$$C_{\alpha} = \{ F : F > F_{1-\alpha;K-1,N-k} \}$$
## ANOVA Table

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among groups</td>
<td>$K - 1$</td>
<td>$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_i - \bar{Y})^2}{K-1}$</td>
<td>MSA/MSW</td>
</tr>
<tr>
<td>Within groups</td>
<td>$N - K$</td>
<td>$s^2_p = \frac{\sum_{i=1}^{K} (n_i - 1)s^2_i}{N-K}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$N - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANOVA: Sum of Squares Proof

• Start with
\[
\sum_{ij}(Y_{ij} - \bar{Y})^2 = \sum_{ij}(Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2
\]

• RHS equivalent to
\[
\sum_{ij}(Y_{ij} - \bar{Y}_i)^2 + \sum_{ij}(\bar{Y}_i - \bar{Y})^2 + 2 \sum_{ij}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}).
\]

• Last term can be written as
\[
2 \sum_i \{(\bar{Y}_i - \bar{Y}) \sum_j (Y_{ij} - \bar{Y}_i)\},
\]
which equals zero since
\[
\sum_j (Y_{ij} - \bar{Y}_i) = 0
\]
for all \(i\).
ANOVA: E(SSW) Proof

\[ E(SSW) = E\{\sum_{ij}(Y_{ij} - \bar{Y}_i)^2\} \]

\[ = E \left\{ \sum_i (n_i - 1) \frac{\sum_j (Y_{ij} - \bar{Y}_i)^2}{n_i - 1} \right\} \]

\[ = \sum_i (n_i - 1) E(s_i^2) \]

\[ = \sum_i (n_i - 1) \sigma^2 \]

\[ = (N - K) \sigma^2 \]
Table 10.1: Distribution of ages (in months) at which infants first walked alone

<table>
<thead>
<tr>
<th>Active Group</th>
<th>Passive Group</th>
<th>No-Exercise Group</th>
<th>Eight-week Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00</td>
<td>11.00</td>
<td>11.50</td>
<td>13.25</td>
</tr>
<tr>
<td>9.50</td>
<td>10.00</td>
<td>12.00</td>
<td>11.50</td>
</tr>
<tr>
<td>9.75</td>
<td>10.00</td>
<td>9.00</td>
<td>12.00</td>
</tr>
<tr>
<td>10.00</td>
<td>11.75</td>
<td>11.50</td>
<td>13.50</td>
</tr>
<tr>
<td>13.00</td>
<td>10.50</td>
<td>13.25</td>
<td>11.50</td>
</tr>
<tr>
<td>9.50</td>
<td>15.00</td>
<td>13.00</td>
<td></td>
</tr>
</tbody>
</table>
### ANOVA: Example

<table>
<thead>
<tr>
<th>Group</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>9</td>
</tr>
<tr>
<td>Passive</td>
<td>10</td>
</tr>
<tr>
<td>None</td>
<td>11</td>
</tr>
<tr>
<td>Control</td>
<td>12</td>
</tr>
</tbody>
</table>

![ANOVA Example Chart](chart.png)
**ANOVA: SAS**

```sas
proc glm data=one; * proc anova data=one;
   class group;
   model age=group;
run;
```

<table>
<thead>
<tr>
<th>Sum of Sources</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>14.77780797</td>
<td>4.92593599</td>
<td>2.14</td>
<td>0.1285</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>43.68958333</td>
<td>2.29945175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>22</td>
<td>58.46739130</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**BIOS 662 25 ANOVA**
ANOVA: SAS

* factor effects model;
data two;
   set one;
   x1=0; x2=0; x3=0;
   if group="active" then x1=1;
   if group="passive" then x2=1;
   if group="no" then x3=1;
   if group="eight" then do; x1=x2=x3=-6/5; end;

proc reg data=two;
   model age = x1 x2 x3;

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>14.77781</td>
<td>4.92594</td>
<td>2.14</td>
<td>0.1285</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>43.68958</td>
<td>2.29945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>22</td>
<td>58.46739</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ANOVA: R

> av <- aov(age ~ group)
> anova(av)

Analysis of Variance Table

Response: age

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>3</td>
<td>14.778</td>
<td>4.926</td>
<td>2.1422</td>
<td>0.1285</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>43.690</td>
<td>2.299</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>