Statistics using R - Lecture for day 3 of block 1

ANOVA

March 2, 2017

Chalmers University of Technology
ANOVA provides a certain type of hypothesis testing.

One of the most widely used methods for testing if the means of different groups are unequal.

The name short for ANalysis Of VAriance. Investigate the (potential) variation in means for different groups via an analysis of the sample variances.

Originally developed by Ronald Fisher (his first application of the method was published in 1921).

See (Dalgaard, 2008, Chapter 7).
One-way ANOVA assumptions

Consider samples taken from \( k \) different groups.

Let \( X_{ij} \) denote observation \( j \) in group \( i \), \( j = 1, \ldots, n_i, \ i = 1, \ldots, k \).

Is assumed that these are independent, normally distributed and all have the same variance (homoscedasticity):

\[
X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).
\]

After the observations have been recorded (switching from RV \( X_{ij} \) to an observed value \( x_{ij} \)) it is useful to decompose them as

\[
x_{ij} = \bar{x}_{..} + (\bar{x}_i - \bar{x}_{..}) + (x_{ij} - \bar{x}_i.).
\]
The ANOVA null hypothesis of equal group means is

\[ H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k. \]

Investigate \( H_0 \) by comparing the variation of the data \textit{between} groups with the variation \textit{within} groups. Let

\[
SSD_T = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2,
\]

\[
SSD_B = \sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_i n_i(\bar{x}_{i.} - \bar{x}_{..})^2,
\]

\[
SSD_W = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2.
\]

Using straightforward algebra, it can be shown that the total variation can be partitioned according to

\[ SSD_T = SSD_B + SSD_W. \]
Mean squares

Regardless of whether $H_0$ holds or not, the usual sample variance estimate implies that

$$E\left[ SSD_W \right] = \sum_i E\left[ \sum_j (X_{ij} - \bar{X}_i)^2 \right] = \sum_i (n_i - 1)\sigma^2 = (N - k)\sigma^2,$$

where $N = \sum_{i=1}^k n_i$.

Under $H_0$, we have $X_{ij} \sim N(\mu + \alpha_1, \sigma^2)$, all independent. Hence,

$$E_{H_0} [SSD_T] = E_{H_0} \left[ \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 \right] = (N - 1)\sigma^2,$$

$$E_{H_0} [SSD_B] = E_{H_0} [SSD_T - SSD_W] = (k - 1)\sigma^2.$$

These motivates defining mean square sums:

$$MS_W = SSD_W / (N - k), \quad MS_B = SSD_B / (k - 1).$$
Define the F-statistic as

\[ F = \frac{MS_B}{MS_W}. \]

- If \( H_0 \) is true, then \( F \) will tend to be close to 1.
- If there is a difference between groups, then \( MS_B \) will tend to be larger than \( \sigma^2 \).
- Therefore, reject \( H_0 \) if \( F \) is large enough.
R commands for ANOVA analysis (1)

- An ANOVA table may be constructed using \texttt{anova}.
- \texttt{anova} is applied to a model object, e.g. a model created by \texttt{lm}.
- Consider the output of \texttt{anova(lm(Gain\sim Amount))}:

  Analysis of Variance Table
  Response: Gain

<table>
<thead>
<tr>
<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></td>
<td>Amount</td>
<td>1299.6</td>
<td>1299.60</td>
<td>5.3949</td>
</tr>
<tr>
<td></td>
<td>Residuals</td>
<td>9153.9</td>
<td>240.89</td>
<td></td>
</tr>
</tbody>
</table>

  Corresponding table using our notation:

<table>
<thead>
<tr>
<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>Amount</td>
<td>k – 1</td>
<td>SSD(_B)</td>
<td>MS(_B)</td>
<td>F</td>
</tr>
<tr>
<td>Residuals</td>
<td>N - k</td>
<td>SSD(_W)</td>
<td>MS(_W)</td>
<td></td>
</tr>
</tbody>
</table>

  Here, \(k = 2\) and \(N = 40\).
If $H_0$ can be rejected, next step is to estimate group means. Effects can be found by solving the system

$$\bar{x}_i = \mu + \alpha_i, \quad \text{under the restriction } \alpha_1 = 0.$$

R output of `summary(lm(Gain~Amount))`:

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 92.950 | 3.471 | 26.783 | <2e-16 *** |
| AmountLow | -11.400 | 4.908 | -2.323 | 0.0256 * |

\[ \bar{x}_1 = \mu + \alpha_1 = [\alpha_1 = 0] = \mu \implies \text{(Intercept)} = \mu = \bar{x}_1. \]

\[ \bar{x}_2 = \mu + \alpha_2 \implies \text{AmountLow} = \alpha_2 = \bar{x}_2 - \bar{x}_1. \]

Use `pairwise.t.test` to all possible two-group comparisons, with possibility of making adjustments for multiple comparisons.
Two-way ANOVA with interaction term

Two factors instead of one. Model for the observations:

\[ X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2). \]

Effects can be found by solving the system

\[ \bar{x}_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \]

under the restrictions \( \alpha_1 = \beta_1 = (\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = 0. \)

- Consider Gain as dependent on Protein and Amount.
- Two-way ANOVA table:
  \[
  \text{anova(lm(Gain~Protein*Amount))}
  \]
- Estimation of \( \mu, \alpha_2, \beta_2 \) and \( (\alpha\beta)_{22} \):
  \[
  \text{summary(lm(Gain~Protein*Amount))}
  \]
What if the ANOVA assumptions are not satisfied?

- Try to transform the data so that the assumptions hold.
- The more equal the variances, the more robust is the method to deviations from normality.
- `shapiro.test`: check for normality.
- `bartlett.test`: check for equality of variances.
- If variances are unequal, an alternative procedure due to Welsh is available: `oneway.test`. Can also be specified in `pairwise.t.test` by setting the argument `pool.sd=F`.
- `kruskal.wallis`: Non-parametric, rank based alternative to one-way ANOVA.
- `friedman.test`: Non-parametric alternative to two-way ANOVA.
References