Chemometrics

MA4605

Week 5. Lecture 9. ANOVA-example2

October 3, 2011

ANOVA example 2

When you collect data in more than 2 groups, you should **not** perform a separate t test for each pair.
Instead compare ALL the groups at once with ANOVA.
Example 2. Four areas in a lake are sampled and the chemical oxygen demand measured. The results are shown below.

Location 1	Location 2	Location 3	Location 4
48	73	51	72
54	63	63	68
57	66	61	71
54	64	54	68
62	74	56	66

- The sample data are separated into 4 groups according to one factor: location.

One way ANOVA tests the claim that the means are equal for $k = 4$ independent groups.
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: not all the means are equal (at least one mean is different).

ANOVA analyzes the variance among values.
When you combine data from several groups, the variance has two components

- variance among the group means= variance between groups

ANOVA analyzes the variance among values.
When you combine data from several groups, the variance has two components

- variance among the group means= variance between groups
- variance among the subjects within each group= variance within groups

ANOVA analyzes the variance among values.
When you combine data from several groups, the variance has two components

- variance among the group means= variance between groups
- variance among the subjects within each group= variance within groups

The first step when calculating the variance is to sum the squares of the differences between each value and the mean. This is called the **sum of squares**.
The variance is the mean sum of squares.

- If the H_0 is correct, then the two estimates of variance (between and within groups) should not differ significantly.
- If it is incorrect, the between-groups variance will be greater than the within group variance.
  To test wether it is significantly greater, a **one-sided F-test** is used.

	Location 1	Location 2	Location 3	Location 4
	48	73	51	72
	54	63	63	68
	57	66	61	71
	54	64	54	68
	62	74	56	66
sample size n	5	5	5	5
sample means	50	68	57	69
sample standard deviation	5.099	5.148	4.95	2.45

$\overline{\overline{x}}=62.25$

The total number of observations n=20.
Total variation = Total Sum of Squares= TSS
$=\sum_{i=1}^{n=20}(x_i - \overline{\overline{x}})^2$
$=(48-62.25)^2+(54-62.25)^2+(57-62.25)^2+(54-62.25)^2+(62-62.25)^2+(73-62.25)^2+(63-62.25)^2+...+(66-62.25)^2$
$=1125.75$
Degrees of freedom= n-1=20-1=19

The number of groups is k=4.
Variation between groups = Sum of Squares Between groups
= SSB $=\sum_{j=1}^{k} n_j(\overline{x}_j - \overline{\overline{x}})^2$
$=5(55-62.25)^2+5(68-62.25)^2+5(67-62.25)^2+5(69-62.25)^2$
= 5 (52.5625 + 33.0625 + 27.5625 + 45.5625)
=793.75
Degrees of freedom= Number of groups-1=k-1=4-1=3

Variation within groups = Sum of Squares Within groups
= Sum of Squared Errors = SSE
$=\sum_{j=1}^{k=4}(n_j - 1)s_j^2$
$=4(5.099)^2 + 4(5.148)^2 + 4(4.95)^2 + 4(2.45)^2$
=4(83) =332
Degrees of freedom$=\sum_{j=1}^{k=4}(n_j - 1)$ = 4 + 4 + 4 + 4 = 16

Rules

- Total Sum of Squares= Sum of Squares between groups + Sum of Squares within groups
  1125.75.92 = 793.75 + 332
- Total df= Between groups df+ Within groups df
  19= 3 +16

Calculate the test statistic

- Mean Square Between groups= MSB = $\frac{SSB}{df\ between} = \frac{793.75}{3} = 264.58$
- Mean Square Within groups= MSE = $\frac{SSE}{df\ within} = \frac{332}{16} = 20.75$
- Test Statistics= F= $\frac{MSB}{MSE} = \frac{264.58}{20.75} = 12.75$
- F $>F_{3,16;0.05} = 3.238872$, hence we reject the null hypothesis.

The critical value $F_{3,16;0.05}$ is obtained from *R* using
$>$ qf(0.95,3,16) or $>$ qf(0.05,3,16,lower.tail=FALSE)
The p-value is obtained from *R* using
$>$ pf(12.751,3,16,lower.tail=FALSE)
[1] 0.0001640662

ANOVA in R

$> y1 < -c(48, 54, 57, 54, 62)$

$> y2 < -c(73, 63, 66, 64, 74)$

$> y3 < -c(51, 63, 61, 54, 56)$

$> y4 < -c(72, 68, 71, 68, 66)$

$> mean(y1)$

$> mean(y2)$

$> mean(y3)$

$> mean(y4)$

$> sd(y1)$

$> sd(y2)$

$> sd(y3)$

$> sd(y4)$

$> y < -c(y1, y2, y3, y4)$

$> mean(y)$

$> index < -c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4)$

$> group < -factor(index)$

$> det < -data.frame(y, group)$
$> model < -aov(y \sim group, det)$
$> summary(model)$

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	3	793.75	264.58	12.751	0.0001641
Residuals	16	332.00	20.75		