Analysis of variance - ANOVA

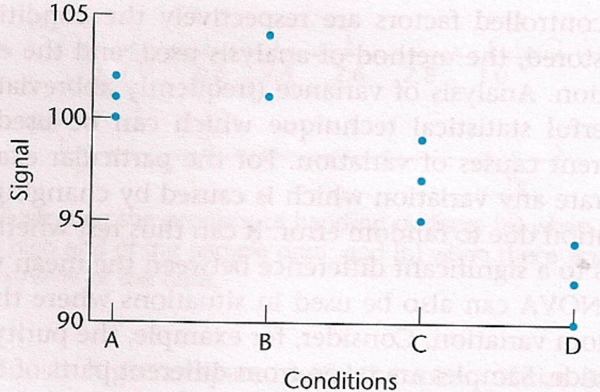
Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

October 7, 2009

ANOVA - Example

Table 3.2 Fluorescence from solutions stored under different conditions

Conditions	Replicate measurements	Mean
A Freshly prepared	102, 100, 101	101
B Stored for 1 hour in the dark	101, 101, 104	102
C Stored for 1 hour in subdued light	97, 95, 99	97
D Stored for 1 hour in bright light	90, 92, 94	92



Within-sample variation

Within-sample variation

- For each sample, say the *j*th sample we compute variance within sample

$$S_j^2 = \frac{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2}{n-1}.$$

Within-sample variation

- For each sample, say the jth sample we compute variance within sample
$$S_j^2 = \frac{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2}{n-1}.$$

- Within variance estimator of variance
$$\hat{\sigma}^2 = \sum_{j=1}^{h} s_j^2/h = \sum_{j=1}^{h}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2/(h(n-1)).$$

Within-sample variation

- For each sample, say the *j*th sample we compute variance within sample
$$S_j^2 = \frac{\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2}{n-1}.$$
- Within variance estimator of variance
$$\hat{\sigma}^2 = \sum_{j=1}^{h} s_j^2/h = \sum_{j=1}^{h}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2/(h(n-1)).$$
- **R** computations

```
x=matrix(c(102,100,101,101,101,104,97,95,99,90,92,94), byrow=T,ncol=3)
s=apply(x,1,var)
mean(s)
```

Between-sample variation

Between-sample variation

- Compute overall mean of the data

$$\bar{x} = \sum_{i=1}^{n} \sum_{j=1}^{h} x_{ij}/(nd).$$

Between-sample variation

- Compute overall mean of the data

$$\bar{x} = \sum_{i=1}^{n} \sum_{j=1}^{h} x_{ij}/(nd).$$

- For each sample, say the jth sample we compute its mean

$$\bar{x}_j = \sum_{i=1}^{n} X_{ji}/m.$$

Between-sample variation

- Compute overall mean of the data

$$\bar{x} = \sum_{i=1}^{n} \sum_{j=1}^{h} x_{ij}/(nd).$$

- For each sample, say the jth sample we compute its mean

$$\bar{x}_j = \sum_{i=1}^{n} X_{ji}/m.$$

- Between variance estimator of variance

$$\tilde{\sigma}^2 = n \sum_{j=1}^{h} (\bar{x}_j - \bar{x})/(h-1).$$

Between-sample variation

- Compute overall mean of the data

$$\bar{x} = \sum_{i=1}^{n} \sum_{j=1}^{h} x_{ij}/(nd).$$

- For each sample, say the jth sample we compute its mean

$$\bar{x}_j = \sum_{i=1}^{n} X_{ji}/m.$$

- Between variance estimator of variance

$$\tilde{\sigma}^2 = n \sum_{j=1}^{h} (\bar{x}_j - \bar{x})/(h-1).$$

- **R** computations

```
n=dim(x)[2]
m=apply(x,1,mean)
n*var(m)
```

F-test and detecting source of differences

We compute the test statistics $F = 62/3 \approx 20.7$ while the 95% quantile of F distribution with 3 and 8 degrees of freedom is given as

```
qf(0.95,3,8)

# 4.066181
```

We clearly see that the test informs us about a significant difference between the means.

F-test and detecting source of differences

We compute the test statistics $F = 62/3 \approx 20.7$ while the 95% quantile of F distribution with 3 and 8 degrees of freedom is given as

```
qf(0.95,3,8)

# 4.066181
```

We clearly see that the test informs us about a significant difference between the means.
But which means are different? The least significant difference method described in Section 3.9:
We compute the least significant difference $s\sqrt{2/n} * t$, where s^2 is within sample estimate of variance and t is the 97.5% quantile of Student-t distribution with $h(n-1)$ degrees of freedom.

```
sqrt(mean(s))*sqrt(2/3)*qt(0.975,8)
#  3.261182
m=apply(x,1,mean)
m
#[1] 101 102  97  92
```

Degrees of freedom and Sum of Squares (SS)

The associated degrees of freedom: for within-sample $h(n-1)$ (in our example $4 * 2 = 8$), for between-sample $h - 1$ (in our example 3).

Degrees of freedom and Sum of Squares (SS)

The associated degrees of freedom: for within-sample $h(n-1)$ (in our example $4*2=8$), for between-sample $h-1$ (in our example 3).
Total number of degrees freedom $hn-1$ and we see

$$hn-1=h(n-1)+h-1.$$

But there is more then the relation between degrees of freedom. Namely

$$SS_T=SS_M+SS_R,$$

where

$$\begin{aligned} SS_T &= \sum_{i=1}^{h}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2, \\ SS_M &= n\sum_{i=1}^{h}(\bar{x}_i-\bar{x})^2 \\ SS_R &= \sum_{i=1}^{h}\sum_{j=1}^{n}(x_{ij}-\bar{x}_j)^2. \end{aligned}$$

Table 3.4

Table 3.4 Summary of sums of squares and degrees of freedom

Source of variation	Sum of squares	Degrees of freedom
Between-sample	$n\sum_i (\bar{x}_i - \bar{x})^2 = 186$	$h - 1 = 3$
Within-sample	$\sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = 24$	$h(n - 1) = 8$
Total	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = 210$	$hn - 1 = 11$

Computations in **R**

```
x=c(102,100,101,101,101,104,97,95,99,90,92,94)
factors=c(rep("A",3),rep("B",3),rep("C",3),rep("D",3))
res=aov(x~factors)
anova(res)
Analysis of Variance Table

Response: x
         Df  Sum Sq Mean Sq F value    Pr(>F)
factors   3    186      62  20.667 0.0004002 ***
Residuals 8     24       3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```