

1.017/1.010 Class 19

Analysis of Variance

Concepts and Definitions

Objective: Identify factors responsible for variability in observed data

Specify one or more **factors** that could account for variability (e.g. location, time, etc.). Each factor is associated with a particular set of populations or **treatments** (e.g. particular sampling stations, sampling days, etc.). **One-way analysis of variance** (ANOVA) considers only a single factor.

Suppose a random sample $[x_{i1}, x_{i2}, \dots, x_{iJ}]$ is obtained for treatment i . There are $i=1, \dots, I$ treatments (e.g. each treatment may correspond to a different sampling location).

Arrange data in a table/array -- rows are treatments, columns are replicates:

$$\begin{bmatrix} x_{11}, x_{12}, \dots, x_{1J} \\ x_{21}, x_{22}, \dots, x_{2J} \\ \vdots \\ x_{I1}, x_{I2}, \dots, x_{IJ} \end{bmatrix}$$

Here we assume each treatment has same number of replicates J . The ANOVA procedure may be generalized to allow different number of replicates for each treatment.

Each random sample has a CDF $F_{x_i}(x_i)$. The different $F_{x_i}(x_i)$ are assumed **identical** except for their means, which may differ. Classical ANOVA also assumes that all data are **normally distributed**.

Each random variable x_{ij} is decomposed into several parts, as specified by the following **one-factor model**:

$$x_{ij} = \mu_i + e_{ij} = \mu + a_i + e_{ij}$$

$\mu_i = E[x_{ij}]$ is unknown mean of x_i (for all j).

μ = unknown **grand mean** (average of μ_i 's).

$a_i = \mu_i - \mu$ = unknown deviation of treatment mean from grand mean (often called an **effect**)

e_{ij} = **random residual** for treatment i , replicate j

$E[e_{ij}] = 0, Var[e_{ij}] = \sigma^2$, for all i, j

Objective is to estimate/test values of a_i 's, which are the unknown distributional parameters of the $F_{xi}(x_i)$'s.

Formulating the Problem as a Hypothesis Test

If the factor does not affect variability in the data then all a_i 's = 0. Use hypothesis test:

$$H_0: a_1 = a_2 = \dots = a_I = 0$$

It is better to test all a_i simultaneously than individually or in pairs. Test that sum-of-squared a_i 's = 0.

$$H_0: \sum_{i=1}^I a_i^2 = 0$$

Derive a test statistic based on sums-of-squares of data.

Sums-of-Squares Computations

Define the sample treatment and grand means:

$$m_{xi} = \frac{1}{J} \sum_{j=1}^J x_{ij} = \bar{x}_i.$$

$$m_x = \frac{1}{IJ} \sum_{i=1}^I \sum_{j=1}^J x_{ij} = \bar{x}_{..}$$

The **total sum-of-squares SST** measures variability of x_{ij} around m_x :

$$SST = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - m_x)^2$$

$$= \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - m_{xi})^2 + \sum_{i=1}^I \sum_{j=1}^J (m_{xi} - m_x)^2$$

$$= SSE + SSTr$$

SST can be divided into **error sum-of-squares SSE** and **treatment sum-of-squares SSTr**.

SSE measures variability of x_{ij} around m_{xi} , within treatments:

$$SSE = \sum_{i=1}^I \sum_{j=1}^J (x_{ij} - m_{xi})^2$$

SSTr measures variability of m_{xi} around m_x , across treatments:

$$SSTr = \sum_{i=1}^I \sum_{j=1}^J (m_{xi} - m_x)^2$$

Error and treatment **mean squared values**:

$$MSE = \frac{SSE}{I(J-1)}$$

$$MSTr = \frac{SSTr}{I-1}$$

$$E[MSE] = \sigma^2$$

$$E[MSTr] = \sigma^2 + \frac{J}{I-1} \sum_{i=1}^I a_i^2$$

MSE is an unbiased estimate of σ^2 , even if a_i 's are not zero.
MSTr is an unbiased estimate of σ^2 , **only if** all a_i 's are zero.

Test Statistic

Use ratio *MSTr*/*MSE* as a test statistic:

$$\mathcal{F}(MSE, MSTr) = \frac{MSTr}{MSE}$$

When H_0 is true and x_{ij} 's are **normally distributed** this statistic follows *F* **distribution** with $v_{Tr} = I - 1$ and $v_E = I(J-1)$ degrees of freedom. Check normality by plotting $(x_{ij} - m_{xi})$ with `normplot`.

One-sided rejection region (rejects only if *MSTr* is large):

$$R_0 : \mathcal{F}(MSE, MSTr) \geq F_{\mathcal{F}, v_{Tr}, v_E}^{-1}[\alpha]$$

One-sided p -value:

$$p = 1 - F_{\mathcal{F}, \nu_{Tr}, \nu_E} [F(MSE, MStr)]$$

Unbalanced ANOVA problems with **different sample sizes for different treatments** can be handled by modifying formulas slightly (see Devore, Section 10.3).

Single Factor ANOVA Tables

Above calculations are typically summarized in an **ANOVA** table:

Source	SS	df	MS	\mathcal{F}	p
Treatments	$SSTr$	$\nu_{Tr} = I-1$	$MSTr = SSTr/\nu_{Tr}$	$\mathcal{F} = MSTr/MSE$	$p = 1 - F_{\mathcal{F}, \nu_{Tr}, \nu_E}(\mathcal{F})$
Error	SSE	$\nu_E = I(J-1)$	$MSE = SSE/\nu_E$		
Total	SST	$\nu_T = IJ-1$	$MST = SST/\nu_T$		

Example -- Effect of Season on Oxygen Level

Consider following set of dissolved oxygen concentration data (x_{ij}) obtained in 4 different seasons/treatments (rows), 6 replicates per season (columns):

5.62	6.12	6.62	6.21	7.08	5.36
7.70	8.31	8.80	8.24	7.87	7.44
2.52	5.44	4.94	2.99	4.39	4.44
6.77	6.65	6.01	6.26	7.09	6.05

Use a single factor ANOVA to determine if season has a significant impact on oxygen variability.

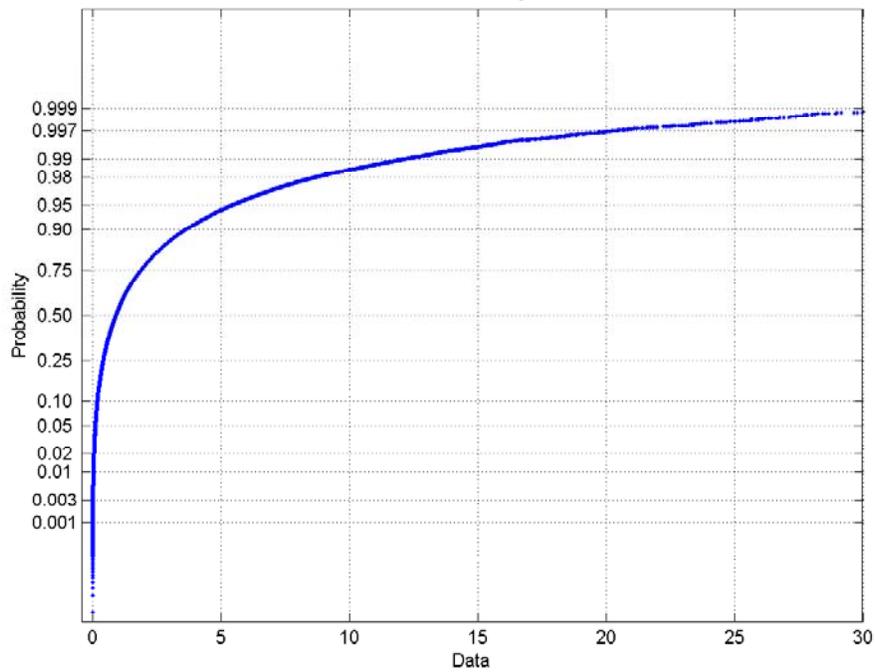
The MATLAB `anova1` function derives the error and treatment sums of squares and computes p value. **When using `anova1` be sure to transpose the data array** (MATLAB requires treatments in columns and replicates in rows).

Results are presented in this standard single factor **ANOVA table**:

Source	SS	df	MS=SS/df	F	p
Treatments	47.1642	3	15.7214	29.8	1.4E-7
Error	10.5518	20	0.5276		
Total	57.716	23			

The very low p value indicates that seasonality is **highly significant** in this case. Note that $MSTr$, which depends on the a_i 's, is much larger than MSE

F CDF, $\nu_{Tr} = 3$, $\nu_E = 5$



Copyright 2003 Massachusetts Institute of Technology
Last modified Oct. 8, 2003