1.017/1.010 Class 21
Multifactor Analysis of Variance

Multifactor Models

We often wish to consider several factors contributing to variability ratherr than just one. Extend concepts of single factor ANOVA to multiple factors. Focus on the two-factor case.

Suppose there I treatments for Factor A and J treatments for factor B., giving IJ random variables described by CDFs $F_{\boldsymbol{x}ij}(x_{ij})$. The different $F_{\boldsymbol{x}ij}(x_{ij})$ are assumed **identical** (except for their means) and **normally distributed** (check this, as in single factor case).

A random sample $[\boldsymbol{x}_{ij1}, \boldsymbol{x}_{ij2}, ..., \boldsymbol{x}_{ijK}]$ of size K is obtained for treatment combination (i, j). Two-factor model describing $\boldsymbol{x}_{ijk}$:

$$\boldsymbol{x}_{ijk} = \mu_{ij} + \boldsymbol{e}_{ijk} = \mu + a_i + b_j + c_{ij} + \boldsymbol{e}_{ijk}$$

$\mu_{ij} = E[\boldsymbol{x}_{ijk}] = \mu + a_i + b_j + c_{ij}$ = unknown mean of $\boldsymbol{x}_{ijk}$ (for all k)
μ = unknown **grand mean** (average of μ_i's).
a_i = unknown **main effects** of **Factor A**
b_j = unknown **main effects** of **Factor B**
c_{ij} = unknown **interactions** between Factors A and B
$\boldsymbol{e}_{ijk}$ = **random residual** for treatment i, replicate j
$E[\boldsymbol{e}_{ijk}] = 0$, $Var[\boldsymbol{e}_{ijk}] = \sigma^2$, for all i, j, k

Note that c_{ij} can only be distinguished from $\boldsymbol{e}_{ijk}$ if number of replicates K>1. Constraints:

$$\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} b_j = 0 \quad \sum_{i=1}^{I} c_{ij} = 0 \quad \forall j \qquad \sum_{j=1}^{J} c_{ij} = 0 \quad \forall i$$

Objective is to estimate/test values of a_i's, b_j's, and c_{ij}'s, which are distributional parameters for the $F_{\boldsymbol{x}ij}(x_{ij})$'s.

Formulating the Problem as a Hypothesis Test

Formulate three sum-of-squares hypotheses that insure that all a_i's, all b_i's, or all c_{ij}'s are zero:

$$\text{H0A}: \sum_{i=1}^{I} a_i^2 = 0$$

$$\text{H0B}: \sum_{j=1}^{J} b_i^2 = 0$$

$$\text{H0AB}: \sum_{j=1}^{J} \sum_{i=1}^{I} c_i^2 = 0$$

Derive test statistics based on sums-of-squares of data.

Sums-of-Squares Computations

Define treatment and grand sample means:

$$m_{xi} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk} = \bar{x}_{i..} \quad m_{xj} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} x_{ijk} = \bar{x}_{.j.}$$

$$m_{xij} = \frac{1}{K} \sum_{j=1}^{K} x_{ijk} = \bar{x}_{ij.}$$

$$m_x = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk} = \bar{x}_{...}$$

Test statistics are computed from sums-of-squares:

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - m_{xij})^2$$

$$SSA = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (m_{xi} - m_x)^2 \quad SSB = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (m_{xj} - m_x)^2$$

$$SSAB = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (m_{xij} - m_{xi} - m_{xj} + m_x)^2$$

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (x_{ijk} - m_x)^2 = SSE + SSA + SSB + SSAB$$

Corresponding mean-sums-of-squares are:

$$MSE = \frac{SSE}{IJ(K-1)}$$

$$MSA = \frac{SSA}{I-1} \qquad MSB = \frac{SSB}{J-1}$$
$$MSAB = \frac{SSAB}{(I-1)(J-1)}$$

Expected values of these mean-sums-of-squares show depends on main effects and interactions:

$$E[MSE] = \sigma^2$$
$$E[MSA] = \sigma^2 + \frac{JK}{I-1}\sum_{i=1}^{I} a_i^2 \qquad E[MSB] = \sigma^2 + \frac{IK}{I-1}\sum_{j=1}^{J} b_j^2$$
$$E[MSAB] = \sigma^2 + \frac{K}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}^2$$

Test Statistic

Use ratios as test statistics for the three hypotheses:

$$\mathcal{F}_A(MSA, MSE) = \frac{MSA}{MSE}$$
$$\mathcal{F}_B(MSB, MSE) = \frac{MSB}{MSE}$$
$$\mathcal{F}_{AB}(MSAB, MSE) = \frac{MSAB}{MSE}$$

When H0 is true each statistic follows **F distribution** with degree of freedom parameters $\nu_A = I - 1$, $\nu_B = J - 1$, $\nu_{AB} = (I - 1)(J - 1)$, and $\nu_E = IJ(K-1)$.

One-sided rejection regions

$$R0A : \mathcal{F}(MSA, MSE) \geq F^{-1}_{\mathcal{F}, \nu_A, \nu_E}[\alpha]$$
$$R0B : \mathcal{F}(MSB, MSE) \geq F^{-1}_{\mathcal{F}, \nu_B, \nu_E}[\alpha]$$
$$R0AB : \mathcal{F}(MSAB, MSE) \geq F^{-1}_{\mathcal{F}, \nu_{AB}, \nu_E}[\alpha] :$$

One-sided p-values:

$$p_A = 1 - F_{\mathcal{F}, \nu_A, \nu_E}[\mathcal{F}(MSA, MSE)]$$

$$p_B = 1 - F_{\mathcal{F}, \nu_A, \nu_E}[\mathcal{F}(MSB, MSE)]$$

$$p_{AB} = 1 - F_{\mathcal{F}, \nu_{AB}, \nu_E}[\mathcal{F}(MSAB, MSE)]$$

Unbalanced ANOVA problems with **different sample sizes for different treatments** can be handled by modifying formulas slightly.

Two-Factor ANOVA Tables

Source	SS	df	MS	$\mathcal{F}$	p
Factor A	SSA	$\nu_A = I - 1$	$MSA = SSA/\nu_A$	$\mathcal{F}_A = MSA/MSE$	$p = 1 - F_{\mathcal{F},\nu A,\nu E}(\mathcal{F})$
Factor B	SSB	$\nu_B = J - 1$	$MSB = SSB/\nu_B$	$\mathcal{F}B = MSB/MSE$	$p = 1 - F_{\mathcal{F},\nu B,\nu E}(\mathcal{F})$
Interaction AB	$SSAB$	$\nu_{AB} = (I-1)(J-1)$	$MSAB = SSAB/\nu_{AB}$	$\mathcal{F}_{AB} = MSAB/MSE$	$p = 1 - F_{\mathcal{F},\nu AB,\nu E}(\mathcal{F})$
Error	SSE	$\nu_E = IJ(K-1)$	$MSE = SSE/\nu_E$		
Total	SST	$\nu_T = IJK-1$			

Exercise: Two Factor ANOVA

Relevant MATLAB functions: `normplot, anova2`

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 $[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{iJ}]$ is obtained for treatment i. There are i =1,..., 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:

$$
\begin{array}{c}
[\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, ..., \boldsymbol{x}_{1J}] \\
[\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, ..., \boldsymbol{x}_{2J}] \\
. \\
. \\
[\boldsymbol{x}_{I1}, \boldsymbol{x}_{I2}, ..., \boldsymbol{x}_{IJ}]
\end{array}
$$

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 $F_{\boldsymbol{x}i}(x_i)$. The different $F_{\boldsymbol{x}i}(x_i)$ are assumed **identical** except for their means, which may differ. Classical ANOVA also assumes that all data are **normally distributed.**

Each random variable $\boldsymbol{x}_{ij}$ is decomposed into several parts, as specified by the following **one-factor model**:

$$\boldsymbol{x}_{ij} = \mu_i + \boldsymbol{e}_{ij} = \mu + a_i + \boldsymbol{e}_{ij}$$

$\mu_i = E[\boldsymbol{x}_{ij}]$ is unknown mean of $\boldsymbol{x}_i$ (for all j).
μ = unknown **grand mean** (average of $\mu_{i's}$).
$a_i = \mu_i - \mu$ = unknown deviation of treatment mean from grand mean (often called an **effect**)
$\boldsymbol{e}_{ij}$ = **random residual** for treatment i, replicate j
$E[\boldsymbol{e}_{ij}] = 0$, $Var[\boldsymbol{e}_{ij}] = \sigma^2$, for all i, j

Objective is to estimate/test values of a_i's, which are the unknown distributional parameters of the $F_{\boldsymbol{x}i}(x_i)$'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:

$$\text{H0: } a_1 = a_2 = \ .... \ = a_I = 0$$

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

$$\text{H0}: \sum_{i=1}^{I} a_i^2 = 0$$

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

Sums-of-Squares Computations

Define the sample treatment and grand means:

$$\boldsymbol{m}_{\boldsymbol{xi}} = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{x}_{ij} = \bar{\boldsymbol{x}}_{i\boldsymbol{.}}$$

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

The **total sum-of-squares** $\boldsymbol{SST}$ measures variability of $\boldsymbol{x}_{ij}$ around $\boldsymbol{m}_{\boldsymbol{x}}$:

$$\begin{aligned}\boldsymbol{SST} &= \sum_{i=1}^{I}\sum_{j=1}^{J} (\boldsymbol{x}_{ij} - \boldsymbol{m}_{\boldsymbol{x}})^2 \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} (\boldsymbol{x}_{ij} - \boldsymbol{m}_{\boldsymbol{xi}})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J} (\boldsymbol{m}_{\boldsymbol{xi}} - \boldsymbol{m}_{\boldsymbol{x}})^2 \\ &= \boldsymbol{SSE} + \boldsymbol{SSTr}\end{aligned}$$

$\boldsymbol{SST}$ can be divided into **error sum-of-squares** $\boldsymbol{SSE}$ and **treatment sum-of-squares** $\boldsymbol{SSTr}$.

$\boldsymbol{SSE}$ measures variability of $\boldsymbol{x}_{ij}$ around $\boldsymbol{m}_{\boldsymbol{xi}}$, within treatments:

$$\boldsymbol{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J} (\boldsymbol{x}_{ij} - \boldsymbol{m}_{\boldsymbol{xi}})^2$$

$\boldsymbol{SSTr}$ measures variability of $\boldsymbol{m}_{\boldsymbol{xi}}$ around $\boldsymbol{m}_{\boldsymbol{x}}$, across treatments:

$$\boldsymbol{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J} (\boldsymbol{m}_{\boldsymbol{xi}} - \boldsymbol{m}_{\boldsymbol{x}})^2$$

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 H0 is true and x_{ij}'s are **normally distributed** this statistic follows ***F* distribution** with $\nu_{Tr} = I - 1$ and $\nu_E = I(J-1)$ degrees of freedom. Check normality by plotting ($x_{ij} - m_{xi}$) with `normplot`.

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

$$R0 : \mathcal{F}(MSE, MSTr) \geq F^{-1}_{\mathcal{F}, \nu_{Tr}, \nu_E}[\alpha]$$

One-sided p-value:

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

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	$SSTr$	$\nu_{Tr} = I\text{-}1$	$MSTr = SSTr/\nu_{Tr}$	$\mathcal{F} = MSTr/MSE$	$p = 1\text{-}F_{\mathcal{F},\nu Tr,\nu E}(\mathcal{F})$
Error	SSE	$\nu_E = I(J\text{-}1)$	$MSE = SSE/\nu_E$		
Total	SST	$\nu_T = IJ\text{-}1$	$MST = SST/\nu_T$		

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