Chapter 1

Chapter 2

Chapter 3

Chapter 4

Analysis of Variance (ANOVA)

4.1 One-factor analysis of variance

Introduction

Analysis of Variance (ANOVA) is a statistical method used to compare the means of multiple groups to determine if the observed differences are significant. Principle: compare the variance **between groups** and the variance **within groups**.

Assumptions

- Independence of observations
- Normality of populations
- Homogeneity of variances

Formally:

$$H_0 : m_1 = m_2 = \cdots = m_P, \quad H_1 : \text{at least one mean is different}$$

Quantitative table of observations

The set of observations can be represented in the following table:

Sample	X1	X2	X3	X4	X5	...	Size n_i	Mean $\bar{X}_i$
E1	X11	X12	X13	X14	X15	...	n_1	$\bar{X}_1$
E2	X21	X22	X23	X24	X25	...	n_2	$\bar{X}_2$
E3	X31	X32	X33	X34	X35	...	n_3	$\bar{X}_3$
$\vdots$								
Ep	Xp1	Xp2	Xp3	Xp4	Xp5	...	n_p	$\bar{X}_p$

Variance decomposition

Factorial sum of squares (between groups)

$$SCEf = \sum_{i=1}^{P} n_i(\bar{X}_i - \bar{X})^2$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{P}\sum_{j=1}^{n_i} X_{ij}$$

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

Residual sum of squares (within groups)

$$SCEr = \sum_{i=1}^{P}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

Total sum of squares

$$SCT = \text{SCEf} + \text{SCEr}$$

Mean squares and F statistic

In one-way ANOVA, mean squares are calculated as:

$$MS_f = \frac{SCE_f}{P-1}, \quad MS_r = \frac{SCE_r}{N-P}$$

The Fisher F test statistic is:

$$F_{\text{obs}} = \frac{MS_f}{MS_r}$$

where:

- P (or k) : number of groups
- N : total number of observations
- SCE_f : factorial sum of squares (between groups)
- SCE_r : residual sum of squares (within groups)

Decision

$$\text{If } F_{\text{obs}} > F_{\alpha,\,P-1,\,N-P} \quad \Rightarrow \quad \text{reject } H_0$$

$$\text{Otherwise} \quad \text{do not reject } H_0$$

where H_0 is the null hypothesis: all group means are equal.

ANOVA summary table

Source of variation	df	Sum of squares (SCE)	Mean squares (MS)	Observed F
Factorial variation	P-1	SCEf	MSf = SCEf/(P-1)	F = MSf/MSr
Residual variation	N-P	SCEr	MSr = SCEr/(N-P)	
Total	N-1	SCT = SCEf + SCEr		

Numerical example

We want to compare the yield (in kg) of 3 wheat varieties (A, B, and C) grown under the same conditions. Each variety was tested on 4 plots (replications).

Variety	Plot 1	Plot 2	Plot 3	Plot 4
A	20	22	19	21
B	25	27	26	28
C	18	17	19	20

Questions

1. Compute the mean of each variety and the overall mean.
2. Compute the sum of squares between groups (SCEf) and the residual sum of squares (SCEr).
3. Compute the mean squares (MSf and MSr) and the F statistic.
4. Test at $\alpha = 0.05$ whether the variety means differ significantly.
5. Complete the ANOVA table.

Solution

1. Means

$$\bar{X}_A = \frac{20 + 22 + 19 + 21}{4} = 20.5$$

$$\bar{X}_B = \frac{25 + 27 + 26 + 28}{4} = 26.5$$

$$\bar{X}_C = \frac{18 + 17 + 19 + 20}{4} = 18.5$$

$$\bar{X} = \frac{20 + 22 + 19 + 21 + 25 + 27 + 26 + 28 + 18 + 17 + 19 + 20}{12} = 21.833$$

2. Sum of squares

SCEf (between groups):

$$SCEf = 4[(20.5 - 21.833)^2 + (26.5 - 21.833)^2 + (18.5 - 21.833)^2] \approx 144.2$$

SCEr (within groups):

For A: $(20 - 20.5)^2 + (22 - 20.5)^2 + (19 - 20.5)^2 + (21 - 20.5)^2 = 5$
For B: $(25 - 26.5)^2 + (27 - 26.5)^2 + (26 - 26.5)^2 + (28 - 26.5)^2 = 5$
For C: $(18 - 18.5)^2 + (17 - 18.5)^2 + (19 - 18.5)^2 + (20 - 18.5)^2 = 5$
$SCEr = 5 + 5 + 5 = 15$

$$SCT = SCEf + SCEr = 144.2 + 15 = 159.2$$

3. Mean squares and F

$$MSf = \frac{SCEf}{P - 1} = \frac{144.2}{2} = 72.1$$
$$MSr = \frac{SCEr}{N - P} = \frac{15}{12 - 3} = 1.667$$
$$F = \frac{MSf}{MSr} = \frac{72.1}{1.667} \approx 43.2$$

4. F test

Degrees of freedom: $df_1 = 2$, $df_2 = 9$

$$F_{0.05,2,9} \approx 4.26$$

Since $F_{\text{obs}} = 43.2 > 4.26 \Rightarrow$ reject H_0. **Conclusion:** The variety means are significantly different.

5. ANOVA table

Source of variation	df	Sum of squares	Mean squares	Observed F
Factorial variation	2	144.2	72.1	43.2
Residual variation	9	15	1.667	
Total	11	159.2		

4.2 Two-Factor Analysis of Variance (Two-Way ANOVA)

We simultaneously study two factors A and B:

- p levels for A: $A_1, \ldots, A_p$
- q levels for B: $B_1, \ldots, B_q$

For each pair (A_i, B_j), we have a sample of size n.

Table of observations

	A_1	A_2	$\cdots$ A_p
B_1	$x_{11,1}, \ldots, x_{11,n}$	$x_{12,1}, \ldots, x_{12,n}$	$\cdots$
B_2	$x_{21,1}, \ldots, x_{21,n}$	$x_{22,1}, \ldots, x_{22,n}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$
B_q	$x_{q1,1}, \ldots, x_{q1,n}$	$x_{q2,1}, \ldots, x_{q2,n}$	$\cdots$

Hypotheses

Factor A:

$$H_{0A} : \mu_{A_1} = \cdots = \mu_{A_p} \qquad H_{1A} : \text{at least two means differ.}$$

Factor B:

$$H_{0B} : \mu_{B_1} = \cdots = \mu_{B_q} \qquad H_{1B} : \text{at least two means differ.}$$

Interaction:

H_{0AB} : no interaction between A and B $\qquad$ H_{1AB} : interaction present.

Means

	A_1	A_2	$\cdots$	A_p	Row means
B_1	$\bar{x}_{11}$	$\bar{x}_{12}$	$\cdots$	$\bar{x}_{1p}$	$\bar{x}_{1.}$
B_2	$\bar{x}_{21}$	$\bar{x}_{22}$	$\cdots$	$\bar{x}_{2p}$	$\bar{x}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_q	$\bar{x}_{q1}$	$\bar{x}_{q2}$	$\cdots$	$\bar{x}_{qp}$	$\bar{x}_{q.}$
Column means	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\cdots$	$\bar{x}_{.p}$	$\bar{X}$

Cell mean:

$$\bar{x}_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ijk}.$$

Row mean (effect of factor B):

$$\bar{x}_{i.} = \frac{1}{p} \sum_{j=1}^{p} \bar{x}_{ij}.$$

Column mean (effect of factor A):

$$\bar{x}_{.j} = \frac{1}{q} \sum_{i=1}^{q} \bar{x}_{ij}.$$

Grand mean:

$$\bar{X} = \frac{1}{pqn} \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} x_{ijk}.$$

Variances

Variance in cell (i, j):

$$S_{ij}^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2.$$

Residual variance:

$$S_R^2 = \frac{1}{pq(n-1)} \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2.$$

Sums of squares

Total sum:

$$SCE_T = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} (x_{ijk} - \bar{X})^2.$$

Residual:

$$SCE_R = \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2.$$

Effect of factor A:

$$SCE_A = nq \sum_{j=1}^{p} (\bar{x}_{.j} - \bar{X})^2.$$

Effect of factor B:

$$SCE_B = np \sum_{i=1}^{q} (\bar{x}_{i.} - \bar{X})^2.$$

Interaction:

$$SCE_{AB} = n \sum_{i=1}^{q} \sum_{j=1}^{p} (\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X})^2.$$

ANOVA Table

Source	df	SCE	MS	F
Factor A	$p-1$	SCE_A	$MS_A = \frac{SCE_A}{p-1}$	$\frac{MS_A}{MS_R}$
Factor B	$q-1$	SCE_B	$MS_B = \frac{SCE_B}{q-1}$	$\frac{MS_B}{MS_R}$
Interaction	$(p-1)(q-1)$	SCE_{AB}	$MS_{AB} = \frac{SCE_{AB}}{(p-1)(q-1)}$	$\frac{MS_{AB}}{MS_R}$
Residual	$pq(n-1)$	SCE_R	$MS_R = \frac{SCE_R}{pq(n-1)}$	-
Total	$pqn-1$	SCE_T	-	-

F Statistics

$$F_A = \frac{MS_A}{MS_R}, \qquad F_B = \frac{MS_B}{MS_R}, \qquad F_{AB} = \frac{MS_{AB}}{MS_R}.$$

Decision

$$F_{\text{calc}} \geq F_{\text{theo}}(df_1, df_2) \Rightarrow \text{Reject } H_0.$$

$$F_{\text{calc}} < F_{\text{theo}}(df_1, df_2) \Rightarrow \text{Do not reject } H_0.$$

Conditions for Applying the F test

The variable must be approximately normal in all populations and must have homogeneous variance.

Case with one observation per cell ($n = 1$)

The residual term disappears because $\bar{x}_{ij} = x_{ij}$.

$$SCE_t = SCE_{f1} + SCE_{f2} + SCE_{int}.$$

$$SCE_{f1} = \frac{1}{q}\sum_{i=1}^{p} x_{i\cdot}^2 - \frac{X_{\cdot\cdot}^2}{pq}, \qquad SCE_{f2} = \frac{1}{p}\sum_{j=1}^{q} x_{\cdot j}^2 - \frac{X_{\cdot\cdot}^2}{pq}.$$

Exercise: Two-Factor ANOVA (A × B)

Problem: We study the effect of:

- Factor A: 3 levels (A_1, A_2, A_3)
- Factor B: 2 levels (B_1, B_2)

For each pair (A_i, B_j), we perform $n = 4$ replications.

Observed Data (yield)

	A_1	A_2	A_3
B1	8, 9, 7, 10	12, 11, 13, 12	14, 15, 13, 14
B2	6, 7, 5, 6	10, 9, 11, 10	12, 11, 12, 13

Questions

Compute: a) The cell means $\bar{x}_{ij}$ b) The row means $\bar{x}_{i\cdot}$ c) The column means $\bar{x}_{\cdot j}$ d) The grand mean $\bar{X}$

Compute the residual variance S_R^2.

Compute the sums of squares (SS): a) Factor A: SCE_A b) Factor B: SCE_B c) Interaction: SCE_{AB} d) Residual: SCE_R

Complete the ANOVA table with degrees of freedom (df), sums of squares (SS), mean squares (MS), and F statistics.

Compute the F statistics: F_A, F_B, F_{AB}.

Test the hypotheses at $\alpha = 0.05$ and conclude for: — Effect of factor A — Effect of factor B — Interaction A × B

1. Means

Cell means (i, j)**:**

$$\bar{x}_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk}$$

$$\bar{x}_{11} = \frac{8+9+7+10}{4} = 8.5, \quad \bar{x}_{12} = \frac{12+11+13+12}{4} = 12, \quad \bar{x}_{13} = \frac{14+15+13+14}{4} = 14$$
$$\bar{x}_{21} = \frac{6+7+5+6}{4} = 6, \qquad \bar{x}_{22} = \frac{10+9+11+10}{4} = 10, \quad \bar{x}_{23} = \frac{12+11+12+13}{4} = 12$$

Row means (effect B):

$$\bar{x}_{i.} = \frac{1}{p}\sum_{j=1}^{p} \bar{x}_{ij}, \quad \bar{x}_{1.} = \frac{8.5+12+14}{3} = 11.5, \quad \bar{x}_{2.} = \frac{6+10+12}{3} \approx 9.33$$

Column means (effect A):

$$\bar{x}_{.j} = \frac{1}{q}\sum_{i=1}^{q} \bar{x}_{ij}, \quad \bar{x}_{.1} = \frac{8.5+6}{2} = 7.25, \quad \bar{x}_{.2} = \frac{12+10}{2} = 11, \quad \bar{x}_{.3} = \frac{14+12}{2} = 13$$

Grand mean:

$$\bar{X} = \frac{1}{pq}\sum_{i,j} \bar{x}_{ij} = \frac{8.5+12+14+6+10+12}{6} \approx 10.42$$

2. Within-cell variances (residual)

$$S_{ij}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij})^2$$
$$S_{11}^2 = \frac{(8-8.5)^2+(9-8.5)^2+(7-8.5)^2+(10-8.5)^2}{3} = 1.667$$
$$S_{12}^2 = \frac{(12-12)^2+(11-12)^2+(13-12)^2+(12-12)^2}{3} = 0.667$$
$$S_{13}^2 = 0.667, \quad S_{21}^2 = 0.667, \quad S_{22}^2 = 0.667, \quad S_{23}^2 = 0.667$$

Residual variance:

$$S_R^2 = \frac{\sum_{i,j}(n-1)S_{ij}^2}{pq(n-1)} = \frac{3(1.667+0.667+0.667+0.667+0.667+0.667)}{6*3} = 1$$

3. Sums of squares

$$SCE_A = nq\sum_j(\bar{x}_{.j}-\bar{X})^2 = 4*2*((7.25-10.42)^2+(11-10.42)^2+(13-10.42)^2) \approx 136.26$$

$$SCE_B = np\sum_i(\bar{x}_{i.} - \bar{X})^2 = 4*3*((11.5-10.42)^2+(9.33-10.42)^2) \approx 28.25$$

$$SCE_{AB} = n\sum_{i,j}(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X})^2 \approx 0.335$$

$$SCE_R = \sum_{i,j}(n-1)S_{ij}^2 = 10$$

$$SCE_T = SCE_A + SCE_B + SCE_{AB} + SCE_R \approx 174.85$$

4. ANOVA Table

Source	df	SCE	MS	F
Factor A	2	136.26	68.13	40.85
Factor B	1	28.25	28.25	16.95
Interaction	2	0.335	0.167	0.10
Residual	6	10	1.667	-
Total	11	174.85	-	-

5. F Statistics

$$F_A = \frac{MS_A}{MS_R} = \frac{68.13}{1.667} \approx 40.85$$

$$F_B = \frac{MS_B}{MS_R} = \frac{28.25}{1.667} \approx 16.95$$

$$F_{AB} = \frac{MS_{AB}}{MS_R} = \frac{0.167}{1.667} \approx 0.10$$

6. Decision at $\alpha = 5\%$

Critical values for $\alpha = 0.05$:

$$F_{\text{theo}}(2, 6) \approx 5.14, \quad F_{\text{theo}}(1, 6) \approx 5.99$$

- Factor A: $F_A = 40.85 > 5.14 \Rightarrow$ **reject** H_{0A}
- Factor B: $F_B = 16.95 > 5.99 \Rightarrow$ **reject** H_{0B}
- Interaction: $F_{AB} = 0.10 < 5.14 \Rightarrow$ **do not reject** H_{0AB}

Conclusion: Factors A and B both have a significant effect on the response variable, while the interaction between A and B is not significant.