

Analysis of Variance

1

One-way ANOVA

- So far we have discussed the comparison between two normal populations with possibly different means but similar variances; i.e., we tested the null hypothesis:

$$H_0 : \mu_1 = \mu_2$$

- Now, it is needed to expand the test to include the comparison of three, or more normal populations; i.e., to test the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_t$$

2

One-way ANOVA Procedure-I

- In one-way classification, the model is given by:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i=1,2,\dots, a$$

$$\varepsilon_{ij} \sim \text{iidN}(0, \sigma^2)$$

Where:

μ = the mean of all treatments

α_i = the mean effect of treatment in the i^{th} group

ε_{ij} = the "random effect" due to individual treatment.

3

One-way ANOVA Procedure-II

- The null hypothesis is given as following (using the mean treatment effect estimated as the difference between grand mean and treatment mean):

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \dots = \alpha_a = 0$$

\neq

- The development of the F-test that follows, comparing the variance **among** groups with the variance **within** groups to test the above hypotheses.
- The data in each group is assumed to be normally distributed and all groups have equal variances (homogeneous).

4

One-way ANOVA Procedure-III

$$\text{MST} = \frac{\sum_i \sum_j (y_{ij} - \bar{y})^2}{na - 1}$$

The mean squared deviation of the observations from the grand mean

$$\text{MSW} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{a(n - 1)}$$

The mean squared deviation of the observations from their respective group mean (the pooled variance)

$$\text{MSA} = n \left[\frac{\sum_i (y_{i.} - \bar{y})^2}{a - 1} \right]$$

The mean squared deviation of the group mean from the grand mean multiplied by the number of observations in each group

5

One-way ANOVA Procedure-IV

- If the null hypothesis is true, the group means will not differ from the overall mean and the within-group variance will be approximately the same as the among-group variance.
- However, if the null hypothesis is false, then the among-group variance will be larger because of the significant deviations of the group means from the grand mean.
- Now let's compute the variances (sums -of-squares) for each group...
- This illustrates how the total sums -of-squares can be partitioned into two parts: among-group and within-group; i.e.,

$$\text{Total SS} = \text{Among SS} + \text{Within SS}$$

6

One-way ANOVA Procedure-V

- The complete and final ANOVA table can be given as:

Source	df	SSq	MSq	F-test
Among group	a-1	SSA	SSA/(a-1)	MSA/MSW
Within group	a(n-1)	SSW	SSW/a(n-1)	
Total	an-1			

If $F\text{-test} > F_{\text{table}}$, then we reject H_0 and conclude that there is at least one level of treatment is significantly different from the others.

7

Multiple Comparison Procedures

- The null hypothesis, which was just rejected, indicates that *there is at least one inequality* among the levels of treatment. However, it is still *unknown which levels (or samples)* are unequal with respect to the others.
- Typically, the researchers wish to make further decisions, particularly if this is a carefully designed experiment with a control and one (or more) treatments.
- Some multiple comparison methods will be described here.

8

Scheffe's Multiple Comparison

- The LSD values for Scheffe's are needed for each pairwise ij comparison and can be defined as follows:

$$LSD_{ij}(s) = \sqrt{(a-1)F_{\alpha, a-1, a(n-1)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) MSE}$$

Where MSE is MSQ of within group (or Error Mean Square).

9

Tukey's Multiple Comparison

- If the sample sizes are equal among all groups, a procedure that is less conservative than Scheffe's MCP is Tukey's MCP:

$$LSD_{ij}(T) = q_{(\alpha, a, a(n-1))} \sqrt{\frac{MSE}{n}}$$

Where $k=a(a-1)/2$ and $q_{(\alpha, a, a(n-1))}$ comes from table of this test.

10

Bonferroni Multiple Comparison

- The Bonferroni modification of the LSD method is also a very good way to control for the experiment-wise error rate and it is defined by:

$$LSD_{ij}(B) = t_{(\alpha, a(n-1), k)} \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j} \right) MSE}$$

Where $k=a(a-1)/2$ and $t_{(\alpha, a(n-1), k)}$ comes from table of this test.

11

Analysis of Variance/Experimental Design

- Analysis of Variance (ANOVA) was originally devised for agricultural statistics on e.g. crop yields. Typically, row and column format, = small plots of a fixed size. The yield $y_{i,j}$ within each plot was recorded.

1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$		
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$		

One Way classification

Model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, s)$
 where μ = overall mean, α_i = effect of the i^{th} factor
 and ε_{ij} = error term.

Hypothesis: $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_m$

12

Factor I

1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$		$y_{1,n1}$	Totals $T_1 = \sum y_{1,j}$	Means $\bar{y}_1 = T_1 / n_1$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$		$y_{2,n2}$	$T_2 = \sum y_{2,j}$	$\bar{y}_2 = T_2 / n_2$
m	$y_{m,1}$	$y_{m,2}$	$y_{m,3}$		$y_{m,nm}$	$T_m = \sum y_{m,j}$	$\bar{y}_m = T_m / n_m$

Overall mean $\bar{y} = \sum \sum y_{ij} / n$, where $n = \sum n_i$

Decomposition (Partition) of Sums of Squares:
 $\sum \sum (y_{ij} - \bar{y})^2 = \sum n_i (\bar{y}_i - \bar{y})^2 + \sum \sum (y_{ij} - \bar{y}_i)^2$

Total Variation (Q) = Between Factors (Q₁) + Residual Variation (Q_E)

Under H₀: $Q / (n-1) \Rightarrow \chi^2_{n-1}$, $Q_1 / (m-1) \Rightarrow \chi^2_{m-1}$, $Q_E / (n-m) \Rightarrow \chi^2_{n-m}$

$Q_1 / (m-1) \Rightarrow F_{m-1, n-m}$
 $Q_E / (n-m)$

13

ANOVA Table

Variation	D.F.	Sum of Squares	Mean Squares	F
Between	m-1	$Q_1 = \sum n_i (\bar{y}_i - \bar{y})^2$	$MS_1 = Q_1 / (m-1)$	MS_1 / MS_E
Residual	n-m	$Q_E = \sum \sum (y_{ij} - \bar{y}_i)^2$	$MS_E = Q_E / (n-m)$	---
Total	n-1	$Q = \sum \sum (y_{ij} - \bar{y})^2$	$Q / (n-1)$	---

14

Two-Way Classification

	Factor I			Means
Factor II	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,n}$
	$y_{m,1}$	$y_{m,2}$	$y_{m,3}$	$y_{m,n}$
Means	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\bar{y}_{\cdot 3}$	$\bar{y}_{\cdot n}$

$\bar{y}_{\cdot \cdot}$ Write as \bar{y}

Partition SSQ:
 $\sum \sum (y_{ij} - \bar{y})^2 = n \sum (\bar{y}_{i\cdot} - \bar{y})^2 + m \sum (y_{\cdot j} - \bar{y})^2 + \sum \sum (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2$

Total Variation Rows Between Columns Residual Variation

Model: $y_{i,j} = \mu + \alpha_i + \beta_j + \epsilon_{i,j}$, $\epsilon_{i,j} \Rightarrow N(0, s)$

H₀: All α_i are equal and all β_j are equal

15

ANOVA Table

Variation	D.F.	Sums of Squares	Mean Squares	F
Between Rows	m-1	$Q_1 = n \sum (\bar{y}_{i\cdot} - \bar{y})^2$	$MS_1 = Q_1 / (m-1)$	MS_1 / MS_E
Between Columns	n-1	$Q_2 = m \sum (\bar{y}_{\cdot j} - \bar{y})^2$	$MS_2 = Q_2 / (n-1)$	MS_2 / MS_E
Residual	(m-1)(n-1)	$Q_E = \sum \sum (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y})^2$	$MS_E = Q_E / ((m-1)(n-1))$	---
Total	mn-1	$Q = \sum \sum (y_{ij} - \bar{y})^2$	$Q / (mn-1)$	---

16

Two-Way Example

Factor I	1	2	3	4	5	Totals	Means	Variation	d.f.	S.S.	F	
Factor II	1	20	18	21	23	20	102	20.4	Rows	3	76.95	18.86**
	2	19	18	17	18	18	90	18.0	Columns	4	8.50	1.57
	3	23	21	22	23	20	109	21.8	Residual	12	16.30	
	4	17	16	18	16	17	84	16.8				
Totals		79	73	78	80	75	385		Total	19	101.75	
Means		19.75	18.25	19.50	20.00	18.75	19.25					

FYI software such as SPSS is designed for analysing data that is recorded with variables in columns and individual observations in the rows. Thus the ANOVA data above would be written as a set of columns or rows, e.g.

Variable	20	18	21	23	20	19	18	17	18	18	23	21	22	23	20	17	16	18	16	17
Factor 1	1	1	1	1	1	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4
Factor 2	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4

17

Examples: Different Designs: What are the Mean Squares Estimating /Testing?

- Factors, Type of Effects**
- 1-Way**

Source	dof	MSQ	E{MS}
Between k groups	k-1	$SS_B / k-1$	$\sigma^2 + n\sigma_{\epsilon}^2$
Within groups	k(n-1)	$SS_W / k(n-1)$	σ^2
Total	nk-1		
- 2-Way- A,B and AB**

	Fixed	Random	Mixed
E{MS A}	$\sigma^2 + nb\kappa_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
E{MS B}	$\sigma^2 + na\kappa_B^2$	$\sigma^2 + n\sigma_{AB}^2 + na\sigma_B^2$	$\sigma^2 + na\sigma_B^2$
E{MS AB}	$\sigma^2 + n\kappa_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$
E{MS Error}	σ^2	σ^2	σ^2

Model here is $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}$

18

Nested Designs

- **Model** $Y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk}$
- **Design** p Batches (A)
 - ↓ ↓ ↓ ↓
 - Trays (B) 1 2 3 4 q
 - Replicates ↓↓↓ ↓↓↓ r per tray
- **ANOVA skeleton**

	dof	E{MS}
Between Batches	$p-1$	$\sigma^2 + r\sigma_B^2 + rq\sigma_A^2$
Between Trays	$p(q-1)$	$\sigma^2 + r\sigma_B^2$
Within Batches		
Between replicates	$pq(r-1)$	σ^2
Within Trays		
Total	$pqr-1$	

19

Examples in Genomic and Trait Models

- Genetic traits may be controlled by No. of genes - usually unknown
Taking this "genetic effect" as one genotypic term, a simple model
 $y_{ij} = \mu + G_i + \varepsilon_{ij}$ for $y \sim N(\mu, \sigma_p^2), G \sim N(0, \sigma_g^2), \varepsilon \sim N(0, \sigma_e^2)$
where the y_{ij} is the trait value for genotype i in replication j , μ is the mean, G_i is the genetic effect for genotype i and ε_{ij} the errors.

- If assume Normality (and interested in Random effects) and zero covariance between genetic effects and error

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2$$

and if the same genotype is replicated b times in an experiment, with phenotypic means used, the error variance is averaged over b .

20

Example contd.

- **HERITABILITY** = Ratio genotypic to phenotypic variance

$$H = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

- Depending on relationship among genotypes, interpretation of genotypic variance differs. May contain additive, dominance, epistatic interactions, variances $\sigma_a^2, \sigma_d^2, \sigma_i^2$ (Above = **broad sense heritability**).
- For some experimental or mating schemes, an additive genetic variance may be calculated. **Narrow sense heritability** then

$$H = \frac{\sigma_a^2}{\sigma_p^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_d^2 + \sigma_i^2 + \sigma_e^2}$$

- Again, if phenotypic means used, can obtain a **mean-based heritability** for b replications.

$$H = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2 / b}$$

21

Extended Example:

$$y_{ij} = \mu_i + G_{ij} + \varepsilon_{ij}$$

$$\text{Have } y_{2j} = \mu_2 + G_{2j} + \varepsilon_{2j}$$

where 1 and 2 denote traits, i the gene and j an individual in population.

Then y is the trait value, μ overall mean, G genetic effect, ε random error.

To quantify relationship between the two traits, the variance covariance matrices for phenotypic, Σ_p genetic Σ_g and environmental effects Σ_e

$$\Sigma_p = \begin{bmatrix} \sigma_{p1}^2 & \sigma_{p12} \\ \sigma_{p12} & \sigma_{p2}^2 \end{bmatrix} = \Sigma_g + \Sigma_e = \begin{bmatrix} \sigma_{g1}^2 & \sigma_{g12} \\ \sigma_{g12} & \sigma_{g2}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{e1}^2 & \sigma_{e12} \\ \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix}$$

So **correlations** between traits in terms of phenotypic, genetic and environmental effects are:

$$\rho_p = \frac{\sigma_{p12}}{\sqrt{\sigma_{p1}^2 \sigma_{p2}^2}}; \quad \rho_g = \frac{\sigma_{g12}}{\sqrt{\sigma_{g1}^2 \sigma_{g2}^2}}; \quad \rho_e = \frac{\sigma_{e12}}{\sqrt{\sigma_{e1}^2 \sigma_{e2}^2}}$$

e.g. due to linkage of controlling genes, same gene controlling both traits

22