## 4. Analysis of Variance

### 4.1 **Introduction**

So far we have looked at how to compare two population means. We now look at a method for comparing several population means at the same time.

Suppose we have k populations of interest. The method of analysis of variance (ANOVA) enables us (under suitable conditions) to test the hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: not\ all\ the\ means\ are\ equal$$

### 4.2 **One way Analysis of Variance (Completely randomized design)**

This refers to a situation where there is one factor with a number of levels e.g. measure baking at 3 temperature levels (low, medium, high), in this case the factor in temperature and the levels are low, medium and high.The levels of a factor are often called treatments.

Note:

For the ANOVA procedure to be valid we either

1. pick random samples from k populations
   or
2. randomly assign subjects or units to k treatments (this balances out known and unknown factors)

#### 4.2.1 Notation

The observations from the k treatments can be represented in a table below where $y_{ij}$ represents the $j^{th}$ observation from the $i^{th}$ treatment.

Treatment

| | 1 | 2 | ... | k | Total |
|---|---|---|---|---|---|
| | $y_{11}$<br>$y_{12}$<br>⋮<br>$y_{1n}$ | $y_{21}$<br>$y_{22}$<br>⋮<br>$y_{2n}$ | ⋯<br>⋯<br><br>... | $y_{k1}$<br>$y_{k2}$<br>⋮<br>$y_{kn}$ | |
| Total | $y_{1\bullet}$ | $y_{2\bullet}$ | ... | $y_{k\bullet}$ | $y_{\bullet\bullet}$ |

where

$$y_{i\bullet} = \sum_{j=1}^{n} y_{ij} \quad , \quad \bar{y}_{i\bullet} = \frac{y_{i\bullet}}{n} \ \ (mean\ for\ treatment\ i)$$

$$y_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} \quad , \quad \bar{y}_{\bullet\bullet} = \frac{y_{\bullet\bullet}}{nk} \ \ (overall\ mean\ )$$

To test the hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: not\ all\ the\ means\ are\ equal$$

we assume that the k populations are independent and normally distributed with means $\mu_1$, $\mu_2, \cdots, \mu_k$ and a common variance $\sigma^2$. This can be written as a model as

$$Y_{ij} = \mu_i + \varepsilon_{ij}\,, i = 1, 2, \ldots, k$$
$$j = 1, 2, \ldots, n$$

where

$Y_{ij} = j^{th}\, response\ from\ i^{th}\ population$
$\mu_i = mean\ of\ the\ i^{th}\ population$
$\varepsilon_{ij} = random\ error$ and
$\varepsilon_{ij} \sim N(0, \sigma^2)$

4.2.2 ANOVA table

To test the hypothesis

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1: not\ all\ the\ means\ are\ equal$$

we construct an ANOVA table as follows:

| Source of variation (Source) | Sum of squares (SS) | Degrees of freedom (df) | Mean squares (MS) | $F^*$ |
|---|---|---|---|---|
| Treatments | SSTrt | $k-1$ | $MSTrt = \frac{SSTrt}{k-1}$ | $\frac{MSTrt}{MSE}$ |
| Error | SSE | $N-k$ | $MSE = \frac{SSE}{N-k}$ | |
| Total | SST | $N-1$ | | |

where

$SSTrt = Treatment\ sum\ of\ squares$
$SSE = Error\ sum\ of\ squares$
$SST = Total\ sum\ of\ squares$
$MSTrt = Mean\ treatment\ sum\ of\ squares$
$MSE = Mean\ square\ error$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{N}$$

$$SSTrt = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$
$$= \sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n} - \frac{y_{\bullet\bullet}^2}{N}$$
$$SSE = SST - SSTrt$$

where $N = nk$ is the total number of observations.

We reject $H_0$ if $F^* > F_\alpha(k-1, N-k)$.

Example

A psychologist was interested in the effects of 3 kinds of drugs on the mean time to complete a task. She randomly assigned 5 subjects to each drug A, B and C and measured the time (in minutes) to complete the task. Are the population mean times the same for each drug?

Drug

| A | B | C |
|---|---|---|
| 20 | 21 | 30 |
| 22 | 26 | 24 |
| 25 | 26 | 26 |
| 24 | 27 | 25 |
| 19 | 25 | 30 |

Soln

Let

$\mu_1 = mean\ time\ for\ drug\ A$
$\mu_2 = mean\ time\ for\ drug\ B$
$\mu_3 = mean\ time\ for\ drug\ C$

1. Testing problem
   $H_0: \mu_1 = \mu_2 = \mu_3$
   $H_1: not\ all\ the\ 3\ means\ are\ equal$

2. Test statistic
   $F = \frac{MSTrt}{MSE}$

3. Computation

| A | B | C | |
|---|---|---|---|
| 20 | 21 | 30 | |
| 22 | 26 | 24 | |
| 25 | 26 | 26 | |
| 24 | 27 | 25 | |
| 19 | 25 | 30 | |
| 110 | 125 | 135 | 370 |

$n = 5 \quad N = 15 \quad y_{1\bullet} = 110 \quad y_{2\bullet} = 125 \quad y_{3\bullet} = 135 \quad y_{\bullet\bullet} = 370$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{N}$$
$$= 20^2 + 22^2 + \cdots + 30^2 - \frac{370^2}{15}$$
$$= 9270 - \frac{370^2}{15}$$
$$= 143.33$$

$$SSTrt = \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n} - \frac{y_{\bullet\bullet}^2}{N}$$
$$= \frac{110^2}{5} + \frac{125^2}{5} + \frac{135^2}{5} - \frac{370^2}{15}$$
$$= 63.33$$

$$SSE = SST - SSTrt = 143.3 - 63.33 = 80$$

| Source | SS | df | MS | $F^*$ |
|---|---|---|---|---|
| Treatments | 63.33 | 2 | 31.665 | 4.75 |
| Error | 80 | 12 | 12 | |
| Total | 143.33 | 14 | | |

4. Critical region
   reject $H_0$ if
   $$F^* > F_\alpha(k-1, N-1) = f_{0.05}(2,12) = 3.89$$

   Note that the level of significance is not given, so we use a standard value of $\alpha = 0.05$
   i.e. $F^* > 3.89$

5. Conclusion
   Reject $H_0$ at the 0.05 level of significance. The mean times are not the same.

Note

1. When the number of observations is the same for all the treatments, we have a balanced design and $N = nk$ as above.
2. When it is not the same then we have an unbalanced design and $N = n_1 + n_2 + \cdots + n_k$ where $n_i = number\ of\ observations\ for\ treatment\ i$.
   The formulae for the sum of squares then become:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\bullet\bullet})^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{N}$$

$$SSTrt = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$$
$$= \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n_i} - \frac{y_{\bullet\bullet}^2}{N}$$
$$SSE = SST - SSTrt$$

Example

A production plant manager claimed that there was no difference in the mean times to complete as assembly line job among plants A, B, C and D. Samples from each of the plants yielded the following times (in minutes) to complete the job.

Plant

| A | B | C | D |
|---|---|---|---|
| 18 | 20 | 23 | 12 |
| 11 | 14 | 16 | 18 |
| 14 | 16 | 21 | 17 |
| 12 | 18 | | 13 |
| 15 | | | |

Test the claim using a 1% level of significance.

Soln

Let

$\mu_1 = mean\ time\ for\ plant\ A$
$\mu_2 = mean\ time\ for\ plant\ B$
$\mu_3 = mean\ time\ for\ plant\ C$
$\mu_4 = mean\ time\ for\ plant\ D$

1. Testing problem
   $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
   $H_1: not\ all\ the\ means\ are\ equal$

2. Test statistic
   $F = \dfrac{MSTrt}{MSE}$

3. Computation

| A | B | C | D |
|---|---|---|---|
| 18 | 20 | 23 | 12 |
| 11 | 14 | 16 | 18 |
| 14 | 16 | 21 | 17 |
| 12 | 18 | | 13 |
| 15 | | | |
| 70 | 68 | 60 | 60 |

$$n_1 = 5 \quad n_2 = 4 \quad n_3 = 3 \quad n_4 = 4 \quad N = 16$$
$$y_{1\bullet} = 70 \quad y_{2\bullet} = 68 \quad y_{3\bullet} = 60 \quad y_{4\bullet} = 60 \quad y_{\bullet\bullet} = 258$$

$$\begin{aligned} SSTrt &= \sum_{i=1}^{k} \frac{y_{i\bullet}^2}{n_i} - \frac{y_{\bullet\bullet}^2}{N} \\ &= \frac{70^2}{5} + \frac{68^2}{4} + \frac{60^2}{3} + \frac{60^2}{4} - \frac{258^2}{16} \\ &= 75.75 \end{aligned}$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{N}$$

$$= 18^2 + 11^2 + \cdots + 13^2 - \frac{258^2}{16}$$

$$= 4338 - \frac{258^2}{16}$$

$$= 177.75$$

$$SSE = SST - SSTrt = 102$$

| Source | SS | df | MS | $F^*$ |
|---|---|---|---|---|
| Treatments | 75.75 | 3 | 25.25 | 2.97 |
| Error | 102 | 12 | 8.5 | |
| Total | 177.75 | 15 | | |

4. Critical region

   reject $H_0$ if

   $$F^* > F_\alpha(k-1, N-1) = f_{0.01}(3,12) = 5.95$$

   i.e. $F^* < 5.95$

5. Conclusion

   Do not reject $H_0$ at the 0.01 level of significance. The mean times in the 4 plants are not significantly different.