Analysis of variance (ANOVA)

One-way analysis of variance

Suppose that we want to see the effect of a categorical variable or "factor" A (with m categories or levels) on a continuous response Y. We have n_i observations on the ith category ($i = 1, 2, \cdots, m$). Let

$$Y_{ij} \sim N(\mu_i, \sigma^2), i = 1, 2, \cdots, m; \;\; j = 1, 2, \cdots, n_i.$$

We want to test the hypotheses

$$H_0: \; \mu_1 = \mu_2 = \cdots = \mu_m$$

$$H_1: \; \text{At least two means are unequal}$$

We can think of a model of the form

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where α_i is the effect of the ith level of the factor (often called the ith "treatment"). In fact, $\mu_i = \mu + \alpha_i$. The error term $\epsilon_{ij} \sim N(0, \sigma^2)$.

The hypotheses written above are equivalent to the following hypothses:

$$H_0: \; \text{All the } \alpha_i \text{ are zero.}$$

$$H_1: \; \text{Not all } \alpha_i \text{ are zero.}$$

The total number of observations:

$$N = \sum_{i=1}^{m} n_i$$

The sample mean for the ith category:

$$\bar{Y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

The overall sample mean:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}$$

It can be shown that

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2+\sum_{i=1}^{m}n_i(\bar{Y}_i-\bar{Y})^2$$

That is,

$$SS_T=SS_E+SS_A$$

Thus, "Total Sum of Squares" can be split into "Error SS" (Within group SS) and "SS due to A" (Between group SS). When the m group means are different, SS_A is significantly more than SS_E.

One-way ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom
A	$SS_A=\sum_{i=1}^{m}n_i(\bar{Y}_i-\bar{Y})^2$	$m-1$
Error	$SS_E=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$	$N-m$
Total	SS_T	$N-1$

The test statistic is

$$F=\frac{SS_A/(m-1)}{SS_E/(N-m)}$$

which follows F distribution with $m-1$ and $N-m$ d.f. under the null hypothesis. F-tests of this type are always upper-tailed.

Estimate of σ^2 is given below (whether null hypothesis is true or not):

$$\widehat{\sigma^2}=\frac{SS_W}{N-m}$$

Example

An auto rental firm is using 15 identical motors that are adjusted to run at a fixed speed to test 3 different brands of gasoline. Each brand of gasoline is assigned to exactly 5 of the motors. Each motor runs on 10 gallons of gasoline until it is out of fuel. The following represents the total mileages obtained by the different motors:

Gas 1: 220, 251, 226, 246, 260
Gas 2: 244, 235, 232, 242, 225
Gas 3: 252, 272, 250, 238, 256

Does the type of gasoline used affect the average mileage obtained?

Solution

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

H_1: At least two means are unequal

Do the calculations yourself and compare with the following results:

Calculated $F = 2.60$

Table value of $F_{0.95,2,12} = 3.89$ (According to textbook, $F_{0.05,2,12} = 3.89$)

We cannot reject H_0. We cannot say that the type of gas affects the mileage.

Note

When the null hypothesis is rejected, we can perform "multiple comparisons", i.e., we can test the equality of each pair of means.

Two-factor analysis of variance

Suppose that we want to see the effect of two categorical variables or factors A and B on a continuous response Y. Factor A has m levels, while factor B has n levels. We have the following data

$$Y_{ij}, \quad i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, n.$$

We can think of a model of the form

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where α_i is the effect of the ith level of factor A, and β_j is the effect of the jth level of factor B. The error term $\epsilon_{ij} \sim N(0, \sigma^2)$.

We want to test the hypotheses related to factor A:

$$H_0\text{: All the } \alpha_i \text{ are zero.}$$

$$H_1\text{: Not all } \alpha_i \text{ are zero.}$$

We may also want to test the hypotheses related to factor B:

$$H_0\text{: All the } \beta_j \text{ are zero.}$$

$$H_1\text{: Not all } \beta_j \text{ are zero.}$$

It can be shown that,

$$SS_T = SS_A + SS_B + SS_E$$

where

$$SS_A = n\sum_{i=1}^{m}(\bar{Y}_{i.} - \bar{Y})^2$$

$$SS_B = m\sum_{i=1}^{n}\left(\bar{Y}_{.j} - \bar{Y}\right)^2$$

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}\right)^2$$

Two-way ANOVA table

Source of Variation	Sum of Squares	Degrees of Freedom
A	$SS_A = n\sum_{i=1}^{m}(\bar{Y}_{i.} - \bar{Y})^2$	$m - 1$
B	$SS_B = m\sum_{i=1}^{n}\left(\bar{Y}_{.j} - \bar{Y}\right)^2$	$n - 1$
Error	$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{\iota}\right)^2$	$mn - m - n + 1$ $= (m-1)(n-1)$
Total	SS_T	$mn - 1$

The test statistic for testing the effect of factor A is

$$F = \frac{SS_A/(m-1)}{SS_E/((m-1)(n-1))}$$

which follows F distribution with $m-1$ and $(m-1)(n-1)$ d.f. under the null hypothesis.

The test statistic for testing the effect of factor B is

$$F = \frac{SS_B/(n-1)}{SS_E/((m-1)(n-1))}$$

which follows F distribution with $n-1$ and $(m-1)(n-1)$ d.f. under the null hypothesis.

Example

Three different washing machines were employed to test four different detergents. The following data give a score of the effectiveness of each washing.

	Machine		
	1	2	3
Detergent 1	53	50	59
Detergent 2	54	54	60
Detergent 3	56	58	62
Detergent 4	50	45	57

Does the detergent used affect the score? Does the machine used affect the score?

Solution

Do it yourself