ANalysis Of Variance (ANOVA)

By- Dr Ekta Khare

ANOVA

- The general statistical rationale behind ANOVA lies in the comparison of two variances:
  - the variance between the treatments and
  - the variance within the treatments.
- If the between- and within-variances are equal, then no treatment effect is present.
- ANOVA variances are calculated by computing the sum of squares, e.g., $\sum(x-\bar{x})^2$, and dividing by the set degrees of freedom, which converts the sum of squares calculation into a variance estimate.
- The mean square error (MS_E) term in ANOVA is the variance estimate, σ^2_E, within the treatments, that is, the random variability.
- The mean square treatment (MS_T) term consists of the variance estimate, σ^2_T, between the treatments; that is, the treatment effect.
- If the H_0 hypothesis is true, there is no treatment effect – then both terms (MS_T and MS_E) are unbiased estimates of the random variability, and no treatment effect is present.
- That is, $MS_T \approx MS_E$

Types of ANOVA Tests

- There are different types of ANOVA tests. The two most common are:
  - a “One-Way” and
  - a “Two-Way.”
- The difference between these two types depends on the number of independent variables in your test.

One-way ANOVA

A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.

The one-way ANOVA test for differences in the means of the dependent variable is broken down by the levels of the independent variable.

An example of a one-way ANOVA includes testing a therapeutic intervention (CBT, medication, placebo) on the incidence of depression in a clinical sample.

Two-way (factorial) ANOVA

A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

An example of a factorial ANOVAs include testing the effects of social contact (high, medium, low), job status (employed, self-employed, unemployed, retired), and family history (no family history, some family history) on the incidence of depression in a population.

One Way ANOVA Example

- A microbiologist is working with a new antibiotic drug, which has proved very effective on bacteria producing β-lactamase, specifically Staphylococcus aureus, Methicillin-resistant strains. Five replicate drug samples will be used with four different preservatives. The study will expose each drug sample to a 10 mL suspension of 1.0 × 105 organisms per mL for twenty (20) min. The bacterial cells will then be cleansed of any extracellular drugs, from baseline will then be calculated and compared, using a completely randomized, one-factor analysis of variance method.
- After the study was conducted, according to the run order, the colony count values were recorded in log10 scale (Table 4.6).

		Treatments			
		A	B	C	D
Replicates	1	3.17	2.06	2.27	4.17
	2	2.91	3.21	3.78	4.01
	3	4.11	2.57	3.59	3.92
	4	3.82	2.31	4.01	4.29
	5	4.02	2.71	3.15	3.72

Treatment A	Treatment B	Treatment C	Treatment D
$(x_{ij}-\bar{x}_{\cdot 1})^2$	$(x_{ij}-\bar{x}_{\cdot 2})^2$	$(x_{ij}-\bar{x}_{\cdot 3})^2$	$(x_{ij}-\bar{x}_{\cdot 4})^2$
$(3.17-3.61)^2 = 0.1936$	$(2.06-2.57)^2 = 0.2601$	$(2.27-3.36)^2 = 1.1881$	$(4.17-4.02)^2 = 0.0225$
$(2.91-3.61)^2 = 0.4900$	$(3.21-2.57)^2 = 0.4096$	$(3.78-3.36)^2 = 0.1764$	$(4.01-4.02)^2 = 0.0001$
$(4.11-3.61)^2 = 0.2500$	$(2.57-2.57)^2 = 0.0000$	$(3.59-3.36)^2 = 0.0529$	$(3.92-4.02)^2 = 0.0100$
$(3.82-3.61)^2 = 0.0441$	$(2.31-2.57)^2 = 0.0676$	$(4.01-3.36)^2 = 0.4225$	$(4.29-4.02)^2 = 0.0729$
$(4.02-3.61)^2 = 0.1681$	$(2.71-2.57)^2 = 0.0196$	$(3.15-3.36)^2 = 0.0441$	$(3.72-4.02)^2 = 0.0900$
$\sum(x_{ij}-\bar{x}_{\cdot 1})^2 = 1.1458$	$\sum(x_{ij}-\bar{x}_{\cdot 2})^2 = 0.7569$	$\sum(x_{ij}-\bar{x}_{\cdot 3})^2 = 1.8840$	$\sum(x_{ij}-\bar{x}_{\cdot 1})^2 = 0.1955$
$\bar{x}_A = \frac{18.03}{5} = 3.61$	$\bar{x}_B = \frac{12.86}{5} = 2.57$	$\bar{x}_C = \frac{16.80}{5} = 3.36$	$\bar{x}_D = \frac{20.11}{5} = 4.02$

Step 1. Find the mean values $(\bar{x}_{\cdot j})$ of each treatment. This is done using the information computed in Table 4.7. In this example, there are 4: $\bar{x}_A$, $\bar{x}_B$, $\bar{x}_C$, and $\bar{x}_D$. Next, find the grand mean, or the mean of the means ($\bar{\bar{x}}$).

$$\bar{\bar{x}} = \frac{\bar{x}_A + \bar{x}_B + \bar{x}_C + \bar{x}_D}{4} = \frac{3.61 + 2.57 + 3.36 + 4.02}{4} = 3.39$$

Step 2. Find $(x_{ij} - \bar{x}_{.j})^2$, or the difference within each treatment, the error estimate. This is the individual treatment values minus the means of the individual treatments for that group. These individual values were determined in Table 4.7.

$$SS_{Error} = \sum\sum(x_{ij} - \bar{x}_{.j})^2 = (1.1458 + 0.7569 + 1.8840 + 0.1955) = 3.982$$

Step 3. Find the SS_{Total}. This is simply the grand mean subtracted from each x_{ij} value squared, then summed.

$$SS_T = \sum\sum(x_{ij} - \bar{\bar{x}})^2$$
$$= (3.17 - 3.39)^2 + (2.91 - 3.39)^2 + \ldots + (4.29 - 3.39)^2 + (3.72 - 3.39)^2$$
$$= 9.563$$

Step 4. Find the $SS_{Treatment}$ by subtraction.

$$SS_{Total} - SS_{Error} = SS_{Treatment}$$
$$9.563 - 3.982 = 5.581 = SS_{Treatment}$$

Source	DF	Sum of squares	Mean square	F_c
Treatment	$c - 1$	$SS_{Treatment}$	$\frac{SS_{Treatment}}{c-1} = MS_{Treatment}$	$\frac{MS_{Treatment}}{MS_{Error}} = F_c$
Error	$c(r-1)$	SS_{Error}	$\frac{SS_{Error}}{c(r-1)} = MS_{Error}$	
Total	$rc - 1$	SS_{Total}		

DF = degrees of freedom. The sum of squares must be averaged by the appropriate degrees of freedom; c = number of columns or treatments; r = number of rows or replicates; $c = 4$; $c - 1 = 4 - 1 = 3$ degrees of freedom; $c(r - 1) = 4(5 - 1) = 4 \times 4 = 16$ degrees of freedom; $rc - 1 = 5 \times 4 - 1 = 20 - 1 = 19$, or just add degrees of freedom for treatments and error, $3 + 16 = 19$.

Source	DF	Sum of squares	Mean square	F_c
Treatment	3	5.581	5.581/3 = 1.860	1.860/0.249 = 7.47
Error	16	3.982	3.982/16 = 0.249	
Total	19	9.563		

The calculated F value is $F_c = 7.47$.

Step 6. Make the decision. If $F_c > F_t$, reject H_0 at the α level.

Because $F_c = 7.47 > F_t = 2.46$, $F_{t(\alpha;\, c-1,\, c(r-1))} = F_{t(0.10;\, 3,\, 16)} = 2.46$ (from Table A.4). Reject H_0 at $\alpha = 0.10$. In fact, looking at the F table (Table A.4) at 3, 16 degrees of freedom, F_c is larger than F_t at $\alpha = 0.01 = 5.29$. Hence, $p < 0.01$, which is highly significant.

Contrasts

- The ANOVA table has clearly demonstrated that significant differences exist among treatments, but which one or ones?
- We will use the Tukey method, for it is neither too liberal, nor too conservative.
- The Tukey method requires the microbiologist to use a Studentized range value (q_{α}; *a, f).*
- where α is the Type I error level, a is the number of treatments or groups, and f = N – a, the total number of observations minus the number oftreatment groups, a.
- All possible treatment pairs are compared to one another.
- If $|\bar{x}_i - \bar{x}_j| > q_{(\alpha;a,f)} s_x$, reject H_0.
- The two treatments $(\bar{x}_i \text{ and } \bar{x}_j)$ differ significantly at α.

The s_x value is computed as:

$$s_x = \sqrt{s^2/n} = \sqrt{MS_E/n}$$

where n = number of replicates per sample set.

- MS_E = 2.49 (from Table)

$$s_x = \sqrt{\frac{MS_E}{n}} = \sqrt{\frac{0.249}{5}} = 0.2232$$

Let us use $\alpha = 0.05$, $a = 4$, $f = N - a = 20 - 4 = 16$, and $q_{(0.05;\ 4,\ 16)} = 4.05$ (from Table A.3). $q\, s_x = 4.05(0.2232) = 0.9040$ = the critical value.
If $|\bar{x}_i - \bar{x}_j| > q\, s_x$, the pairs differ at α.

Mean Difference	$q\, s_x$	Significant/ Not Significant at α
$\lvert\bar{x}_A - \bar{x}_B\rvert = \lvert 3.606 - 2.572\rvert = 1.034$	> 0.9040	Significant
$\lvert\bar{x}_A - \bar{x}_C\rvert = \lvert 3.606 - 3.360\rvert = 0.300$	$\ngtr 0.9040$	Not Significant
$\lvert\bar{x}_A - \bar{x}_D\rvert = \lvert 3.606 - 4.022\rvert = 0.416$	$\ngtr 0.9040$	Not Significant
$\lvert\bar{x}_B - \bar{x}_C\rvert = \lvert 2.572 - 3.360\rvert = 0.788$	$\ngtr 0.9040$	Not Significant
$\lvert\bar{x}_B - \bar{x}_D\rvert = \lvert 2.572 - 4.022\rvert = 1.450$	> 0.9040	Significant
$\lvert\bar{x}_C - \bar{x}_D\rvert = \lvert 3.606 - 4.022\rvert = 0.660$	$\ngtr 0.9040$	Not Significant