⇒ Analysis of variance

(1)

Introduced by Sir R A Fisher (1890-1962) in 1923. (Abbreviated as ANOVA)

Analysis of variance is a technique that partitions the total variation - a term distinct from variance and measured by the sum of squares of deviations from the mean - into its component parts each of which is associated with a differente source of variation.

The analysis of variance ~~pred~~ procedure therefore compares two different estimates of variance by using F-distribution to determine wether the Population means are equal.

⇒ One-way Analysis of variance

It is also called the one-variable of classification analysis of variance.

Hypothesis

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

H_1: Not all means are equal.

Data General table:

Observation	Samples (or Treatments) 1	2	...	j	...	K	Total
1	x_{11}	x_{12}		x_{1j}		x_{1K}	
2	x_{21}	x_{22}		x_{2j}		x_{2K}	
⋮	⋮	⋮		⋮		⋮	
i	x_{i1}	x_{i2}		x_{ij}		x_{iK}	
⋮	⋮	⋮		⋮		⋮	
r	x_{r1}	x_{r2}	...	x_{rj}		x_{rK}	
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	...	$T_{\cdot j}$	...	$T_{\cdot K}$	$T_{\cdot\cdot}$
Mean	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$		$\bar{X}_{\cdot j}$		$\bar{X}_{\cdot K}$	$\bar{X}_{\cdot\cdot}$ → Grand mean

∵ k samples of equal size r

Analysis of variance Table

Source of variation	d.f	Sum of squares (SS)	Mean Square (MS)	computed F
Between samples	$k-1$	$SSB = \frac{\Sigma T_{\cdot j}^2}{r} - C.F$	$S_b^2 = \frac{SSB}{k-1}$	$F = \frac{S_b^2}{S_\omega^2}$
Within samples (Error)	$n-k$	$SSE = TSS - SSB$	$S_\omega^2 = \frac{SSE}{n-k}$	
Total	$n-1$	$SST = \Sigma\Sigma X_{ij}^2 - C.F$	$S_T^2 = \frac{SST}{n-1}$	

C.F (correction factor) $= \frac{T_{..}^2}{n}$

Six step

1) $H_0 : \mu_1 = \mu_2 = \cdots \mu_K$
   H_1 : Not all k means are equal

2) Decide upon a significance level α

3) Test statistics: $F = \frac{S_b^2}{S_\omega^2}$

S_b^2 and S_ω^2 are the two estimates of the common variance σ^2, if H_0 is true. has an F-dist with $v_1 = k-1$ and $v_2 = n-k$ degree of freedom.

4) Compute the necessary sums of square and complete the analysis of variance table.

5) Determine the critical region which will consist of all values greater than or equal to $F_\alpha(k-1, n-k)$

6) Decide
   Reject H_0 if F falls in the critical region, accept H_0 otherwise.

...by choosing a convenient origin as all the SS are independent of origin.

Example 20.1 Given the data below, test the hypothesis that the means of the three populations are equal. Let $\alpha = 0.05$.

Sample 1	Sample 2	Sample 3
40	70	45
50	65	38
60	66	60
65	50	42

(i) We state our null and alternative hypotheses as

$H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.* all the three means are equal, and

H_1 : Not all three means are equal.

(ii) The significance level is set at $\alpha = 0.05$.

(iii) The test-statistic to use is

$$F = \frac{s_b^2}{s_w^2},$$

which, if H_0 is true, has an F-distribution with $v_1 = k-1$ and $v_2 = n-k$ degrees of freedom.

	Sample 1 X_{i1} (X_{i1}^2)	Sample 2 X_{i2} (X_{i2}^2)	Sample 3 X_{i3} (X_{i3}^2)	Total	$\sum_j X_{ij}^2$
	40 (1600) 50 (2500) 60 (3600) 65 (4225)	70 (4900) 65 (4225) 66 (4356) 50 (2500)	45 (2025) 38 (1444) 60 (3600) 42 (1764)	--- --- --- ---	8525 8169 11556 8489
$T_{\cdot j}$	215	251	185	651	36739
$T_{\cdot j}^2$	46225	63001	34225	143451	↑
$\sum_i X_{ij}^2$	11925	15981	8833	36739	← check

$$\text{Correction Factor } (C.F.) = \frac{T_{..}^2}{n} = \frac{(651)^2}{12} = 35316.75$$

Total $SS = \sum_i \sum_j X_{ij}^2 - C.F.$

$$= 36739 - 35316.75 = 1422.25$$

Between $SS = \dfrac{\sum_j T_{\cdot j}^2}{r} - C.F.$

$$= \frac{143451}{4} - 35316.75 = 546.00, \text{ and}$$

Within SS = Total SS − Between SS = 1422.25 − 546.00 = 876.25.

The Analysis of Variance table is:

Source of Variation	$d.f.$	Sum of Squares	Mean Square	Computed F
Between Samples	2	546.00	273.00	$\frac{273.00}{97.36} = 2.80$
Within Samples	9	876.25	97.36	----
Total Variation	11	1422.25	---	---

(v) The critical region is $F \geq F_{0\cdot 05}(2, 9) = 4.26$

(vi) **Conclusion.** Since the calculated value of $F = 2.80$ does not fall in the critical region, so we accept our null hypothesis and conclude that all the three means are equal.

Table 19.1 Percent Points of the F-Distribution

5 Per cent Points of F, i.e. $F_{0.05}(v_1, v_2)$

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.73	1.52	1.00

... found by interchange of v_1 and v_2, i.e., v_1