ANOVA
(ANALYSIS OF VARIANCE)

Dr. D. K. Paul

Associate Professor & Course Coordinator
M. Sc. Environmental Science & Management
Department of Zoology, Patna University
Member, SEAC (for EIA), Bihar, Constituted by
MoEF & CC, Govt. of India
dkpaul.pat31@gmail.com

INTRODUCTION

- Having understood the variance ratio test, it is much more important test called the analysis of variance test which is not confined to comparing two sample means but more than two samples drawn from corresponding normal population.
- The statistical technique used to compare means of variations of more than two populations is called "Analysis of Variance"(ANOVA).
- The term "analysis of variance" is used because the total variability in the set of data can be broken into the sum of variability among the sample means & the variability with in the samples.
- The analysis of variance is a statistical technique specially designed to test whether the means of more than 2 quantities populations are equal or not.
- This technique was developed by R.A. Fisher in 1920.

Assumption in Analysis of Variance

The assumptions in analysis of variance are the same as F-test.

i. **Normality**: i.e. The values in each group are normally distributed.

ii. **Homogeneity**: i.e. the variance within each group should be equal for all groups. ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots \sigma_x^2$).

iii. **Independence of error**: It states that the error (variation of each value around its own group mean) should be independent for each value.

Technique of Analysis of Variance

For the sake of clarity the technique of ANOVA has been discussed separately for

a. One – Way Classification
b. Two – Way Classification

One – Way Classification

In one way classification, the data are classified according to only one criteria. The null hypothesis is

$H_0 : \mu_1 = \mu_2 = \mu_3 \ldots\ldots\ldots \mu_k$ (All means are equal).

$H_A : \mu_1 \neq \mu_2 \neq \mu_3 \ldots\ldots\ldots \mu_k$ (All means are not equal).

That is, the arithmetic means of populations from which the 'K' samples were randomly drawn are equal to one another. The steps in carrying out analysis are:

1. **Calculate variance between the samples (groups)**

Steps:

a. Calculate the means of each sample i.e. $\bar{X}_1$, $\bar{X}_2$ etc….

b. Calculate the grand average $\bar{\bar{X}}$ (x double bar) by

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{N}$$ where, N= no. of samples

c. Take the difference between the means of the various samples & the grand average.
d. Square these deviations & obtain the total which will give sum of squares between the samples denoted as SSC.
e. Divide the total obtained in step (d) by df (degrees of freedom) as (number of samples-1) i.e. if there are 4 samples, then df = 4-3 = 3 or v=k-1, where, k= number of samples.

2. Calculate variance within the samples

Steps:

a. Calculate the mean value of each sample i.e. $\bar{X}_1, \bar{X}_2, \bar{X}_3$,etc….

b. Take the deviations of the various items in a sample from the mean values of the respective samples.

c. Square these deviations & obtain the total which gives the sum of the square within the samples denoted as SSE.

d. Divide the total obtained in step (c) by df. The df is obtained by deduction from the total number of items i.e. the number of samples i.e. v_2 = N – K, where K = no. of samples, N= total no. of all the observations.

3. Calculate the ratio F as follows.

$$F = \frac{Between\ column\ variance}{Within\ column\ variance}$$

Symbolically $F = \frac{s_1^2}{s_2^2}$

Compare the calculated value of F with the table value of F.

If the calculated value of F is greater than the table value, it is concluded that the difference in samples means is significant i.e. it could not have arisen due to fluctuations of sample sampling.

Analysis of Variance (ANOVA) Table: One way classification model

Source of variation	SS (Sum of squares)	V (degree of freedom)	MS (Mean square)	Variance Ratio (F)
Between samples	SSC	$v_1 = c - 1$	MSC = SSC/(c-1)	MSC
Within samples	SSE	$v_2 = n - c$	MSE = SSE/(n-c)	MSE
Total	SST	N - 1		

Where, SST = Total sum of squares of variations
SSC = Sum of squares between samples (columns)
SSE = Sum of squares within samples (rows)
MSC = Mean sum of squares between samples
MSE = Mean sum of squares within samples

Example of One – Way ANOVA

Example 1: To assess the significance of possible variation in performance in a certain test between the some departments (A, B, C, D) of university, a common test was given to a number of students taken at random from the senior P.G. class of each of the four departments concerned. The result are given below. Make an analysis of variance of data.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution

Step – I : Sum of squares between the samples.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

$\sum X_1 = 45$ $\sum X_2 = 50$ $\sum X_3 = 60$ $\sum X_4 = 65$

$\bar{X}_1 = 9$ $\bar{X}_2 = 20$ $\bar{X}_3 = 12$ $\bar{X}_4 = 13$

Grand mean($\bar{\bar{X}}$) $= \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3+\bar{X}_4}{N}$ (N = no. of samples)

$= \frac{9+10+12+13}{4} = \frac{44}{4} = 11$

Grand mean of all samples ($\bar{\bar{X}}$)== $\frac{45+50+60+65}{20} = \frac{220}{20} = 11$

Step II : Square of deviation of the various samples from the grand average.

Sample A $(\overline{X_1} - \overline{\overline{X}})^2$	Sample B $(\overline{X_2} - \overline{\overline{X}})^2$	Sample C $(\overline{X_3} - \overline{\overline{X}})^2$	Sample D $(\overline{X_4} - \overline{\overline{X}})^2$
$4\ (9 - 11)^2 = 4$	$1\ (10 - 11)^2 = 1$	$1\ (12 - 11)^2 = 1$	$4\ (13 - 11)^2 = 4$
4 =4	1 =1	1 =1	4 = 4
4 =4	1 =1	1 =1	4 = 4
4 =4	1 =1	1 =1	4 = 4
4 =4	1 =1	1 =1	4 = 4
20	5	5	20

Sum of the squares between the samples, SSC = 20+5+20+5= 50

Mean sum of squares between the samples, MSC = 50/(4-1) = 16.7

(df = 4-1 = 3)

Variance between samples = 16.7

Step III: Sum of squares within the samples

Sample A		Sample B		Sample C		Sample D	
X_1	$(X_1 - \overline{X_1})^2$	X_2	$(X_2 - \overline{X_2})^2$	X_3	$(X_3 - \overline{X_3})^2$	X_4	$(X_4 - \overline{X_4})^2$
8	$(8-9)^2=1$	12	$(10-12)^2=4$	18	$(12-18)^2=36$	13	$(13-13)^2=0$
10	1	11	1	12	0	9	16
12	9	9	1	16	16	2	1
8	1	14	16	6	36	16	9
7	4	4	36	8	16	15	4
$\overline{X_1}= 9$	$\sum(X_1 - \overline{X_1})^2=16$	$\overline{X_2}= 10$	$\sum(X_2 - \overline{X_2})^2=58$	$\overline{X_3}= 12$	$\sum(X_3 - \overline{X_3})^2=104$	$\overline{X_4}=13$	$\sum(X_4 - \overline{X_4})^2=30$

Total sum of the squares within the samples= 16+58+104+30=208

Mean sum of the squares within the samples=208/(20-4)=208/16=13

V_2 (v=N-K)= 20-4=16, $\overline{\overline{X}}$ =11 (df=20-4=16)

Step IV : Total variation

Sample A		Sample B		Sample C		Sample D	
X_1	$(\overline{X_1} - \overline{\overline{X}})^2$	X_2	$(\overline{X_2} - \overline{\overline{X}})^2$	X_3	$(\overline{X_3} - \overline{\overline{X}})^2$	X_4	$(\overline{X_4} - \overline{\overline{X}})^2$
8	$(8-11)^2=9$	12	1	18	49	13	4
10	1	11	0	12	1	9	4
12	1	9	4	16	25	12	1
8	9	14	9	6	25	16	25
7	16	4	49	8	9	15	16
$\sum(\overline{X_1} - \overline{\overline{X}})^2$ =36		$\sum(\overline{X_2} - \overline{\overline{X}})^2$ =63		$\sum(\overline{X_3} - \overline{\overline{X}})^2$ =109		$\sum(\overline{X_4} - \overline{\overline{X}})^2$ =50	

Step IV

- Total sum of squares (SSE) =

$$\sum(\overline{X_1}-\overline{\overline{X}})^2+\sum(\overline{X_2}-\overline{\overline{X}})^2+\sum(\overline{X_3}-\overline{\overline{X}})^2+\sum(\overline{X_4}-\overline{\overline{X}})^2$$

=36+63+109+50= 258

df= 20-1 = 19

Step V:

Source of variation	Sum of squares	df	Mean square
Between sample	50	3	167
Within sample	208	16	130
Total	258	19	

$$F = \frac{Variance\ between\ samples}{Variance\ within\ samples} = \frac{16.7}{13} = 1.285$$

Decision: The table value of F for $v_1 = 3$ & $V_2 = 16$ at 5% level of significance = 3.24. The calculated value of F is less than the table value & hence the difference in the mean values of the sample is not significant.

Inference:

The sample could have come from the same universe \ population .

Short - Cut Method

The above method of calculating the sum of squares for variance between samples & variance within the samples is not generally followed in practice because it is time consuming. So an easier method known as short- cut method is usually followed

Sample A		Sample B		Sample C		Sample D	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	164	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\sum X_1=45$	$\sum X_1^2=421$	$\sum X_2=50$	$\sum X_2^2=558$	$\sum X_3=60$	$\sum X_3^2=824$	$\sum X_4=65$	$\sum X_4^2=875$

Sum of the all the items of various samples=

$$\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4$$

$$= 45+50+60+65$$

$$= 220 = T$$

Correction factor = $T^2/N = (220)^2/N = 48400/20 = 2{,}420$

Total sum of squares = $\sum X_1 + \sum X_2 + \sum X_3 + \sum X_4 - (T^2/N)$

$$= 421+558+824+875-2420$$

$$= 2678-2420 = 258$$ (as before in long method)

Sum of squares between the samples

$$\frac{(\Sigma X_1)^2}{N} + \frac{(\Sigma X_2)^2}{N} + \frac{(\Sigma X_3)^2}{N} + \frac{(\Sigma X_4)^2}{N} - \frac{T^2}{N}$$

$$= \frac{(45)^2}{5} + \frac{(60)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420$$

$$= \frac{12350}{5} - 2420 \quad = 2470 - 2420 = 50 \text{ (as before)}$$

Sum of squares within the samples = Total sum of squares – Sum of square between samples

= 258 – 58 = 205 (as before)

Source of variation	Sum of squares	df	Mean square
Between samples	50	4 – 1 = 3	16.7
Within samples	208	20 – 4 = 16	13.0

$$F = \frac{Variance\ between\ samples}{Variance\ within\ samples} = \frac{16.7}{13} = 1.285$$

Decision: Calculated value 1.285 is less than table value (3.24, v = 4-1=3, V_2= 20-4 = 16). So, the difference between the sample is not significant.

Interference: The sample have come from the same Universe.

Two – Way ANOVA

In a, one factor analysis of variance, the treatments constitute different levels of a single factor which is controlled in the experiments. When it is believed that two independent factors might have an effect on the response variable of interest, two way ANOVA is used. In a two way classification, following table is used.

Source of variance	Sum of squares	df	Mean sum of squares	Ratio of F
Between samples	SSC	C – 1	MSC = SSC/C – 1	MSE/MSC
Between rows	SSR	r – 1	MSR = SSR/r – 1	MSE/MSR
Residual or error	SSE	(C – 1) (r – 1)	MSE = SSE/(r – 1) (C – 1)	
Total	SST	(n – 1)		

Where, SSC = Sum of squares between columns

SSR = Sum of squares between rows

SSE = Sum of squares due to error

SST = Total sum of squares

SSE = Total sum of squares – Sum of squares between columns – Sum of squares between rows

$F(V_1V_2) = \frac{MSC}{MSE}$

Where, $V_1 = C - 1$, $V_2 = (C - 1)(r - 1)$

It should be carefully noted that V_1 may not be same in both cases : in one case $V_1 = (C_1 - 1)$ & in another case $V_1 = (r - 1)$

The calculated value are compared with the table values. It the calculated value of F is greater than the table value at pre – designed level of significance, the null hypothesis is rejected, otherwise accepted.

Residual error or square = Sum of total squares – Sum of squares within samples – Sum of squares between the samples

Example of Two – Way ANOVA

A Tea company appoints 4 salesmen A , B, C, D & observes their sales in 3 seasons – Summer, Winter & Monsoon. The figures (in lakhs) are given in following data.

Seasons	Salesman				Season's Total
	A	B	C	D	
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Salesman's Total	90	93	93	96	360

1. Does the salesman significantly differ in performance?
2. Is there significant difference between seasons?

Solution: Coding of data → Subtracting 30 from each figure; we get

Seasons	A	B	C	D	Season's Total
Summer	+6	+6	-9	+5	+8
Winter	-2	-1	+1	+2	0
Monsoon	-4	-2	-1	-1	-8
Total	0	3	-9	6	Grand Total T = 0

Correction factor = $T^2/N = 0^2/12 = 0$

Correction factor = $\frac{T^2}{N} = \frac{0^2}{12} = 0$

Sum of squares between Salesman → $\frac{0^2}{3} + \frac{3^2}{3} + \frac{(-9)^2}{3} + \frac{6^2}{3} - \frac{T^2}{N}$

=42 – 0 = 42

Here v= 4-1 = 3 (for each salesman A,B,C,D there are three observations)

Sum of squares between seasons → For summer/winter/monsoon→ (4) observation

$$\frac{8^2}{4} + \frac{0^2}{4} + \frac{(-8)^2}{4} - \frac{T^2}{N}$$

=32-0 =32

Here v= 3-1 =2, for summer /winter /monsoon there are four observations

Total sum of squares → $(+6)^2+ (-2)^2+ (-4)^2+ (+6)^2+ (-1)^2+ (-2)^2+ (-9)^2+ (+1)^2+$

$(-1)^2+ (+5)^2+ (+2)^2+ (-1)^2 - \frac{T^2}{N}$

= 36+4+16+36+1+4+81+1+1+25+4+1-0

= 210-0 = 210

=V = 12 – 1 = 11

Residual or error sum of square → 210-42-32 = 210- 74 = 136

Source of variation	Sum of square	Df	Mean squares
Between columns(Salesman)	42	3	14 (MSC)
Between row (Season)	32	2	16 (MSR)
Residual	136	6	22.67 (MSE)
Total	210	11	

$F(V_1 = 3, V_2 = 6) = \frac{MSE}{MSC}$

$= \frac{22.67}{14}$

= 1.619 ; at 5% level, table value = 4.76.

$F(V_1 = 2, V_2 = 6) = \frac{MSE}{MSR}$

$= \frac{22.67}{16}$

= 1.417; at 5% level, table value = 5.14

Hence, there is no significant difference in the seasons as far as sales among different salesmen. So, salesmen & seasons are alike so far as the sales are concerned.

Thanks