For the students of

M. Com. (Applied Economics) Sem. IV

Paper: Research Methodology (Unit IV)

Note: Study material may be useful for the courses wherever Research Methodology paper is being taught.

Prepared by:
Dr. Anoop Kumar Singh
Dept. of Applied Economics,
University of Lucknow

Topic: F-TEST and Analysis of Variance (ANOVA)

Introduction

Analysis of variance (ANOVA) is statistical technique used for analyzing the difference between the means of more than two samples. It is a parametric test of hypothesis. It is a step wise estimation procedures (such as the "variation" among and between groups) used to attest the equality between two or more population means .

ANOVA was developed by statistician and eugenicist Ronald Fisher. Though many statisticians including Fisher worked on the development of ANOVA model but it became widely known after being included in Fisher's 1925 book “Statistical Methods for Research Workers”. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. ANOVA provides an analytical study for testing the differences among group means and thus generalizes the *t*-test beyond two means. ANOVA uses F-tests to statistically test the equality of means.

Concept of Variance

Variance is an important tool in the sciences including statistical science. In the Theory of Probability and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Actually, it is measured to find out the degree to which the data in series are scattered around its average value. Variance is widely used in statistics, its use is ranging from descriptive statistics to statistical inference and testing of hypothesis.

Relationship Among Variables

under the said analysis, we use to examine the differences in the mean values of the

dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables.

We take the null hypothesis that there is no significant difference between the means of different populations. In its simplest form, analysis of variance must have a dependent variable that is metric (measured using an interval or ratio scale). There must also be one or more independent variables. The independent variables must be all categorical (non-metric). Categorical independent variables are also called factors. A particular combination of factor levels, or categories, is called a treatment.

What type of analysis would be made for examining the variations depends upon the number of independent variables taken into account for the study purpose. One-way analysis of variance involves only one categorical variable, or a single factor. If two or more factors are involved, the analysis is termed n-way (eg. Two-Way, Three-Way etc.) Analysis of Variance.

F Tests

F-tests are named after the name of Sir Ronald Fisher. The F-statistic is simply a ratio of two variances. Variance is the square of the standard deviation. For a common person, standard deviations are easier to understand than variances because they're in the same units as the data rather than squared units. F-statistics are based on the ratio of mean squares. The term "mean squares" may sound confusing but it is simply an estimate of population variance that accounts for the degrees of freedom (DF) used to calculate that estimate.

For carrying out the test of significance, we calculate the ratio F, which is defined as:

$F = \frac{S_1^2}{S_2^2}$, where $S_1^2 = \frac{(X_1 - \bar{X}_1)^2}{n_1 - 1}$

And $S_2^2 = \frac{(X_2 - \bar{X}_2)^2}{n_2 - 1}$

It should be noted that S_1^2 is always the larger estimate of variance, i.e., $S_1^2 > S_2^2$

$F = \frac{Larger\ estimate\ of\ variance}{Smaller\ estimate\ of\ variance}$

$\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$

ν_1 = degrees of freedom for sample having larger variance.

ν_2 = degrees of freedom for sample having smaller variance.

The calculated value of F is compared with the table value for ν_1 and ν_2 at 5% or 1% level of significance. If calculated value of F is greater than the table value then the F ratio is considered significant and the null hypothesis is rejected. On the other hand, if the calculated value of F is less than the table value the null hypothesis is accepted and it is inferred that both the samples have come from the population having same variance.

Illustration 1: Two random samples were drawn from two normal populations and their values are:

A	65	66	73	80	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance.

(Given: F=3.36 at 5% level for ν_1=10 and ν_2 =8.)

Solution: Let us take the null hypothesis that the two populations have not the same variance.

Applying F-test:

F=$\frac{S_1^2}{S_2^2}$

A X_1	$(X_1\text{-}\bar{X}_1)$ x_1	x_1^2	B X_2	$(X_2\text{-}\bar{X}_2)$ x_2	x_2^2
65	-15	225	64	-19	361
66	-14	196	66	-17	289
73	-7	49	74	-9	81
80	0	0	78	-5	25
82	2	4	82	-1	1
84	4	16	85	2	4
88	8	64	87	4	16
90	10	100	92	9	81
92	12	144	93	10	100
			95	12	144
			97	14	196
$\sum X_1$= 720	$\sum x_1$=0	$\sum x_1^2$-=798	$\sum X_2$=913	$\sum x_2$=0	$\sum x_2^2$=1298

$\bar{X}_1 = \frac{\sum X_1}{n_1} = \frac{720}{9} = 80$;

$\bar{X}_2 = \frac{\sum X_2}{n_2} = \frac{913}{11} = 83$

$S_1^2 = \sum x_1^2/n_1\text{-}1 = \frac{798}{9-1} = 99.75$

$S_2^2 = \sum x_2^2 / n_2 - 1 = \frac{734}{11-1} = 129.8$

$F = \frac{S_1^2}{S_2^2} = \frac{99.75}{129.8} = 0.768$

At 5 percent level of significance, For $\nu_1 = 10$ and $\nu_2 = 8$, the table value of $F_{0.05} = 3.36$.

The calculated value of F is less than the table value. The hypothesis is accepted. Hence the two populations have not the same variance.

TESTING EQUALITY OF POPULATION (TREATMENT) MEANS:

ONE-WAY CLASSIFICATION

In one way classification, following steps are carrying out for computing F- ratio through most popular method i.e. short-cut method:

1. Firstly get the squared value of all the observation for different samples (column)

2. Get the sum total of sample observations as $\sum X_1$, $\sum X_2$,……. $\sum X_k$ in each column.

3. Get the sum total of squared values for each column as $\sum X_1^2$, $\sum X_2^2$,……. $\sum X_K^2$ in each column.

4. Finding the value of "T" by adding up all the sums of sample observations i.e. $T = \sum X_1 + \sum X_2 + \ldots\ldots. \sum X_k$

5. Compute the Correction Factor by the formula:

$C.F. = \frac{T^2}{N}$

6. Find out Total sum of Squares (SST) through squared values and C F:

$SST = \sum X_1^2 + \sum X_2^2, + \ldots\ldots. \sum X_K^2$ - CF

7. Find out Sum of square between the samples SSC by following formula:

$SSC = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \ldots\ldots\ldots\ldots\ldots \frac{(\sum X_K)^2}{n_k} - CF$

8. Finally, find out sum of squares within samples i.e. SSE as under:

SSE= SST-SSC

ANALYSIS OF VARIANCE (ANOVA) TABLE

Source of Variation	Sum of squares (SS)	Degrees of freedom (v)	Mean square (MS)	Variance ratio of F
Between samples	SSC	ν_1 = C-1	MSC= SSC/ν_1	

(Treatments)				F= $\frac{MSC}{MSE}$
Within samples (error)	SSE	ν_2= N-C	MSE SSE/ν_2	
Total	SST	n-1		

SSC= Sum of squares between samples (Columns)

SST= Total sum of the squares of Variations.

SSE= Sum of squares within the samples.

MSC= Mean sum of squats between samples

MSE= Mean sum of squares within samples

Illustration 2: To test the significance of variation in the retail prices of a commodity in three principal cities, Mumbai, Kolkata, and Delhi, four shops were chosen at random in each city and the prices who lack confidence in their mathematical ability observed in rupees were as follows:

Kanpur	15	7	11	13
Lucknow	14	10	10	6
Delhi	4	10	8	8

Do the data indicate that the price in the three cities are significantly different?

Solution: Let us take the null hypothesis that there is no significant difference in the prices of a commodity in the three cities.

Calculations for analysis of variance are us under:

Sample 1		Sample 2		Sample 3	
Kanpur		Lucknow		Delhi	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
15	225	14	196	4	16
7	49	10	100	10	100
11	121	10	100	8	64
13	169	6	36	8	64
$\sum x_1 = 46$	$\sum x_1^2$ =564	$\sum x_2$=40	$\sum x_2^2$ =432	$\sum x_3 = 30$	$\sum x_3^2$ =244

There are r = treatments (samples) with n_1=4, n_2= 4, n_3 = 4, and n= 12.

T= Sum of all the observations in the three samples

= $\sum x_1$+ $\sum x_2$ +$\sum x_3$ = 46+40+30 = 116

CF = Correction Factor = $\frac{T^2}{n} = \frac{(116)^2}{12}$ = 1121.33

SST = Total sum of the squares

= $(\sum x_1^2 + \sum x_2^2 + \sum x_3^2)$ – CF = (564+ 432+ 244)-1121.33 = 118.67

SSC= Sum of the squares between the samples

= $[\frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3}]$ – CF

=$[\frac{(46)^2}{4} + \frac{(40)^2}{4} + \frac{(30)^2}{4}]$ – 1121.33

= $[\frac{2116}{4} + \frac{1600}{4} + \frac{900}{4}]$ – 1121.33

= $\frac{4616}{4}$ – 1121.33 = 32.67

SSE= SST-SSC= 118.67-32.67= 86

Degrees of freedom: df_1 = r-1= 3-1 = 2 and df_2 =n-r= 12-3=9

Thus MSTR= $\frac{SSTR}{df_1} = \frac{32.67}{2}$ = 16.33 and MSE= $\frac{SSE}{df_2} = \frac{86}{9}$ = 9.55

ANOVA TABLE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Test-Statistic
Between Samples	32.67 = SSTR	2=r-1	MSC= $\frac{SSTR}{r-1}$ = $\frac{32.67}{2}$ = 16.335	F= $\frac{MSC}{MSE}$ =$\frac{16.335}{9.55}$ = 1.71
Within Samples	86= SSE	9=n-r	MSE= $\frac{SSE}{n-r} = \frac{86}{9}$ =9.55	
Total	118.67=SST	11=n-1		

The table value of F for df_1 =2 , df_2 = 9, and α = 5% level of significance is 4.26. Since calculated value of F is less than its critical (or table) value, the null hypothesis is accepted. Hence we conclude that prices of a commodity in three cities have no significant difference.

TESTING EQUALITY OF POPULATION (TREATMENT) MEANS: TWO-WAY CLLASIFICATION

ANOVA TABLE FOR TWO-WAY CLASSIFICATION

Source of Variation	Sum of square	Degrees of Freedom	Mean Square	
Between columns	SSC	c-1	MSC= SSTR/(c-1)	$F_{treatment}$ =MSC/MSE
Between rows	SSR	r-1	MSR= SSR/(r-1)	F_{blocks}= MSR/MSE
Residual error	SSE	(c-1)(r-1)	MSE= SSE/(c-1)(r-1)	
Total	SST	n-1		

Total variation consists of three parts: (i) variation between columns, SSC; (ii) variation between rows, SSR; and (iii) actual variation due to random error, SSE. That is,

SST=SSC+(SSR+SSE).

The degrees of freedom associated with SST are cr-1, where c and r are the number of columns and rows, respectively.

Degrees of freedom between columns= c-1

Degrees of freedom between rows= r-1

Degrees of freedom for residual error=(c-1)(r-1)

The test-statistic F for analysis of variance is given by:

$F_{treatment}$ = MSC/MSE; MSC > MSE or MSE/MSC; MSE > MSC

F_{blocks} = MSR/MSE; MSR > MSE or MSE/MSR; MSE>MSR.

Illustration 3: The following table gives the number of refrigerators sold by 4 salesmen in three months May, June and July:

Month	Salesman			
	A	B	C	D
March	50	40	48	39
April	46	48	50	45
May	39	44	40	39

Is there a significant difference in the sales made by the four salesmen? Is there a significant difference in the sales made during different months?

Solution: Let us take the following null hypothesis:

H_O : There is no significant difference in the sales made by the four salesmen.

H_O : There is no significant difference in the sales made during different months.

The given data are coded by subtracting 40 from each observation. Calculations for a two-criteria-month and salesman-analysis of variance are shown below:

Two-way ANOVA Table

Month	Salesman								Row Sum
	A(x_1)	x_1^2	B(x_2)	x_2^2	C(x_3)	x_3^2	D(x_4)	x_4^2	
March	10	100	0	0	8	64	-1	1	17
April	6	36	8	64	10	100	5	25	29
May	-1	1	4	16	0	0	-1	1	2
Column sum	15	137	12	80	18	164	3	27	48

T= Sum of all observations in three samples of months= 48

CF= Correction Factor= $\frac{T^2}{n}$= $\frac{(48)^2}{12}$ = 192

SSC= Sum of squares between salesmen (columns)

= [$\frac{(15)^2}{3}$ + $\frac{(12)^2}{3}$ + $\frac{(18)^2}{3}$ + $\frac{(3)^2}{3}$] – 192

= (75+48+108+3)-192= 42

SSR= Sum of squares between months (rows)

= [$\frac{(17)^2}{4}$ + $\frac{(29)^2}{4}$ + $\frac{(2)^2}{4}$]-192

= (72.25 +210.25+1) -192 = 91.5

SST= Total sum of squares

= ($\sum x_1^2$ +$\sum x_2^2$ +$\sum x_3^2$ +$\sum x_4^2$)-CF

= (137+80+164+27)-192 = 216

SSE= SST-(SSC+SSR) = 216-(42+91.5) = 82.5

The total degrees of freedom are df= n-1=12-1=11.

So df_c= c-1 = 4-1 = 3, df_r = r-1=3-1=2; df =(c-1)(r-1)= 3x2=6

Thus, MSC= SSC/(c-1) = 42/3=14

MSR= SSR/(r-1)= 91.5/2= 45.75

MSE= SSE/(c-1)(r-1) = 82.5/6=13.75

The ANOVA table is shown below:

Source of variation	Sum of squares	Degrees of freedom	Mean Squares	Variance Ratio
Between Salesmen	SSC=42.0	c-1=3	MSC=SSC/(c-1) =14.00	$F_{Treatment}$ = MSC/MSE =14/13.75 =1.018 F_{Block}= MSR/MSE =45.75/13.75 =3.327
Between months	SSR=91.5	r-1=2	MSR=SSR/(r-1) =45.75	
Residual error	SSE=82.5	(c-1)(r-1)=6	MSE=SSE/(c-1)(r-1) =13.75	
Total	SST=216	n-1=11		

(a) The table value of F = 4.75 for df_1 =3, df_2 = 6, and α =5%. Since the calculated value of $F_{Treatment}$ = 1.018 is less than its table value, the null hypothesis is accepted. Hence we conclude that the sales made by the salesmen do not differ significantly.

(b) The table value of F= 5.14 for df_1=2, df_2=6, and α= 5%. Since the calculated value of F_{Block}= 3.327 is less than its table value, the null hypothesis is accepted. Hence we conclude that sales made during different months do not differ significantly.