Analysis of Variance "ANOVA"

Analysis of Variance procedure (called ANOVA) is a way to make multiple comparisons(or is used to compare the means of several populations).

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots = \mu_n$, for n means.

H_a = at least one mean is not equal to the others.

F distribution

ANOVA procedures utilize a distribution called the ***F distribution.*** **A** given F distribution has **two** separate degrees of freedom, represented by $\mathbf{df_1}$ and $\mathbf{df_2}$. The first, df_1, is called the degrees of freedom for ***the numerator*** (samples or groups) and the second, df_2, is called degrees of freedom for ***the denominator*** (of error)**.**

One-way analysis of variance

Suppose that the goal of a research project is to discover whether there are differences in the means of several independent groups. The problem is how we will measure the extent of differences among the means. If we had two groups, we would measure the difference by the distance between

sample means (x _ y) and ***use the two-sample t test***. Here we have more than two groups; we could take all possible pairs of means and do many two-sample t tests. What is the matter with this approach of doing many two-sample t tests, one for each pair of samples? As the number of groups increases, so does the number of tests to perform; for example, we would have to do 45 tests if we have 10 groups to compare. Obviously, the amount of work is greater, but that should not be the critical problem, especially with technological aids such as the use of calculators and computers. So what is the problem? The answer is that performing many tests increases the probability that one or more of the comparisons will result in a type I error (i.e., α significant test result when the null hypothesis is true). This statement should make sense intuitively. For example, suppose that the null hypothesis is true and we perform 100 tests—each has a 0.05 probability of resulting in a type I error; then 5 of these 100 tests would be statistically significant as the result of type I errors. Of course, we usually do not need to do that many tests; however, every time we do more than one, the probability that at least one will result in a type I error exceeds 0.05, indicating a falsely significant difference! What is needed is a different way to summarize the differences between several means and a method of simultaneously comparing these means in one step. This method is called ANOVA or one-way ANOVA, an abbreviation of analysis of variance.

Some necessary definitions and notations

$\mathbf{X_{ij}}$ = observation in group i or sample i .

i = the number of samples.

$\mathbf{n_i}$ = the sample size of sample i .

* **Dot notation** : a dot that replaces an index stands for the mean for the observations the dot replaces.

$\mathbf{X_i .}$ = sum for sample i .

In summation notation , the dot looks like :

$$\mathbf{X_i . = X_{i1} + X_{i2} + \ldots\ldots + X_{ij}}$$

$$\overline{\mathbf{X}}_{\mathbf{i}} \mathbf{.} = \frac{\sum \boldsymbol{Xij}}{\boldsymbol{ni}}$$

The total number of observation is :.

$$n^* = \sum ni$$

X.. = all observations

$$\overline{X}.. = \frac{\sum(ni)(Xi.)}{\sum ni} = \text{mean of all observations.}$$

Sum of squares

* Sum of square total (SST)

SST = $\sum(Xij - X..)^2$

Or

SST = ΣXij^2 – CF

CF = Correction factor

CF = $\frac{(X..)^2}{n*}$

df. $_{\textbf{total}}$**=** n* - 1 or (kr -1)

if k : the number of samples or groups.

r : the number of observations.

* The within sum of squares (SSW)

SSW = $\sum(Xij - Xi.)^2$

= $\sum(ni - 1)Si^2$

df.$_{\textbf{(within samples)}}$= $\sum(ni - 1)$

= n – k or k (r-1)

* The between sum of squares (SSB)

SSB = $\sum ni\ (Xi. - X..)^2$

Or **SSB** = $\frac{\sum Xi.^2}{r}$ – C.F.

df.(between samples)= k-1

Note :. ***(SST = SSB + SSW)***

@ The variance between mean square (MSB)

MSB = $\frac{SSB}{df(between)}$

= $\frac{SSB}{k-1}$

@ The variance within mean square (MSW)

MSW = $\frac{SSW}{df.(within)}$

= $\frac{SSW}{n-k}$ or $\frac{SSW}{k(r-1)}$

$F_{statistic} = \frac{MSB}{MSW}$

Observations (j)	Samples (i)				
	(1)	(2)	.....	(k)	
1	X_{11}	X_{21}	.....	X_{k1}	
2	X_{12}	X_{22}	.....	X_{k2}	
.	.	.	.....	.	
.	.	.	.....	.	
r	X_{1r}	X_{2r}	.....	X_{kr}	
Total	$X_{1.}$	$X_{2.}$	.....	$X_{k.}$	$X_{..}$
Mean	$\overline{X}_{1.}$	$\overline{X}_{2.}$	.....	$\overline{X}_{k.}$	$\overline{X}_{..}$

ANOVA table

Source of Variance S.O.V.	df.	SS	MS	$F_{\cdot statistic}$	F_{table}
Between samples	k-1	SSB	MSB	MSB/MSW	
Within samples	n-k	SSW	MSW		
Total	n-1	SST			

Decisions are made by referring the observed value of the test statistic F to the F table with (k-1 , n-k) degrees of freedom.

In fact, when k = 2 , we have

$$F = t^2$$

Where t is the test statistic for comparing the two population means. In other words, when k = 2 , the F test is equivalent to the two – sided two – sample t test.

Note:. $F_{\alpha\,(df1\,,df2)} = 1 / F_{(1-\alpha)\,(df2\,,\,df1)}$

Or :. $F_{(1-\alpha)\,(df2\,,\,df1)} = 1 / F_{\alpha\,(df1\,,df2)}$

Example:.

A study was conducted to test the question as to whether cigarette smoking is associated with reduced serum-testosterone levels in men aged 35 to 45. The study involved the following four groups :.

1- Nonsmokers who had never smoked.
2- Former smokers who had quit for at least six months prior to the study.
3- Light smokers, defined as those who smoked 10 or fewer cigarettes per day.
4- Heavy smokers, defined as those who smoked 30 or more cigarettes per day.

Each group consisted of 10 men and table below shows raw data, where serum-testosterone levels were measured in μg/dL.

Serum-testosterone levels measure in μg/dL.

Nonsmokers	Former smokers	Light smokers	Heavy smokers
0.44	0.46	0.37	0.44
0.44	0.50	0.42	0.25
0.43	0.51	0.43	0.40
0.56	0.58	0.48	0.27
0.85	0.85	0.76	0.34
0.68	0.72	0.60	0.62
0.96	0.93	0.82	0.47
0.72	0.86	0.72	0.70
0.92	0.76	0.60	0.60
0.87	0.65	0.51	0.54

Answer:.

$X_{..} = 24.03$

$\overline{X}_{..} = \frac{24.03}{40} = 0.60075$

$X_{1.} = 0.44 + 0.44 + \ldots\ldots\ldots + 0.87 = 6.87$

$\overline{X}_{1.} = \frac{6.87}{10} = 0.687$

$X_{2.} = 0.46 + 0.50 + \ldots\ldots\ldots + 0.65 = 6.82$

$\overline{X}_{2.} = \frac{6.82}{10} = 0.682$

$X_{3.} = 0.37 + 0.42 + \ldots\ldots\ldots + 0.51 = 5.71$

$\overline{X}_{3.} = \frac{5.71}{10} = 0.571$

$X_{4.} = 0.44 + 0.25 + \ldots\ldots\ldots + 0.54 = 4.63$

$\overline{X}_{4.} = \frac{4.63}{10} = 0.463$

$$\mathbf{CF} = \frac{(X..)^2}{n*}$$

$$= \frac{(24.03)^2}{4 \times 10} = 14.43602$$

$$\mathbf{SST} = \Sigma Xij^2 - \mathrm{CF}$$

$$= (0.44)^2 + \ldots\ldots + (0.54)^2 - 14.43602$$

$$= 15.8461 - 14.43602$$

$$= 1.41008$$

$$\mathbf{SSB} = \frac{\Sigma Xi.^2}{r} - \mathrm{C.F.}$$

$$= [(6.87)^2 + (6.82)^2 + (5.71)^2 + (4.63)^2 / 10\ ] - 14.43602$$

= 14.77503 – 14.43602

= 0.33901

SSW = SST – SSB

= 1.41008 – 0.33901

= 1.07107

S.O.V.	df.	SS	MS	$F_{statistic}$	F_{table} 0.01	F_{table} 0.05
Between Samples	4-1=**3**	0.33901	0.113	3.8047	4.39	2.872*
Within Samples	40-4=**36**	1.07107	0.0297			
Total	40-1=**39**	1.41008				

The resulting F test indicates that the overall differences between the four population means is statistically significant at $\alpha = 0.05$ level but not at $\alpha = 0.01$ level.

Example:.

A study was conducted to investigate the risk factors for peripheral arterial disease among persons 55 to 74 years of age. The bellowing table provides data on LDL cholesterol levels (mmol/L) from four different subgroups of subjects. Test to compare the three groups simultaneously. Name your test and state clearly your null and alternative hypotheses and choice of test size.

Groups	n	$\overline{Xi.}$	S
1- Patients with intermittent claudication.	73	6.22	1.62
2- Major asymptotic disease cases.	105	5.81	1.43
3- Minor asymptotic disease cases.	240	5.77	1.24
4- Those with no disease.	1080	5.47	1.31

Answer :.

$$n^* = \sum ni$$

$$= 1498$$

$$\overline{X}.. = \frac{\sum(ni)(Xi.)}{n*}$$

$$= \frac{(73\times6.22)+(105\times5.81)+(240\times5.77)+(1080\times5.47)}{1498}$$

$$= 5.578$$

SSB $= \sum ni(\overline{Xi}. - \overline{X}..)^2$

$= 73(6.22 - 5.578)^2 + 105(5.81 - 5.578)^2 + 240(5.77 - 5.578)^2 + 1080(5.47 - 5.578)^2$

$= 57.184$

SSW $= \sum(ni - 1)Si^2$ ***When[test the three groups simultaneously]***

$= (73 -1)(1.62)^2 +(105 -1)(1.43)^2 +(240 -1)(1.24)^2$

$= 769.113$

SST = SSB + SSW

= 57.184 + 769.113

= 826.297

ANOVA table

S.O.V.	df.	SS	MS	$F_{statistic}$	$F_{tab.\ 0.01}$	$F_{tab.0.05}$
Between Samples	3	57.184	19.061	37.026	3.78	2.60
Within Samples	1494	769.113	0.515			
Total	1497	826.297				

The resulting F test indicates that the overall differences between the three groups simultaneously means is statistically significant at $\alpha = 0.01$ and 0.05 level.