-4-
Analysis of variance

4.1 Introduction

Consider a population consisting of many items that can be classified in two ways. For example, let a factory have three production lines for the same item (Light insulating boards). Suppose that samples were chosen from each line and tested for their density (kg/m^3). The results are shown in the following table

Line A	Line B	Line C
1052	1065	1061
1037	1054	1057
1062	1043	1049
1055	1050	1050
1061	1048	1039
1048	1066	1044
1070	1052	1036

An analysis of variance answers the following question: How far is the following null hypothesis correct H_o: The quality of boards does not depend on the production line. In other words, the mean values of densities from each line are equal: $\mu_A = \mu_B = \mu_C$.

To this aim, two estimates for the variance are calculated: $\hat{s}_1^2$ and $\hat{s}_2^2$. They are then compared by dividing the higher value on the lower value. This ratio, known as the *F* – ratio, is then compared to some critical value. This will be detailed in the following section.

4.2 One - way ANOVA

4.2.1 Estimates of variances

The previous example is one of a one – way ANOVA (Analysis Of Variance). A first estimate of variance ($\hat{s}_1^2$) is obtained by considering the mean value of columns: $\bar{x}_A$, $\bar{x}_B$ and $\bar{x}_C$.

The standard deviation of universe is related to that of the means by

$$\sigma = \sigma_{\bar{x}}.\sqrt{N}$$

Squaring we get $\sigma^2 = \sigma_{\bar{x}}^2.N$

Using the estimate $\hat{s}_1^2$, an estimate for the variance of the universe is:

$$\sigma_1^2 = \hat{s}_1^2.n \qquad (4.1)$$

This represents the variance of the universe based on an estimated variance of column means.

The different computational steps are best understood by referring to the following table.

The variance of column means $\hat{s}_1^2 =$ 14.33

Hence the estimated variance of universe = 7×14.33 = **100.33**

Another estimate $\hat{s}_2^2$ of universe variance (σ_2^2) is obtained by considering the mean of columns variance.

The column variances are obtained from the function VAR.S and equal respectively 115.3, 73.67 and 82.67 with mean $\hat{s}_2^2$ = 90.56

The ratio between these two variances is used to test the hypothesis enunciated at the beginning of this section. It is called the *F* – ratio.

The *F* – ratio is therefore 100.33 / 90.56 = **1.108**

4.2.2 The *F* - distribution

This distribution was first proposed by Fisher. It is a continuous distribution similar in shape to the t – distribution, except that its density function depends on two parameters rather than one: *n* and *d*, which are the number of degrees of freedom of numerator and denominator of the *F* value respectively.

To compare the calculated value to tabulated values, one has to include two entries: the number of degrees of freedom of numerator and that of denominator. If the calculated value of *F* is lower than some critical value (at a specified significance level) then the hypothesis is not rejected.

In that case, the number of degrees of freedom is given separately for both estimated variances: For the numerator it is given by:

$(d.f.)_n = c - 1$ (4.2)

For the denominator, it is:

$(d.f.)_d = c.(r - 1)$ (4.3)

Where, *c* is the number of columns and *r* the number of rows.

In the present case, $(d.f.)_n = 3 - 1 = 2$

And, $(d.f.)_d = 3.(7 - 1) = 18$

The value of *F* corresponding to the two aforementioned degrees of freedom at any significance level can be readily obtained from the function: FINV available on the statistical list in EXCEL: $\boldsymbol{F_{crit.}}$ **= 3.55** at $\alpha = 0.05$.

Since the obtained value is less than the critical value, then the hypothesis is accepted and the differences between means are only due to chance.

A further assessment regarding the reliability of the hypothesis can be obtained by calculating the ***p* – value**. This represents the maximum significance level that would not reject the null hypothesis. In general, the higher the *p* – value the more reliable is the hypothesis.

This can be obtained using the goal – seek module by equating the critical value of *F* to the calculated value.

4.2.3 One way ANOVA tables

Using the DATA ANAYSIS module available on EXCEL ADDINS, it is possible to obtain directly the ANOVA table including a lot of statistical data that were not cited above. The values of *F* and F_{crit} are compared to decide about the acceptance or rejection of the

hypothesis. The following table displays the ANOVA table obtained for the previous example.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	7	7385	1055	115.333
Column 2	7	7378	1054	73.6667
Column 3	7	7336	1048	82.6667

ANOVA

Source of	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	200.667	2	100.333	**1.10798**	0.35173	**3.55456**
Within Groups	1630	18	90.5556			
Total	1830.67	20				

Example 4.1

A company possesses four production lines that deliver their waste water to a treatment plant. The BOD figures ($mg.L^{-1}$) for the four streams were monitored on monthly basis during one year to obtain the following table.

A	**B**	**C**	**D**
325	415	356	378
366	410	389	388
314	388	377	420
355	402	402	412
327	400	400	425
402	395	357	377
389	388	366	380
388	420	388	375
410	390	404	420
390	375	390	380
375	415	370	375
395	400	345	415

Check the hypothesis: All four effluents have comparable mean BOD levels.

Solution:

The ANOVA table obtained using the EXCEL module is shown below and reveals that the hypothesis cannot be accepted. In order not to reject this hypothesis, one has to use a significance level of 0.00648 corresponding to $L = 0.9935$

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	12	4436	369.67	1055.33
Column 2	12	4798	399.83	181.061
Column 3	12	4544	378.67	392.606
Column 4	12	4745	395.42	431.720

ANOVA

Source of	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7206.563	3	2402.19	**4.6628**	0.00648	**2.816**
Within Groups	22667.9	44	515.18			
Total	298711.48	47				

4.3 Two way ANOVA

In many instances the problem is complicated by the presence of more than one variable. Consider for example the mean marks obtained by male and female students in three different exams. Two factors are in play: type of exam and gender. The following table shows the scores.

	A	**B**	**C**
M	52	62	58
F	53	66	60

The problem at hand is to decide whether there are any differences between the mean marks in each exam and the mean mark of different genders.

This is a typical case of a two way ANOVA without replication. Replication will arise if instead of the mean marks we would have recorded individual marks of a sample of students.

Usually, tedious calculations can be avoided as the ANOVA table can be readily obtained from the DATA ANALYSIS module using the ANOVA – two factors function. If no data are replicated, then one uses the option (without replication).

Example 4.2

A study was carried out to follow up the effect of exposure time on the mean mass of lead in liver (mg) at four lead casting units. The results were compiled in the following table.

Exposure, days	**A**	**B**	**C**	**D**
28	0.875	0.788	0.900	0.778
56	1.155	0.989	1.095	0.987
84	1.236	1.254	1.222	1.205

Test the hypotheses concerning the dependence of the amount of lead stored in liver on the time of exposure and the nature of casting unit.

Solution:

The two ways ANOVA table is obtained:

Anova: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Row 1	4	3.341	0.83525	0.00376
Row 2	4	4.226	1.0565	0.006857
Row 3	4	4.917	1.2292	0.00044
Column 1	3	3.266	1.08867	0.03588
Column 2	3	3.031	1.01033	0.05463
Column 3	3	3.217	1.0723	0.02631
Column 4	3	2.97	0.99	0.04559

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	0.31204	2	0.15602	**73.295**	0.00006	**5.413**
Columns	0.02038	3	0.00679	**3.1915**	0.105	**4.575**
Error	0.01012	6	0.00213			
Total	0.34519	11				

This table shows that for rows (Exposure duration), $F_{crit} > F$ and that in order to accept the hypothesis of equal means, one has to use a significance level = 0.00006 $(L = 0.99994)$. This means that the level of poisoning is definitely dependent on the duration of exposure. On the other hand, for columns (Nature of unit), $F_{crit} < F$ and the hypothesis of independence can be accepted.

4.4 Statistical interpretation of the p – value

As expressed in the above problem, the p – value indicates the maximum significance level that can be used for the independence hypothesis not to be rejected. This value can also be viewed in another way:

- If a hypothesis is accepted, then the farther are the values of F and F_{crit}, the higher will be the p – value. In the above example, this value was 0.105 for the accepted hypothesis that the concentration of lead is independent of the casting unit. This means that if the test was repeated, say, one hundred times, then, even if it was proved than in 10 times of these, the level of lead was effectively dependent on the casting unit, then this will simply be due to chance.

- On the other hand, if a hypothesis is rejected, then the p – value simply represents the probability that this conclusion is wrong, that is, the hypothesis should have been accepted.

4.5 Exercise problems

In all forthcoming problems it will be assumed that the significance level = 0.05

(1) The following data represent the daily number of units turned out by 5 workmen using 4 different types of machines. Test the following hypothesis: the mean productivity is the same for the 4 different machine types.

	Machine type			
Workman	M_1	M_2	M_3	M_4
1	44	38	47	36
2	46	40	52	43
3	34	36	44	32
4	43	38	46	33
5	38	42	49	39

(2) The following data give the percent yield of a chemical product that resulted from using 4 different catalysts for the process. Test whether the mean yield depends on the type of catalyst.

A	**B**	**C**	**D**
36	35	34	34
33	37	39	31
35	36	37	35
34	35	38	32
32	37	39	34

(3) The following table relates to the mean marks of male and female students in 5 exams. Test the hypothesis that the mean mark does not depend on gender.

Exam	**Male**	**Female**
A_1	36	42
A_2	33	37
A_3	35	35
A_4	34	33
A_5	32	37

(4) In a study performed to investigate the effect of age category and gender on the annual salaries of employees in a large firm, the following data was obtained for mean salaries.

	20 – 30	30 - 40	40 - 50	50 – 60	>60
Male	58000	78000	125000	164000	105000
Female	54000	74000	118000	160000	84000

Perform a two way ANOVA analysis to test the hypotheses: the mean salary is independent of both age and gender.

(5) The following data were obtained from a survey performed by graduating students (Fall 2011) to test the presence of any relation between the mean marks of students (expressed as percentages) in different departments (labeled here A, B, C, D ,E) and the academic level (1^{st}, 2^{nd}, etc…).

Perform a two way ANOVA analysis to test the presence of any dependence of mean marks on both factors.

	A	B	C	D	E
Year 1	58	62	60	54	68
Year 2	60	63	62	58	65
Year 3	62	63	67	61	64
Year 4	70	68	68	66	71