Chapter 6

Analysis of variance and experimental designed

6.1 Introduction

In this chapter, we are interested in comparing several means which are more than two means. As before , we can use the T-test but we have to compare every two means which gives us a large number of tests(if we have 6 means, we need 10 tests) and also, we cannot control the error level overall the individual tests which lead to false conclusions overall that decreasing the accuracy of the test.

Thus, we will use a statistical technique called analysis of variance **(ANOVA)** which take place in designed or planed experiments in which the levels of one or more variables**(called factors)** are different in chosen way or design. Then the response of another variable to these changes is measured. Our aim is to compare the effect of the levels of the factor(factors) on the **response variable** , which we will denote by Y. This response variable Y will be separated into different sources, some of these sources can be controlled **(the factors)** and the other uncontrolled sources **(the error)** . For example, the growth of a plant changes if there are changes in the soil moisture, the temperature, the nitrogen in the soil , etc.

We will assume in this chapter that the levels of all factors are chosen or fixed by the experimenter. While it is possible to have a random sample of levels of factors(in that case the factor is said to be random), such cases is not considered in this chapter. The tests that will be discussed depend on the assumption that the factor levels are fixed.

6.2 One Way ANOVA (one factor study)

Assume we have one factor(qualitative or quantitative) which we choose certain values of interest. This factor is called treatment because such studies happened when different units in an experiment were given different treatments. For example, if we are interested in the weight of a particular type of animal(response variable) , and we want to know the effect of different types of animals foods (factor) on this weight. The factor (animal foods) is called a treatment since we give (or treat) different animals with different animal foods.

In a one factor study, we want to compare the means of the response variable Y for each level of the treatment (factor). If there exist only 2 levels, we can use the t- statistic used in chapters 3 and 4 to compare two means. In this section, we are dealing with factors containing 3 or more levels, so, we will use the technique of the analysis of variance for one factor. The data in one factor study (contains a treatments), each treatment i, we measure the response variable Y for n_i different units giving data as follows:

	Treatments					
	1	2	3	…	a	
	y_{11}	y_{21}	y_{31}		y_{a1}	
	y_{12}	y_{22}	y_{32}		y_{a2}	
	⋮	⋮	⋮		⋮	
	y_{1n_1}	⋮	⋮		y_{an_k}	
		y_{2n_2}	⋮			
			y_{3n_3}			
Total	$T_{1.}$	$T_{2.}$	$T_{3.}$	…	$T_{a.}$	$T_{..}$

Now, we want to test whether the population treatment means are all equal or, i.e., that the effects of all treatments are the same. Thus the hypothesis can be written as:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ (all treatments have the same effect)

H_1: $not\ all\ \ \mu_i\ are\ equal$(some treatments have different effects)

To test these hypothesis, we need to use the analysis of variance which depend on dividing the **total sum of squares** (SST) into two parts, one part due to treatments and the other due to error since the sources of variation divided into treatments and error. This will be as follows:

SST=Total sum of squares= $\sum\sum y_{ij}^2 - \frac{(T_{..})^2}{N}$

SSt_r= Treatment sum of squares=$\sum \frac{T_{i.}^2}{n_{i.}} - \frac{(T_{..})^2}{N}$

SSE=Error sum of squares= SST- SSt_r

Where

$N = n_1 + n_2 + \cdots + n_a,$

$T_{..} = \sum\sum y_{ij} = \sum_i T_{i.}$, and

$$T_{i.} = \sum_j y_j$$

Also, we have the following assumptions:

1- Independent random samples from the (a) populations.
2- Treatments populations are normally distributed and equal variances.

6.2.1 The Steps of The One Way ANOVA Test

1- Data N, α, a

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ (all treatments have the same effect)

H_1: $not\ all\ \ \mu_i\ are\ equal$ (some treatments have different effects)

3-The test statistic:

$$F = \frac{MST_r}{MSE}$$

Where $MST_r = \frac{SSt_r}{a-1}$ and $MSE = \frac{SSE}{N-a}$

4-The table value:

$$F_{1-\alpha,\ a-1,\ N-a}$$

5-the decision:

We reject H_0 and accept H_1 , if $F > F_{1-\alpha,\ a-1,\ N-a}$

i.e., there is a significant difference effects of treatments.

Also, we can arrange the values above into a table called ANOVA table (as given in computer results) as follows:

ANOVA Table

Source	df	SS	MS	F
Factor	a-1	SSt_r	$MST_r = \frac{SSt_r}{a-1}$	$F = \frac{MST_r}{MSE}$
Error	N-a	SSE	$MSE = \frac{SSE}{N-a}$	
Total	N-1	SST		

<u>EX(1)</u>

In department of animal production, they are interested in discovering the effect of three enzymes A,B,C for increasing daily milk of a specified type of cows. 18 cows randomly Chosen with the same circumstance and each 6 cows of them given one of the three enzymes, then the increasing in milk is measured with liters as follows:

A	B	C
16	9	14
17	13	19
11	12	13
15	11	11
18	15	13
19	12	14

Is there exist a significant difference between the three enzymes in increasing the average of daily milk at $\alpha = 0.05$.

solution

1- Data $N = 18, \alpha = 0.05, a = 3$

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3$ (three enzymes have the same effect)

$H_1: at\ least\ one\ mean\ is\ different$ (some of 3 enzymes have different effects)

	A	B	C	
	16	9	14	
	17	13	19	
	11	12	13	
	15	11	11	
	18	15	13	
	19	12	14	
Total	$T_{1.}$=96	$T_{2.}$=72	$T_{3.}$=84	$T_{..}$=252

$\sum\sum y_{ij}^2 = (16)^2 + (17)^2 + (11)^2 + \cdots + (14)^2 = 3672$

SST= $\sum\sum y_{ij}^2 - \frac{(T_{..})^2}{N} = 3672 - \frac{(252)^2}{18} = 144$

SSt_r= $\sum \frac{T_{i.}^2}{n_{i.}} - \frac{(T_{..})^2}{N} = \left[\frac{(96)^2}{6} + \frac{(72)^2}{6} + \frac{(84)^2}{6}\right] - \frac{(252)^2}{18} = 48$

SSE= SST- SSt_r=144-48=96

ANOVA Table

Source	df	SS	MS	F
Factor	2	48	24	3.75
Error	15	96	6.4	
Total	17	144		

3-The test statistic:

$$F=\frac{MST_r}{MSE}=3.75$$

4-The table value:

$$F_{1-\alpha,\ a-1,\ N-a}=F_{0.95,2,15}=3.68$$

5-the decision:

We reject H_0 and accept H_1 , if $F=3.75>3.68=F_{0.95,2,15}$

i.e., there is a significant difference between the 3 enzymes in increasing the average of daily milk.(some of 3 enzymes have different effects)

EX(2)

In study on the effect of Nitrogen fertilization on cereal crops, plots of a particular variety of wheat were randomly given fertilizer at one of four rates: 0, 50, 100, 150. At a certain date, plants were randomly selected from the plots and the plants height (in cm) was measured and recorded as then we obtain the following ANOVA table:

Source	Df	SS	MS	F
Treatments		329.842	109.82733	
Error		44.6		
Total	16	374.082		

a) Complete the table and then find:
b) What is the variable?

c) What are the treatments? How many?

d) Can we conclude that all 4 rates of fertilizers have the same (equal) effects on the average plant height at $= 0.1$?

Solution

a) Complete the table and then find:

Source	Df	SS	MS	F
Treatments	3	329.842	109.82733	32.012
Error	13	44.6	3.43077	
Total	16	374.082		

b) the variable is: the plant height

c) the treatments are : the fertilizers rates , there exist 4 rates.

d) 1- Data $N = 17, \alpha = 0.1, a = 4$

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (4 rates have the same effect)

$H_1: at\ least\ one\ mean\ is\ different$ (some of the 4 rates have different effects)

3-The test statistic: $F = \frac{MST_r}{MSE} = 32.012$

4-The table value:

$$F_{1-\alpha,\ a-1,\ N-a} = F_{0.9,3,13} = 2.56$$

5-the decision:

We reject H_0 and acceptH_1 , if $F = 32.012 > 2.56 = F_{0.95,2,15}$

i.e., there is a significant difference between the 4 rates in the average of plant height.(some of 4 rates have different effects on the average of plant height)

6.3 Two Way ANOVA (two factor studied without interaction)

Assume we have two factor effect on a response variable Y. these two factors with no interactions which means that the effect of a particular level of one factor does not depend on what level of the other factor is used. The first factor A has levels (a) and the second factor contains the levels (b). For example, if we are studying the affect of the training methods and the IQ level on the scientific understanding level.

In this study, we can have two cases:

1) If we are interested in one factor A (called treatments), and not interested in the factor B(called the blocks). Thus, We use the blocks to improve the accuracy of the experiment by making sure that any differences found in treatments are not really due to differences in the blocks. So we have only one test for treatments.
2) If we are interested in the two the factors with no interactions, the only difference is that we have the effects of the both factors to test. So, we can make test for factor A and another test for factor B.

In the previous two cases, if we Suppose that there are (a) treatments and (b) blocks, then the data appears as follows:

	Factor A (Treatments)					total
Factor B (Blocks)	1	2	3	…	a	
1	y_{11}	y_{21}	y_{31}		y_{a1}	$T_{.1}$
2	y_{12}	y_{22}	y_{32}		y_{a2}	$T_{.2}$
⋮	⋮	⋮	⋮		⋮	
b	y_{1b}	y_{2b}	y_{3b}	…	y_{ab}	$T_{.b}$
total	$T_{1.}$	$T_{2.}$	$T_{3.}$	…	$T_{a.}$	$T_{..}$

Now, we will study the second case when we are interested in the 2 factors. we have two tests:

i)Test whether the means of the factor A are all equal or, (i.e., the effects of all levels factor A are the same). Thus the hypothesis can be written as:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ (all levels of factor A have the same effect)

$H_1: not\ all\ \ \mu_i\ are\ equal$(some levels of factor A have different effects)

ii) Test whether the means of the factor B are all equal or, (i.e., the effects of all levels of factor B are the same). Thus the hypothesis can be written as:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_b$ (all levels of factor B have the same effect)

$H_1: not\ all\ \ \mu_i\ are\ equal$(some levels of factor B have different effects)

To test the previous hypothesis, we need to use the analysis of variance which depend on dividing the **total sum of squares** (SST) into three parts, the first part due to factor A(SSA) ,the second part due factor B(SSB) ,and the last part due to error(SSE), since the sources of variation divided into factor A, factor B and the error. This will be as follows:

SST=Total sum of squares= $\sum\sum y_{ij}^2 - \frac{(T_{..})^2}{N}$

SSA= sum of squares of factor A= $\sum \frac{T_{i.}^2}{b} - \frac{(T_{..})^2}{N}$

SSB= sum of squares of factor B= $\sum \frac{T_{.j}^2}{a} - \frac{(T_{..})^2}{N}$

SSE=Error sum of squares= SST- SSA-SSB

Where $N = a * b$

$T_{..} = \sum\sum y_{ij} = \sum_i T_{i.}$,

$T_{.j} = \sum_i y_{ij}$ and $T_{i.} = \sum_j y_{ij}$

Also, we have the following assumptions:

1- Independent random samples from the (a b) populations.
2- These populations are normally distributed and equal variances.
3- There is no interaction between the two factors.

6.3.1 The Steps Of the test of factor A

1- Data $N, \alpha,$ a

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$ (all levels of factor A have the same effect)

$H_1:$ $not\ all\ \ \mu_i\ are\ equal$ (some levels of factor A have different effects)

3-The test statistic:

$$F_1 = \frac{MSA}{MSE}$$

4-The table value:

$$F_{1-\alpha,\ a-1,\ (a-1)(b-1)}$$

5-the decision:

We reject H_0 and acceptH_1 , if $F_1 > F_{1-\alpha,\ a-1,\ (a-1)(b-1)}$

i.e., there is a significant difference in the levels of factor A.

Also, we can arrange the values above into a table called ANOVA table (as given in computer results) as follows:

6.3.2 The Steps Of the test of factor B

1- Data $N, \alpha,$ a

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_b$ (all levels of factor B have the same effect)

$H_1:$ $not\ all\ \ \mu_i\ are\ equal$ (some levels of factor B have different effects)

3-The test statistic:

$$F_2 = \frac{MSB}{MSE}$$

4-The table value:

$$F_{1-\alpha\,,\ b-1,\ (a-1)(b-1)}$$

5-the decision:

We reject H_0 and accept H_1 , if $F_2 > F_{1-\alpha\,,\ b-1,\ (a-1)(b-1)}$
i.e., there is a significant difference in the levels of factor B.

we can summarize the previous two tests in the two-way ANOVA table as follows:

ANOVA Table

Source	df	SS	MS	F
Factor A	a-1	SSA	$MSA = \frac{SSA}{a-1}$	$F_1 = \frac{MSA}{MSE}$
Factor B	b-1	SSB	$MSB = \frac{SSB}{a-1}$	$F_2 = \frac{MSB}{MSE}$
Error	(a-1)(b-1)	SSE	$MS = \frac{SSE}{(a-1)(b-1)}$	
Total	ab-1	SST		

EX(3)

If we have 4 types of fertilizers denoted A, B,C,D, and 4 variety of wheat seeds numbered 1,2,3,4. After harvest, the wheat was ground into flour, and the increasing in the production was measured:

		Fertilizers type			
		A	B	C	D
Variety of wheat seeds	1	9.3	9.4	9.2	9.7
	2	9.4	9.3	9.4	9.6
	3	9.6	9.8	9.5	10
	4	10	9.9	9.7	10.2

a) Test whether 4 types of fertilizers have different effects on the average of increasing production of wheat.
b) Test whether the four varieties of wheat seeds have the same average of increasing production of wheat.

Use $\alpha = 0.05$ and assume that no interaction among factors.

Solution

		Fertilizers type				**Total**
Variety of wheat seeds		A	B	C	D	
	1	9.3	9.4	9.2	9.7	**37.6**
	2	9.4	9.3	9.4	9.6	**37.7**
	3	9.6	9.8	9.5	10	**38.9**
	4	10	9.9	9.7	10.2	**39.8**
Total		**38.3**	**38.4**	**37.8**	**39.5**	**154**

$$SST = 1483.5 - \frac{(154)^2}{16} = 1.25$$

$$SSA = \frac{1}{4}[(38.3)^2 + (38.4)^2 + (37.8)^2 + (39.5)^2] - \frac{(154)^2}{16} = 0.385$$

$$SSB = \frac{1}{4}[(37.6)^2 + (37.7)^2 + (38.9)^2 + (39.8)^2] - \frac{(154)^2}{16} = 0.825$$

SSE=SST-SSA-SSB=1.25-0.385-0.825=0.04

ANOVA Table

Source	df	SS	MS	F
Factor A	3	0.385	0.1283	$F_1 = 28.896$
Factor B	3	0.825	0.2750	$F_2 = 62.5$
Error	9	0.04	0.0044	
Total	15	1.25		

a) Test whether 4 types of fertilizers have different effects on the average of increasing production of wheat.

1- Data $N = 16, \alpha = 0.05, a = 4$

2-The hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (all fertilizers have the same effect)

$H_1: not\ all\ \ \mu_i\ are\ equal$ (some fertilizers have different effects)

3-The test statistic:

$$F_1 = \frac{MSA}{MSE} = 28.896$$

4-The table value:

$$F_{1-\alpha\ ,\ \ a-1,\ \ (a-1)(b-1)} = F_{0.95\ ,3,9} = 3.86$$

5-the decision:

We reject H_0 and acceptH_1 , if $F_1 = 28.896 > 3.86 = F_{0.95\ ,3,9}$
i.e., there is difference in the types of fertilizer in the effect on the average of increasing production of wheat.

b) Test whether the four varieties of wheat seeds have the same average of increasing production of wheat.

1- Data $N = 16, \alpha = 0.05, b = 4$

2- The hypothesis:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (all varieties of wheat seeds have the same effect)

$H_1: not\ all\ \ \mu_i\ are\ equal$ (some wheat seeds have different effects)

3-The test statistic:

$$F_2 = \frac{MSB}{MSE} = 62.5$$

4-The table value:

$$F_{0.95\ ,3,9} = 3.86$$

5-the decision:

We reject H_0 and acceptH_1 , if $F_2 = 62.5 > 3.86 = F_{0.95,3,9}$
i.e., there is a significant difference in the average of increasing the wheat Production for the 4 varieties of wheat seeds.