الجمهورية العربية المتحدة

معهد التخطيط القومي

نسخة واحدة فقط

Memo. No. 607

Analysis of Variance

By

Dr. Abd El Aziem Anis

December 1965

Operations Research Center.

القاهرة

٣ شارع محمد مظهر - بالزمالك

Analysis of Variance

I) Def.: The analysis of variance is a statistical technique by which the total variation of the variable being studied can be separated into components that are of experimental interest or importance.

example: as an illustration of the type of problem for which the analysis of variance is useful, consider a gunnary problem experiment in which 4 different brands of shells are to be tested to see whether they are equally satisfactory in quality.

The experiment consists of having 6 different marksmen fire on equal number of rounds with each brand of shell and recording the scores made by each marksman for each brand.

Then scores (the variable) may be arranged in a rectangular array containing 6 rows and 4 columns.

For the purpose of considering other problems, let the scores be displayed in a rectangular array containing a rows and b columns as shown in table.

Table I

x_{11}	x_{12}	.....	x_{1j}	..... x_{1b}	$\bar{x}_{1.}$
x_{21}	x_{22}	.....	x_{2j}	..... x_{2b}	$\bar{x}_{2.}$
.....					
.....					
x_{i1}	x_{i2}	.....	x_{ij}	..... x_{ib}	$\bar{x}_{i.}$
x_{a1}	x_{a2}	.....	x_{aj}	..... x_{ab}	$\bar{x}_{a.}$
$\bar{x}_{.1}$	$\bar{x}_{.2}$		$\bar{x}_{.j}$	$\bar{x}_{.b}$	$\bar{x}_{..}$

The location of the dot in the index show whether the mean is a row mean or a column mean.

i.e. $$\bar{x}_{i\cdot} = \sum_{r=1}^{b} x_{ir}/b$$

$$\bar{x}_{\cdot j} = \sum_{r=1}^{a} x_{rj}/a$$

$$\bar{x}_{\cdot\cdot} = \frac{1}{a\,b} \sum_{j=1}^{b} \sum_{i=1}^{a} x_{ij}$$

Hence it is obvious that

$$\bar{x}_{\cdot\cdot} = \frac{1}{a} \sum_{i=1}^{a} \bar{x}_{i\cdot}$$

$$= \frac{1}{b} \sum_{j=1}^{b} \bar{x}_{\cdot j}$$

II) Two well known mathematical models are available for application in experiments of this type:

(1) the "Linear hypothosis" model

(2) the "components of variance" model.

the essential difference between the two models lies in the assumptions made.

Notice: that x_{ij} is regarded as a set of ab random variables for which the observed values are the values resulting from a single random experiment.

The linear hypothesis model: First assumption: (α)

(α) This model assumes that the random variable x_{ij} has a mean μ_{ij} which can be written in the form

$$\mu_{ij} = a_i + b_j + c$$

where

$$c = E(\bar{x}) ,$$
$$a_i = E(x_{i\cdot} - \bar{x}_{\cdot\cdot})$$
$$b_j = E(x_{\cdot j} - \bar{x}_{\cdot\cdot}) \qquad (1)$$

but

$$\sum_{i=1}^{a} (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) = o , \quad \sum_{j=1}^{b} (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = o$$

Hence

$$\sum_{i=1}^{a} a_i = o , \quad \sum_{j=1}^{b} b_j = o \qquad (2)$$

(1.1) Assumption (1) merely states that the mean of the variable x_{ij} is the sum of a general mean c, a row effect a_i, a column effect b_j.

Application in case of gunnary experiment: This means that if the i-th markman was superior, his mean score would be expected to exceed the mean more for all six marksmen by a + ve quantity a_i whereas if he were an inferior marksman, a_i would be - ve.

Also b_j is a number (+ ve or - ve) which measures the superiority or of the brand j with respect to all brands.

Important criticism: The additive feature of (1) is restrictive. For example, if the rows of table (1) correspond to different amounts of a chemical compound added to the soil, whereas the columns correspond to different quantities of a second chemical compound added, one would not expect the effects of those compounds on crop production to operate independently in this manner.

Second assumption β:

In addition to (1), this model assumes that the variables x_{ij} are independently and normally distributed with the same variances σ^2. Now if we want to test the hypothesis that the

brands of shells are equally good, this would be expressed in the form that

$$H_o : b_1 = b_2 = b_3 = \ldots\ldots b_b$$

In view of (2), this mean

$$H_o : b_j = o \qquad (j=1, 2, \ldots b) \tag{3}$$

(8) Proceedings of the analysis of variance

Under the foregoing assumptions, we proceed to prove that

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\bar{x})^2 = \sum_i\sum_j(\bar{x}_{i\circ}-\bar{x})^2 + \sum_i\sum_j(\bar{x}_{\circ j}-\bar{x})^2 + \sum_i\sum_j(x_{ij}-\bar{x}_{i\circ}-\bar{x}_{j\circ}+\bar{x})^2 \tag{4}$$

The importance of this formulae lies in the fact that its shows that the total variation of the variable x_{ij} could be broken down into three components:

(1) the first component measuring the variation of row means (ie variation in the marksmen ability).

(2) the second component measuring the variation of column means (ie the variation in the shell brand's effect)

(3) the third component measuring the variation in the variables x_{ij} after the row and column effects have be eliminated.

Proof of (4):

$$(x_{ij}-\bar{x})^2 = \left[(\bar{x}_{i\circ}-\bar{x}) + (\bar{x}_{\circ j}-\bar{x}) + (x_{ij}-\bar{x}_{i\circ}-\bar{x}_{\circ j}+\bar{x})\right]^2$$

$$= (\bar{x}_{i\circ}-\bar{x})^2 + (\bar{x}_{\circ j}-\bar{x})^2 + (x_{ij}-\bar{x}_{i\circ}-\bar{x}_{\circ j}+\bar{x})^2$$

$$+ 2(\bar{x}_{i\circ}-\bar{x})(\bar{x}_{\circ j}-\bar{x}) + 2(\bar{x}_{i\circ}-\bar{x})(x_{ij}-\bar{x}_{i\circ}-\bar{x}_{\circ j}+\bar{x}) + 2(\quad)(\quad)$$

Hence $$\sum_i\sum_j(x_{ij}-\bar{x})^2 = \sum\sum(\bar{x}_{i\circ}-\bar{x})^2 + \sum\sum(\bar{x}_{\circ j}-\bar{x})^2 + \sum\sum(x_{ij}-\bar{x}_{i\circ}-\bar{x}_{\circ j}+\bar{x})^2$$

$$+ 2\sum\sum(\bar{x}_{i\circ}-\bar{x})(\bar{x}_{\circ j}-\bar{x}) + 2\sum\sum(\quad)(\quad) + 2\sum\sum(\quad)(\quad)$$

Now

$$\sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{x}_{i.}-\bar{\bar{x}})(\bar{x}_{.j}-\bar{\bar{x}}) = \sum_{i=1}^{a} (\bar{x}_{i.}-\bar{\bar{x}}) \sum_{j=1}^{b} (\bar{x}_{.j}-\bar{\bar{x}}) = o$$

Since

$$\sum_{j} (\bar{x}_{.j}-\bar{\bar{x}}) = o \quad , \quad \sum_{i} (\bar{x}_{i.} - \bar{\bar{x}}) = o$$

This applies also to the other two lest sums.
Hence (4) is proved.

(8) Application of the F - test:

Before we apply the F-test to these components in (4), we have to convert them into χ^2 variables.

Let us take variable $\bar{x}_{.j}$: It is a linear combination of the basic variables x_{ij} which as are assumed to be normal.

Hence $\bar{x}_{.j}$ is a normal variable

Also

$$E(\bar{x}_{.j}) = E\left(\frac{1}{a} \sum_{i=1}^{a} x_{ij}\right)$$

$$= \frac{1}{a} \sum_{i=1}^{a} E(x_{ij}) = \frac{1}{a} \sum_{i=1}^{a} (a_i + b_j + c)$$

$$= b_j + c \quad (\text{since} \sum_{i=1}^{a} a_i = o)$$

But when H_o is true $b_j = o$

Hence

$$E(\bar{x}_{.j}) = c$$

Also since $\bar{x}_{.j}$ is a mean of a independent variables

$$V(\bar{x}_{.j}) = \sigma^2 / a$$

Hence all this shows that the variables $\bar{x}_{\cdot j}$ are independently and normally distributed with mean c and variance σ^2/a when H_o is true.

It follows then that

$$\sum_{j=1}^{b} (\bar{x}_{\cdot j} - \bar{x})^2 / \sigma^2/a = \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{x}_{\cdot j} - \bar{x})^2 / \sigma^2 \qquad (5)$$

will possess a χ^2 distribution with (b-1) degrees of Freedom. (on the basis of the assumption H_o).

Now

$\frac{1}{\sigma^2} \sum\sum (x_{ij} - \bar{x})^2$ obviously is χ^2 with (ab-1) D.F.

$\frac{1}{\sigma^2} \sum\sum (\bar{x}_{i\cdot} - \bar{x})^2$ is χ^2 with (a - 1) D.F.

Hence

$$\sum\sum (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 \qquad (6)$$

has a χ^2 distribution with D.F.

$$(ab-1) - \left[(a-1)+(b-1)\right] = ab - (a+b) + 1 = (a-1)(b-1)$$

Hence if (5) is divided by (b-1), and (6) is divided by (a-1)(b-1), the ratio of the resulting quantities will have an F distribution. It is clear that (5) should be used in testing H_o because it measures the variation of column mean, and this variation should prove excessively large when H_o is true as compared to its value when H_o is true. (6) also should be used because it measures the variation in any other factors and thus should prove useful as a basis for comparison.

Summary of the linear hypothesis F test

If the variables x_{ij} are independently and normally distributed with means $\mu_{ij} = a_i + b_j + c$ and variance σ^2, the

hypothesis H_o : $b_j = o$ (j=1, 2, .. b) may be tested by using F distribution where

$$F_{\nu_1,\nu_2} = (a-1)\sum\sum(\bar{x}_{\cdot j} - \bar{\bar{x}})^2 \;/\; \sum\sum(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}})^2$$

and where $\nu_1 = b-1$, $\nu_2 = (a-1)(b-1)$.

Notice: The equality of the row means can be tested in the same manner, only F in this case would take a different expression.

ex/ Four plots of lands growing potatoes were divided into 5 subplots each. For each plot 5 treatments were assigned at random to the 5 subplots. The following table was given. Test whether the 5 treayments are equally effective with respect to mean yield.

(Treatment)

		A	B	C	D	E
	1	310	353	366	299	367
	2	284	293	335	264	314
Plots	3	307	306	339	311	377
	4	267	308	312	266	342

Test of column mean

$$\sum_{j=1}^{5} (\bar{x}_{\cdot j} - \bar{\bar{x}})^2 = 3178, \qquad \sum_{i=1}^{4}\sum_{j=1}^{5} (\bar{x}_{\cdot j} - \bar{\bar{x}})^2 = 4x3178$$

$$\sum_{i=1}^{4} (\bar{x}_{i\cdot} - \bar{\bar{x}})^2 = 1286, \qquad \sum_{j=1}^{5}\sum_{i=1}^{4} (\bar{x}_{i\cdot} - \bar{\bar{x}})^2 = 5x1286$$

$$\sum_{i=1}^{4}\sum_{j=1}^{5} (x_{ij} - \bar{\bar{x}})^2 = 21,530$$

Hence by (4)

$$(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = 21,530 - 4X3178 - 5X1286$$

$$= 2388$$

$$F = \frac{3X4(3178)}{2388} = 16.0 \quad , \quad \nu_1 = 4 \quad , \quad \nu_2 = 12$$

The result is significant.

Hence the 5 treatments undoubtedly differ in their effect on yield. row mean:

$$H_o : a_i = o \qquad (i = 1, 2, \ldots, a)$$

$$F = \frac{4X5(1286)}{2388} = 10.8 \quad , \quad \nu_1 = 3, \quad \nu_2 = 12$$

This result is also significant.

This means that the 4 plots differ in fertility.

The results are usually displayed in a table form

Source of variation	Sum. Sq.	D.F.	M.S.	F
Columns	12,712	4	3178	16.0
rows	6,430	3	2143	10.8
Remainder	2,388	12	199	
Total	21,530	19		

One way classification:

Now suppose that we are conducting the gunnary experiment to test only whether the different brands of shells are of equal quality.

i.e. Suppose that instead of using six men, we use one man in the test. Clearly here our previous assumptions will reduce to

$$\mu_{ij} = b_j + c$$

and we would be testing the hypothesis H_o: $b_j = o$ $(j=1, 2, \ldots b)$.

The analysis of variance formulae would be reduced to

$$\sum_i \sum_j (x_{ij}-\bar{\bar{x}})^2 = \sum_i \sum_j (\bar{x}_j-\bar{\bar{x}})^2 + \sum_i \sum_j (x_{ij}-\bar{x}_j)^2 \qquad (1)$$

This could be proved in the following way:

$$(x_{ij}-\bar{\bar{x}})^2 = (x_{ij}-\bar{x}_j+\bar{x}_j-\bar{\bar{x}})^2$$

$$= (\bar{x}_j-\bar{\bar{x}})^2 + (x_{ij}-\bar{x}_j)^2 + 2(\bar{x}_j-\bar{\bar{x}})(x_{ij}-\bar{x}_j)$$

$$\sum_i \sum_j (x_{ij}-\bar{\bar{x}})^2 = \sum_i \sum_i (\bar{x}_j-\bar{\bar{x}})^2 + \sum_i \sum_j (x_{ij}-\bar{x}_j)^2 + 2\sum\sum(x_j-\bar{\bar{x}})(x_{ij}-\bar{x}_j)$$

Now $\sum_j \sum_i (x_{ij}-\bar{x}_j) = o$

Since $\bar{x}_j = \frac{1}{a} \sum_{i=1}^{a} x_{ij}$

Hence formula (1) follows.

It is clear that $\sum\sum(x_{ij}-\bar{\bar{x}})^2/\sigma^2$ is a χ^2 with (ab-1) D.F.

and that $\sum\sum(\bar{x}_j-\bar{\bar{x}})^2/\sigma^2$ is a χ^2 with (b-1) D.F.

Hence $\sum\sum(x_{ij}-\bar{x}_j)^2/2$ is a χ^2 with (ab-b) D.F.

and hence

$$F = \frac{a(b-1)}{b-1} \frac{\sum\sum(\bar{x}_j-\bar{\bar{x}})^2}{\sum\sum(x_{ij}-\bar{x}_j)^2}$$

This could be summarised as follows:

Source of variation	D.F.	Sum of Squares	Mean Square
Between class means	b-1	$\sum\sum(\bar{x}_j - \bar{\bar{x}})^2$	$\sum\sum$ /b - 1
Within classes	b(a-1)	$\sum\sum(x_{ij} - \bar{x}_j)^2$	$\sum\sum$ /b(a-1)
Total	ab - 1	$\sum\sum(x_{ij} - \bar{\bar{x}})^2$	

Ex. 35, plots of approximately equal fertility were sown with 7 different variaties of wheat, 5 plots to each variaty, the distribution of variaties among the plots being random. The following table gives the yields of grain in bushels per acre, the 7 columns corresponding to the different variaties. Do the data (fictitious) indicate a significant difference in the yields of the variaties?

13	15	14	14	17	15	16
11	11	10	10	15	9	12
10	13	12	15	14	13	13
16	18	13	17	19	14	15
12	12	11	10	12	10	11

Answer:

Source of variation	D.F.	S.S.	M.S.	F.
Between variation	6	41.6	6.933	1.1
Within variaties	28	174	6.214	
Total	34	215.6		

$\nu_1 = 6, \quad \nu_2 = 28$

ex.: The following table given the results obtained from dye trials on each of 5 preparations of Naphthalene Black 12 B made from each of 6 samples of H acid intermadiate.

Yields of Naphthalene Black 12B

Sample of H acid	1	2	3	4	5	6
	1440	1490	1510	1440	1515	1445
Individual yield in	1440	1495	1550	1445	1595	1450
grams of standard	1520	1540	1560	1465	1625	1455
colour	1545	1555	1595	1545	1630	1480
	1580	1560	1605	1595	1635	1520

Does the use of different intermediates gives significantly different yields?

Notice on computation:

Computation could be simplified in many cases:

$$\sum\sum(x_{ij}-\bar{x})^2 = \sum\sum x_{ij}^2 - T^2/ab \qquad (1)$$

where

$$T = \sum\sum x_{ij}$$

Also

$$\sum_j \sum_i (x_{ij}-\bar{x}_j)^2 = \sum_j \left\{ x_{ij}^2 - T_j^2 / a \right\}$$

Where T_j is the sum of the values in the j class.

Hence

$$\sum\sum(x_{ij}-\bar{x}_j)^2 = \sum_i \sum_j x_{ij}^2 - \sum_j T_j^2/a \qquad (2)$$

Hence

$$\sum\sum(x_{ij}-\bar{x}_j)^2 = \sum_j T_j^2/a - T^2/ab \qquad (3)$$

Since the deviations from the means are independent of the choice of origin the results obtained in (1), (2), (3) are unaltered by chance of origin. This would simplify the arithmetic.

ex.: In the first example on fertelises, diminish all the yields by (2)

	1	3	2	2	5	3	4
	-1	-1	-2	-2	3	-3	o
	-2	1	o	3	2	1	1
	4	6	1	5	7	2	3
	o	o	-1	-2	o	-2	-1
T_j	2	9	o	6	17	1	7
$\bar{x}_j$	0.4	1.8	o	1.2	3.4	0.2	1.4

$$T = \sum T_j = 42$$

$$\sum_i \sum_j = 266 \quad , \quad T^2/ab = 50.4, \quad \frac{1}{a}\sum T_i^2 = 92.$$

Hence by (1) , (2) , (3)

$$\sum\sum (x_{ij} - \bar{\bar{x}})^2 = 266 - 50.4 = 215.6$$

$$\sum\sum (x_{ij} - \bar{x})^2 = 266 - 92 = 174$$

$$\sum\sum (\bar{x}_j - \bar{\bar{x}})^2 = 92 - 50.4 = 41.6$$

etc...

"Components of Variance" Model

I) Assumptions involved:

This model makes a linearity assumption about the basic variables x_{ij} rather than about its mean/ ij.

In place of (1) (in case of L.H.M.), it is assumed that x_{ij} could be expressed: