IV-2. Structure and handling of data

IV-2-1. Separation of variances

In reality, results are combined effect of multiple factors. As an example, growth of the plant is affected by temperature, luminance, fertilizer, moisture and so on. In such case, we have to consider influences of combined effect of factors separately from single effect of each factors. The relation between factors and result is complicate, because several factors sometimes have relations. In other words, factors may not independent each other, and impacts of factors are not always orthogonal each other. We have to discuss correlations in several cases. This will be discussed in later paragraph IV-3-3. Simple linear regression and correlation. In this paragraph, we will discuss the case when multiple factors have combined effect though they are independent each other.

For basic understanding, we try to make data set composed from plural factors independent each other.

There is a sub sample population A

$$\mathrm{A}: \{1,5,6\}$$

$$A_1 = 1, A_2 = 5,\ A_3 = 6$$

Allegorical story of this data is follow.

There were 3 pots. The amounts of applied fertilizer are different in nitrogen level among the pots. Height of grass planted 3 different pots were 1cm, 5cm and 6cm.

The data can be summarized as follows.

Sample size (number of the data)

$$n_A = 3$$

Degree of freedom

$$\mathrm{df}_A = n_A - 1 = 3 - 1 = 2$$

Average

$$M_A = \frac{Sum_A}{n_A} = \frac{\sum_{i=1}^{n_A} A_i}{n_A} = \frac{1+5+6}{3} = 4$$

Sum of square

$$SS_A = (1-4)^2 + (5-4)^2 + (6-4)^2 = 14$$

Variance

$${\sigma_A}^2 = \frac{SS_A}{\mathrm{df}_A} = \frac{(1-4)^2 + (5-4)^2 + (6-4)^2}{3-1} = 7$$

Standard deviation

$$\sigma_A = \sqrt{{\sigma_A}^2} = \sqrt{7}$$

On the precondition of normal distribution of data, average and variance can represent the population. Data varies depending on factors, though we assume that we can expect normal distribution and data are obtained without bias.

The other sub sample population is B

$$\mathrm{B}: \{1,5,6,8\}$$

$$B_1 = 1,\ B_2 = 5,\ B_3 = 6, B_4 = 8$$

Allegorical story of this data is as follow.

There were 4 pots. The amounts of applied fertilizer are different in phosphate level among the pots. Height of grass planted 3 different pots were 1cm, 5cm, 6cm and 8cm.

The data can be summarized as follows.

Sample size

$$n_B = 4$$

Degree of freedom

$$\mathrm{df}_B = 4 - 1 = 3$$

Average

$$M_B = \frac{1+5+6+8}{4} = 5$$

Sum of square

$$SS_B = (1-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2 = 26$$

Variance

$$\sigma_B{}^2 = \frac{26}{3} = 8.66667$$

Standard deviation

$$\sigma_B = \sqrt{8.66667}$$

Comparison of tow sub sample population.

A: average $M_A = 4$ degree of freedom $\mathrm{df}_A = 2$ $SS_A = 14$ variance $\sigma^2{}_A = 7$

B: average $M_B = 5$ degree of freedom $\mathrm{df}_B = 3$ $SS_B = 26$ variance$\sigma^2{}_B = 8.66667$

For the comparison of spread of the data by each factor, we calculate the ration of variances.

$$F = \frac{\sigma_B{}^2}{\sigma_B{}^2} = \frac{7}{8.66667} = 0.807692$$

From this result, most of common people sensuously think that the spread by impact of nitrogen is slightly smaller than that of phosphate, though the difference is not prominent. Here we consider sub sample population C of which spread is one tenth of that of B.

$$C: \{0.1, 0.5, 0.6, 0.8\}$$

$$C_1 = 0.1,\ C_2 = 0.5,\ C_3 = 0.6,\ C_4 = 0.8$$

Allegorical story of this data is as follow.

There were 4 pots on which face phot of different persons are pasted. The differences of the height of the grass from a standard height were 0.1cm, 0.5cm, 0.6cm and 0.8cm. Some may think this experiment is no meaning, though we should not make any blind assumption before experiment.

The sub-sample population of C is summarized as follows.

Sample size

$$n_C = 4$$

Degree of freedom

$$\mathrm{df}_C = 4 - 1 = 3$$

Average

$$M_C = \frac{0.1 + 0.5 + 0.6 + 0.8}{4} = 0.5$$

SS

$$SS_C = (0.1 - 0.5)^2 + (0.5 - 0.5)^2 + (0.6 - 0.5)^2 + (0.8 - 0.5)^2 = 0.26$$

Variance

$${\sigma_C}^2 = \frac{0.26}{3} = 0.0866667$$

Standard deviation

$$\sigma_C = \sqrt{0.0866667}$$

Comparison between A and c is as follows

A: average $M_A = 4$ degree of freedom $\mathrm{df}_A = 2$ $SS_A = 14$ variance ${\sigma^2}_A = 7$

C: average $M_C = 0.5$ degree of freedom $\mathrm{df}_C = 3$ $SS_C = 0.26$ variance${\sigma^2}_C = 0.0866667$

$$F = \frac{{\sigma_A}^2}{{\sigma_C}^2} = \frac{7}{0.0866667} = 80.7692$$

The data of A is total height of grass and the data C is difference from a standard point, so comparison of average has no meaning. However, we can conclude that data spread by factor A is nearly hundred times larger than that by C. Little person says that the impact by factor A and C is similar. If we set threshold of F value to reject hypothesis that the impacts of A and C are similar beforehand by F distribution, we can judge the difference is statistically significant for rejection of the hypothesis by comparison of threshold and observed value of F ratio. This is F test. However, we should consider which variances we should compare, and we need to extract the variances for comparison from data. Logically, we can select one factor as main factor which gives impact on the data and can consider other factors as meaningless repeat. However, we generally do not have preliminary knowledge before analysis and it is not scientific to judge by biased preliminary information. Actual data includes various factors. For the discussion of treatment of such data, we try to make dataset produced by overlapping of two additively acting factors. At first, we consider round robin addition of sub population A and B.

Table 1. round robin addition of two factors.

	1	5	6	Sum	Mean
1	1 + 1 = 2	5 + 1 =6	6 + 1 = 7	15	5
5	1 + 5=6	5 + 5 =10	6 + 5 = 11	27	9
6	1 + 6 =7	5 + 6 = 11	6 + 6 = 12	30	10
8	1 + 8 = 9	5 + 8 = 13	6 + 8 = 14	36	12
Sum	24	40	44	108	9
Mean	6	10	11	9	

Values written in red figures are the part determined by factor A and values written in blue figures are the part determined by factor B.

A: average $M_A = 4$ degree of freedom $\mathrm{df}_A = 2$ $SS_A = 14$ variance $\sigma^2{}_A = 7$

B: average $M_B = 5$ degree of freedom $\mathrm{df}_B = 3$ $SS_B = 26$ variance$\sigma^2{}_B = 8.66667$

Total Average is $M = M_A + M_B = 9$

However, the persons who do not know the process of making this data set will calculate statistic values as follow.

Sum

$$Sum_{\ total} = 2 + 6 + 7 + 6 + 10 + 11 + 7 + 11 + 12 + 9 + 13 + 14 = 108$$

Average

$$M_{total} = \frac{Sum_{\ total}}{n_{total}} = \frac{Sum_{\ total}}{n_A n_B} = \frac{108}{12} = 9$$

Sum of square

$$\begin{aligned} SS_{\ total} &= (2-9)^2 + (6-9)^2 + (7-9)^2 + (6-9)^2 + (10-9)^2 + (11-9)^2 + (7-9)^2 \\ &+ (11-9)^2 + (12-9)^2 + (9-9)^2 + (13-9)^2 + (14-9)^2 \\ &= 49 + 9 + 4 + 9 + 1 + 4 + 4 + 4 + 9 + 0 + 16 + 25 = 134 \end{aligned}$$

Degree of freedom

$$\mathrm{df}_{total} = (3 \times 4) - 1$$

Variance

$$\sigma_{total}{}^2 = \frac{SS_{total}}{\mathrm{df}_{total}} = \frac{134}{11} = 12.18182$$

Following table is process of calculation of $SS_{\ total}$

Table 2. Calculation of SS

	1	5	6	Sum
1	49＊	9	4	62
5	9	1	4	14
6	4	4	9	17
8	0	16	25	41
Sum	62	30	42	134

Total SS (134) is obtainable by summing of sum of line or sum of column.

When we focus on the asterisk, the calculation of this cell is $(1+1-9)^2 = 49$. However, this calculation is originally $\{(1-4)+(1-5)\}^2$, when we separate the calculation by original values and averages in original sub sample populations. As the result, sum of first line can be expressed as follow.

$$\{(1-4)+(1-5)\}^2 + \{(5-4)+(1-5)\}^2 + \{(6-4)+(1-5)\}^2 = 62$$

We expand this equation

$$\{(1-4)+(1-5)\}^2 + \{(5-4)+(1-5)\}^2 + \{(6-4)+(1-5)\}^2$$
$$= (1-4)^2 + (5-4)^2 + (6-4)^2 + (1-5)\{(1-4)+(5-4)+(6-4)\} + 3(1-5)^2$$

Here,

$$SS_{\hat{A}} = (1-4)^2 + (5-4)^2 + (6-4)^2 = 14$$
$$(1-4)+(5-4)+(6-4) = 0$$

$(1-4), (5-4), (6-4)$ are distance from average, so the sum or them should be 0.

Conclusively, sum of square in first lien is as follow.

$$SS_A + n_A(1-5)^2 = 14 + 3 \times 16 = 62$$

Similarly,

Second line

$$SS_A + n_A(5-5)^2 = 14 + 3 \times 0 = 14$$

Third line

$$SS_A + n_A(6-5)^2 = 14 + 3 \times 1 = 17$$

Fourth line

$$SS_A + n_A(8-5)^2 = 14 + 3 \times 9 = 41$$

$SS_{\ total}$ is sum of these calculation

$$SS_{\ total} = n_B SS_A + n_A\{(1-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2\}$$

Here,

$$SS_B = (1-5)^2 + (5-5)^2 + (6-5)^2 + (8-5)^2$$

So,

$$SS_{\ total} = n_B SS_A + n_A SS_B = 4 \times 14 + 3 \times 26 = 134$$

Conclusively.

$$SS_{\ total} = n_B SS_A + n_A SS_B$$

This formula is explaining the structure of composing elements of $SS_{\ total}$. Statistical analysis of data in this structure is named Two way analysis of valiance without replication (Two way ANOVA without replication). This analysis is used in analysis of data obtained by experiment combined two factors. Following is an example of such experiment. There 12 planting pots. The pots are divided to 3 groups. each group has 4 pots, and each group is fertilized in different level of nitrogen (A1, A2, A3）. Each pot in in a group is fertilized in different level of phosphate (B1, B2, B3, B4）. The experimental condition of each pots can be expressed as A1B1, A1B2, and so on. Generally, we plant plural number of plants in a pot, so this example is not common, but it is not inconceivable.

Table 3. Example of obtained data (same as table 1)

	A1	A2	A3	Sum	Mean	Square
B1	2	6	7	15	5	$(5-9)^2 *$
B2	6	10	11	27	9	$(9-9)^2$
B3	7	11	12	30	10	$(10-9)^2$
B4	9	13	14	36	12	$(12-9)^2$
Sum	24	40	44	108		24
Mean	6	10	11		9	
Square	$(6-9)^2$	$(10-9)^2$	$(11-9)^2$	14		

Following the previous explanation, we try to calculate without knowledge of origin of the data.

Average of sub sample population $\widehat{\mathrm{A}}$

$$\frac{6+10+11}{3} = 9$$

SS of sub sample population $\widehat{\mathrm{A}}$

$$(6-9)^2 - (10-9)^2 - (11-9)^2 = 14$$

Average of sub sample population $\widehat{\mathrm{B}}$

$$\frac{5+9+10+12}{4} = 9$$

SS of sub sample population $\widehat{\mathrm{B}}$

$$(5-9)^2 - (9-9)^2 - (10-9)^2 + (12-9)^2 = 26$$

We could confirm similarity of calculated variance and original variance of each factor. The author supposes most of readers understand the mechanism of the similarity, though the author explains the mechanism for accurate understanding.

Following is calculation process of sum of the first line tracking back to original data.

$$\{1+1\}+\{5+1\}+\{6+1\}=(1+5+6)+3\times(1)$$

The average is obtained by dividing the sum by number of data.

$$\frac{Sum_A+n_A\times 1}{n_A}=M_A+1$$

Similarly,

Second line

$$\frac{Sum_A+n_A\times 5}{n_A}=M_A+5$$

Third line

$$\frac{Sum_A+n_A\times 6}{n_A}=M_A+6$$

Fourth line

$$\frac{Sum_A+n_A\times 8}{n_A}=M_A+8$$

As shown in above calculation, average in a line is sum of original data of $\widehat{\mathrm{B}}$ and average of factorA. So, the sum of averages of line is

$$n_BM_A=1+5+6+8$$

About column

$$M_B=\frac{1+5+6+8}{n_B}$$

And

$$n_BM_A+1+5+6+8{=}n_BM_A+n_BM_B=n_B(M_A+M_B)$$

Average of average of line is

$$M=\frac{n_BM_A+n_BM_B}{n_B}=M_A+M_B$$

Similarly,

Average of average of column is

$$M=\frac{n_AM_A+n_AM_B}{n_A}=M_A+M_B$$

Then we consider sum of square.

Calculation of square of first line (asterisk) is as follow

$$\{(M_A+1)-(M_A+M_B)\}^2$$

It can be transformed as follow

$$\{(M_A+1)-(M_A+M_B)\}^2=(1-M_B)^2$$

Second line

$$(5 - M_B)^2$$

Third line

$$(6 - M_B)^2$$

Fourth line

$$(8 - M_B)^2$$

Sum of square among four lines in average column is as follow

$$(1 - M_B)^2 + (5 - M_B)^2 + (6 - M_B)^2 + (8 - M_B)^2$$

This is $SS_{\hat{B}}$

$$SS_B = (1 - M_B)^2 + (5 - M_B)^2 + (6 - M_B)^2 + (8 - M_B)^2$$

Similarly,

$$SS_A = (1 - M_A)^2 + (5 - M_A)^2 + (6 - M_A)^2$$

We can confirm the similarity of calculated variance from combined data and original variance of each factor. This result means that we can obtain the variances of original sub sample population from combined data set.

Using this conclusion, the variance of each factor and total variance is as follows

$$\sigma_A{}^2 = \frac{SS_A}{\mathrm{df}_A} = \frac{14}{2} = 7$$

$$\sigma_A{}^2 = \frac{SS_B}{\mathrm{df}_B} = \frac{26}{3} = 8.66667$$

$$\sigma_{total}{}^2 = \frac{SS_{total}}{\mathrm{df}_{total}} = \frac{134}{11} = 12.18182$$

This result clearly shows

$$\sigma_{total}{}^2 \neq \sigma_{\hat{A}}{}^2 + \sigma_{\hat{B}}{}^2$$

Same relation is existing in degree of freedom and sum of square

$$\mathrm{df}_{\hat{A}} = 2$$

$$\mathrm{df}_{\hat{B}} = 3$$

$$\mathrm{df}_{total} = 11$$

$$\mathrm{df}_{total} \neq \mathrm{df}_A + \mathrm{df}_B$$

$$SS\ _{total} \neq SS_A + SS_B$$

The correct relation is as follow

$$SS\ _{total} = n_B SS_A + n_A SS_B$$

$$n_B SS_A = 56, \qquad SS_A = 14$$

$$n_A SS_B = 78. \qquad SS_B = 26$$

$$\sigma_A{}^2 = \frac{SS_A}{n_A - 1} = \frac{14}{2} = 7$$

$$\sigma_B{}^2 = \frac{SS_B}{n_B - 1} = \frac{26}{3} = 8.6667$$

When we consider the number of the other factor as number of repeats, the variance expands by number of repeats in each factor.

We could obtain the variance of each factor composing total variance, the results of calculation is as follows.

$$\sigma_{total}{}^2 = \frac{SS_{total}}{\mathrm{df}_{total}} = \frac{134}{11} = 12.18182$$

$$\sigma_A{}^2 = \frac{SS_A}{\mathrm{df}_A} = \frac{14}{2} = 7$$

$$\sigma_B{}^2 = \frac{SS_B}{\mathrm{df}_B} = \frac{26}{3} = 8.6667$$

Then we compare the variance

$$F_{B-A} = \frac{\sigma_A{}^2}{\sigma_B{}^2} = \frac{7}{8.6667} = 0.807689$$

Then we refer table of F distribution at degree of freedom of numerator 3 and degree of freedom dominator 2. We can judge that the null hypothesis of $F_{B-A} = 1$ is not rejected. When we consider this analysis as two-way ANOVA, F_{B-A} has no meaning. We should compare variance caused by factors with some meaningless variance caused by random fluctuation. However, the author cannot show that in this case, the data not include such fluctuation and the variance to be compared is 0. Variance obtained from actual data includes positive variance caused by unknown minor factors, and each factor is not completely independent from other factors and each observed value include uncertain fluctuation. As an example of comparison between impact of factor and other meaningless minor factor, we try to make combined data of factor A and C

$$C: \{0.1, 0.5, 0.6, 0.8\}$$

$$C_1 = 0.1,\ C_2 = 0.5,\ C_3 = 0.6,\ C_4 = 0.8$$

Sample size

$$n_C = 4$$

Degree of freedom

$$\mathrm{df}_C = 4 - 1 = 3$$

Average

$$M_C = \frac{0.1 + 0.5 + 0.6 + 0.8}{4} = 0.5$$

SS

$$SS_C = (0.1 - 0.5)^2 + (0.5 - 0.5)^2 + (0.6 - 0.5)^2 + (0.8 - 0.5)^2 = 0.26$$

Variance

$$\sigma_C{}^2 = \frac{0.26}{3} = 0.0866667$$

Standard deviation

$$\sigma_C = \sqrt{0.0866667}$$

In previous discussion, we made following round robin table

Table 4 round robin table of adding data

	A1	A2	A3	Sum	Mean
C1	1.1	5.1	6.1	12.3	4.1
C2	1.5	5.5	6.5	13.5	4.5
C3	1.6	5.6	6.6	13.8	4.6
C4	1.8	5.8	6.8	14.4	4.8
Sum	6.0	22.0	26.0	54.0	4.5
Mean	1.5	5.5	6.5	4.5	

In previous discussion we consider that the factor A is significantly affect the growth of plant because F is larg and data spread by factor A is enough large comparing to the spread by factor C. However, when the impact of factor C is significant, we cannot say factor A is insignificant only because of smallness of spread by factor A comparing to the spread by factor C. So, we have considered that fluctuation in factor C was random fluctuation. In more accurate discussion, we need to use model closer to our assumption. Honestly speaking, the author considers that photographs on the plot has no impact to the growth of plant in the pot. It may be a mark to clarify the person responsible to the care of the pot. If so, C is only repeats, and labels of C1, C2, C3 and C4 have no meaning. We can remake the table as follow.

Table 5. Table of One-way ANOVA with repeats

	A1	A2	A3
	1.1	5.1	6.1
	1.5	5.5	6.5
	1.6	5.6	6.6
	1.8	5.8	6.8
Sum	6.0	22.0	26.0
Mean	1.5	5.5	6.5

- Row has no meaning and there is no concept of average of line.

However, there is no change in average and SS

$$\mathrm{M}_{total} = 4.5$$

$$SS_{total} = 56.78$$

Here, we calculate variance of random fluctuation by repeat ($\sigma_{residual}{}^2$) and variance by factor A (σ_A) using the theory that total degree of freedom and SS is sum or partial degree of freedom and SS.

$$\mathrm{df}_{total} = \mathrm{df}_A + \mathrm{df}_{residual}$$

$$\mathrm{df}_{residual} = \mathrm{df}_{total} - \mathrm{df}_A = (n_A n_B - 1) - (n_A - 1) = n_A(n_B - 1)$$

$$SS_{total} = SS_A + SS_{residual}$$

$$\begin{aligned} SS_{residual} &= \{(1.1-1.5)^2 + (1.5-1.5)^2 + (1.6-1.5)^2 + (1.8-1.5)^2\} \\ &\quad + \{(5.1-5.5)^2 + (5.5-5.5)^2 + (5.6-5.5)^2 + (5.8-5.5)^2\} \\ &\quad + \{(6.1-6.5)^2 + (6.5-6.5)^2 + (6.6-6.6)^2 + (6.8-6.5)^2\} \\ &= 3 \times 0.26 = 0.78 \end{aligned}$$

Confirm this value is $n_A SS_{\hat{C}}$

$$n_C SS_A = SS_{total} - SS_{residual} = 56.78 - 0.78 = 56$$

$$n_C SS_A = 4SS_A = 56$$

$$SS_A = 14$$

$$\sigma_{residual} = \frac{SS_{residual}}{\mathrm{df}_{residual}} = \frac{0.78}{3(4-1)} = \frac{0.78}{9} = 0.086667$$

$$\sigma_A = \frac{SS_A}{\mathrm{df}_A} = \frac{14}{2} = 7$$

$$F_{A-residua} = \frac{\sigma_A{}^2}{\sigma_{residual}{}^2} = \frac{7}{0.086667} = 80.76892$$

This ratio is huge, We do not need to confirm table of F distribution to judge significance of factor A. (impact of factor A is enough large comparing to random fluctuation). This is method of one-way ANOVA. Some careful readers may notice that 0.78 is 3×0.26 ($n_A SS_C$). which is obtainable by sum of SS of each column. It is safer to memorize to calculate SS of residual at first and then calculate SS of factor by deduction of SS of residual from total SS. The reader should confirm the sensitivity of detection increase by one-way ANOVA,