IV-2-2. Variance of sum, Variance of difference

IV-2-2-1. Variance of sum

Paired t test is the method to compare paired data such as the difference of length of right hand and left hand in a person, the difference between foot and hand in person difference of the growth two plants planted in a pot. In this case the differences is secondary data obtainable by deduction of the value of from the paired data. However, most of datasets are not paired. For example, it is not possible to compare effect of fertilizer putting two fertilizer in a pot. We should compare growth of plants planting them in different pots. In those case we need to make secondary data expressing the difference from primary data.

There are two data groups A:$\{A_1, A_2, \cdots A_m\}$ and B$\{B_1, B_2, \cdots B_n\}$. We consider secondary data of A + B, A − B, A × B. In case of paired data, we can make secondary data by adding both of pair. A + B: $\{A_1 + B_1, A_2 + B_2, \cdots A_n + B_n\}$. However, in the case of non-paired data, we have no idea to select pair to make sum. A possible solution is to make sum of all possible combination of both groups. the number of the secondary data is mn. This calculation has no practical meaning, though it is useful understanding the understanding the structure of variance.

Sum of data

Following is $m \times n$ round robin

Table 6

	A_1		A_i		A_m	sum	Average
B_1	$A_1 + B_1$		$A_i + B_1$		$A_m + B_1$	$\sum_{i=1}^{m} A_i + mB_1$	$M_A + B_1$
B_j	$A_1 + B_j$		$A_i + B_j$		$A_m + B_j$	$\sum_{i=1}^{m} A_i + mB_j$	$M_A + B_j$
B_n	$A_1 + B_n$		$A_i + B_n$		$A_m + B_n$	$\sum_{i=1}^{m} A_i + mB_n$	$M_A + B_n$
Sum	$nA_1 + \sum_{j=1}^{n} B_j$		$nA_i + \sum_{j=1}^{n} B_j$		$nA_m + \sum_{j=1}^{n} B_j$	$n\sum_{i=1}^{m} A_i + m\sum_{j=1}^{n} B_j$	
Average	$A_1 + M_B$		$A_i + M_B$		$A_m + M_B$		$M_A + M_B$

Example

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

Table 7. Example dataset of summing data

Sum and average

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

Total sum:108, total average M=108/12=9

Average of A: $M_A = 4$ $SS_A = 14$ variance ${\sigma^2}_A = 7$

Average of$B$: $M_B = 5$ $SS_B = 26$ variance ${\sigma^2}_B = 8.66667$

Here, A: original sub sample population used to make combined data set

B: original sub sample population used to make combined data set

Total average is sum of average $\mathrm{M} = \mathrm{M}_A + M_B$ We consider the method to make SS for estimation of variance around mean of parent population. Target SS is SS_{A+B}, and target variance is σ_{A+B}. At first, we do not consider degree of freedom

The data can be express as follow

$$x_i = M + e_i$$

$$M: mean$$

$$e_i: deviation\ from\ mean$$

Formula 38

Secondary moment of sample population is $\mathrm{E}\{(x-M)^2\}$.

The data deviate averagely $\sqrt{\mathrm{E}\{(x-M)^2\}}$ from mean.

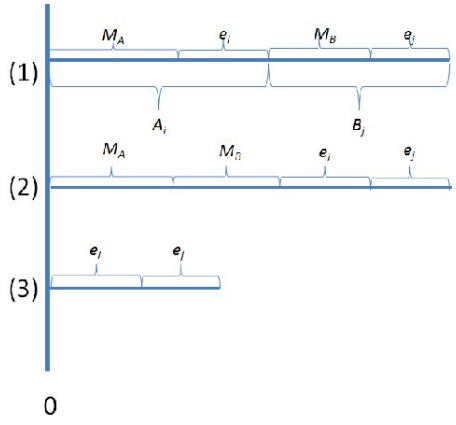

Fig. 32. Structure of combined data

We consider elements composing $A_i + B_j$. Figure 32 shows elements of $A_i + B_j$. The horizontal bar in (1) is number line the end of left side is 0. The line is composed by addition of the element, so the total length is not changed by swapping of the value. And M_A and M_B are average and common among all data. When we consider deviation, we can remove M_A and M_B from total as (3). In this example both e_i and e_j are positive, though both can be negative, and bar in (3) can expand to left from 0.

When we express the length as $e_{A+B_{ij}}$.

$$e_{A+B_{ij}} = e_{A_i} + e_{B_j}$$

Table 8 is result of removing averages in table 6.

Table 8. result of removing average in table 6

	A 1		Ai		Am	sum	Average
B 1	$e_{A_1} + e_{B_1}$		$e_{A_i} + e_{B_1}$		$e_{A_m} + e_{B_1}$	$\sum_{i=1}^{m} e_{A_i} + me_{B_1}$	e_{B_1}
							e_{B_2}
B j	$e_{A_1} + e_{B_j}$		$e_{A_i} + e_{B_j}$		$e_{A_m} + e_{B_j}$	$\sum_{i=1}^{m} e_{A_i} + me_{B_j}$	e_{B_j}
B n	$e_{A_1} + e_{B_n}$	$e_{A_2} + e_{B_n}$	$e_{A_i} + e_{B_n}$		$e_{A_m} + e_{B_n}$	$\sum_{i=1}^{m} e_{A_i} + me_{B_n}$	e_{B_n}
Sum	$ne_{A_1} + \sum_{j=1}^{n} e_{B_j}$	$ne_{A_2} + \sum_{j=1}^{n} e_{B_j}$	$ne_{A_i} + \sum_{j=1}^{n} e_{B_j}$		$ne_{A_m} + \sum_{j=1}^{n} e_{B_j}$	0	0
Average	e_{A_1}	e_{A_2}	e_{A_i}		e_{A_m}	0	0

Total sum and average are 0, because the data are deviation from average. Here, we consider secondary moment of sample population of A+B. Secondary moment is average of sum of square of difference from average. The value of variance of parent population is $\frac{\text{SS}}{degree\ of\ freedom}$, and variance of sample population is $\frac{\text{SS}}{numberof\ data}$.

Total SS is as follow

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{ij}}^2}{mn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(e_{A_i} + e_{B_i}\right)^2}{mn}$$

For simplification,

$$\overline{{e_A}^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{A_{ij}}}^2}{mn}$$

$$\overline{{e_{ij}}^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{ij}}^2}{mn}$$

$\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{A_i}}^2$ is not include elements of B, and $\sum_{j=1}^{n} {e_{A_i}}^2$ is meaning replication times (n).

$$\overline{{e_A}^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{A_i}}^2}{mn} = \frac{n\sum_{i=1}^{m} {e_{A_i}}^2}{mn} = \frac{\sum_{i=1}^{m} {e_{A_i}}^2}{m}$$

Similarly,

$$\overline{{e_B}^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} {e_{B_j}}^2}{mn} = \frac{m\sum_{j=1}^{n} {e_{B_j}}^2}{mn} = \frac{\sum_{j=1}^{n} {e_{B_j}}^2}{n}$$

The relation of averages,

$$\overline{e_{A+B}} = \overline{e_A} + \overline{e_B}$$

Secondary moments is

$$\overline{{e_{A+B}}^2} = \overline{(e_A + e_B)^2} = \overline{{e_A}^2} + 2\overline{e_A e_B} + \overline{{e_B}^2}$$

$\overline{e_A}$、 $\overline{e_B}$ is average of deviation. So, $\overline{e_A} = 0$, $\overline{e_B} = 0$

$$\overline{{e_{A+B}}^2} = \overline{{e_A}^2} + \overline{{e_B}^2}$$

Our purples is estimation of quadratic moment of estimated mean around true mean of parent population ${\sigma_{A+B}}^2$.

$${\sigma_{A+B}}^2 \neq \overline{{e_{A+B}}^2}$$

$$SS_{total} = \sum_{i=1}^{m}\sum_{j=1}^{n} {e_{ij}}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \left(e_{A_i} + e_{B_i}\right)^2$$

$$\overline{e_{A+B}}^2 = \frac{SS_{total}}{mn}, \quad \overline{e_A}^2 = \frac{SS_A}{m}, \quad \overline{e_B}^2 = \frac{SS_B}{n}$$

$$\frac{SS_{A+B}}{mn} = \frac{SS_A}{m} + \frac{SS_B}{n}$$

$$SS_{A+B} = nSS_A + mSS_B$$

Someone who does not know that the values in table 7 is produced by adding two data group, calculates refit side of equation by summing square of vale obtained by deduction of mean from each data. The process is as shown in table 9.

Table 9. Calculation of SS_{total}

	A 1		A i		A m
B 1	$\left(e_{A_1} + e_{B_1}\right)^2$		$\left(e_{A_i} + e_{B_1}\right)^2$		$\left(e_{A_m} + e_{B_1}\right)^2$
B j	$\left(e_{A_1} + e_{B_j}\right)^2$		$\left(e_{A_i} + e_{B_j}\right)^2$		$\left(e_{A_m} + e_{B_j}\right)^2$
B n	$\left(e_{A_1} + e_{B_n}\right)^2$		$\left(e_{A_i} + e_{B_n}\right)^2$		$\left(e_{A_m} + e_{B_n}\right)^2$

Table 10. Expansion of table 9

	A 1		A i		A m	合計
B 1	$e_{A_1}{}^2 + 2e_{A_1}e_{B_1} + e_{B_1}{}^2$		$e_{A_j}{}^2 + 2e_{A_j}e_{B_1} + e_{B_1}{}^2$		$e_{A_m}{}^2 + 2e_{A_m}e_{B_1} + e_{B_1}{}^2$	$SS_{\hat{A}} + me_{B_1}{}^2$
B j	$e_{A_1}{}^2 + 2e_{A_1}e_{B_j} + e_{B_j}{}^2$		$e_{A_i}{}^2 + 2e_{A_i}e_{B_j} + e_{B_j}{}^2$		$e_{A_m}{}^2 + 2e_{A_m}e_{B_j} + e_{B_j}{}^2$	$SS_{\hat{A}} + me_{B_m}{}^2$
B n	$e_{A_1}{}^2 + 2e_{A_1}e_{B_n} + e_{B_n}{}^2$		$e_{A_i}{}^2 + 2e_{A_i}e_{B_n} + e_{B_n}{}^2$		$e_{A_m}{}^2 + 2e_{A_m}e_{B_n} + e_{B_n}{}^2$	$SS_{\hat{A}} + me_{B_m}{}^2$
Sum	$ne_{A_1}{}^2 + SS_{\hat{B}}$		$ne_{A_i}{}^2 + SS_{\hat{B}}$		$ne_{A_m}{}^2 + SS_{\hat{B}}$	$nSS_{\hat{A}} + mSS_{\hat{B}}$

Calculation of yellow part in table 10.

$$\sum_{i=1}^{m} e_{A_i}{}^2 + 2e_{B_j}\sum_{i=1}^{m} e_{A_i} + \sum_{i=1}^{m} e_{B_j}{}^2$$

First term is SS, second term is sum of deviation $\sum_{i=1}^{m} e_{A_i} = 0$, and the third term does not include i

$$\sum_{i=1}^{m} e_{A_i}{}^2 + 2e_{B_j}\sum_{i=1}^{m} e_{A_i} + \sum_{i=1}^{m} e_{B_j}{}^2 = SS_A + me_{B_m}{}^2$$

When we put the values in the table 7, table 11 is obtained

Table 11. An example of 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

As an example, 49* is calculated as $(2-9)^2$, though same result can be obtained by $\big((2-6)+(2-5)\big)^2$

Value marked by yellow (62) is 14+3×16=62

$nSS_A + mSS_B$ is 4×14+3×26＝134

Please try the calculation from the data (2, 6, 7, 6, 10, 11, 7, 11,12, 9, 13,14) to obtain average and SS, it will be the same as above result.

$$SS_{total} = nSS_A + mSS_B$$

Formula 39

We go back to formula 39 and remember the purpose of calculation of SS. The purpose is estimation of secondary moment of parent population.

$$\sigma^2 = \frac{SS}{numbe\ of\ data - 1}$$

Our purpose is estimation of $\sigma_{A+B}{}^2$

When we go back to Table 8.

Table 8. result of removing average in table 7

	A 1		A i		A m	sum	Average
B 1	$e_{A_1}+e_{B_1}$		$e_{A_i}+e_{B_1}$		$e_{A_m}+e_{B_1}$	$\sum_{i=1}^{m} e_{A_i}+me_{B_1}$	e_{B_1}
							e_{B_2}
B j	$e_{A_1}+e_{B_j}$		$e_{A_i}+e_{B_j}$		$e_{A_m}+e_{B_j}$	$\sum_{i=1}^{m} e_{A_i}+me_{B_j}$	e_{B_j}
B n	$e_{A_1}+e_{B_n}$	$e_{A_2}+e_{B_n}$	$e_{A_i}+e_{B_n}$		$e_{A_m}+e_{B_n}$	$\sum_{i=1}^{m} e_{A_i}+me_{B_n}$	e_{B_n}

Averages of each row are e_{B_j}, and deviation of each cell from average of row is e_{A_i}. Thus sum of square in each row is

$$\sum_{i=1}^{m} {e_{A_i}}^2 = SS_A$$

Quadratic moment around mean is $\frac{SS_A}{m-1}$. This is a estimated value of $\sigma_{A+B}{}^2$

$$\sigma_{A+B}{}^2 = \frac{SS_A}{m-1} = \sigma_A{}^2$$

Similarly, Averages of each column is e_{A_j}. This is another estimation of $\sigma_{A+B}{}^2$.

$$\sigma_{A+B}{}^2 = \frac{SS_B}{m-1} = \sigma_B{}^2$$

Actually, the value of $\sigma_A{}^2$, and $\sigma_B{}^2$ is different, though both $\sigma_A{}^2$ and $\sigma_B{}^2$ are expectation value of $\sigma_{A+B}{}^2$. From upper formulas

$$(m-1)\sigma_{A+B}{}^2 = SS_A = (m-1)\sigma_A{}^2$$
$$(n-1)\sigma_{A+B}{}^2 = SS_B = (n-1)\sigma_B{}^2$$

Combine both equations.

$$(m+n-2)\sigma_{A+B}{}^2 = SS_A + SS_B = (m-1)\sigma_A{}^2 + (n-1)\sigma_B{}^2$$

$$\sigma_{A+B}{}^2 = \frac{SS_A + SS_B}{m+n-2} = \frac{(m-1)\sigma_A{}^2 + (n-1)\sigma_B{}^2}{m+n-2}$$

From this we can conclude that

$$SS_{A+B} = SS_A + SS_B$$
$$df = m+n-2$$
$$\sigma_{A+B}{}^2 = \frac{(m-1)\sigma_A{}^2 + (n-1)\sigma_B{}^2}{m+n-2}$$

Formula 40

This formula is explained as weighted mean by degree of freedom. This is correct. Several text books explain this as follows. When we calculate deviation, we use average. The degree of freedom decreases with the calculation of average. In this case we did averaging twice. One is the process of making SS of group A, the other is in the process of making SS of group B. This explanation is intuitively acceptable. However, it is not logical explanation. The process used in this text to make variance of combined group is natural and logical. However, the author put a trick in the process of explanation. This trick is necessary to make simple story. Before the explanation of the trick, some readers may already notice an uncomfortable feeling.

The sample size of A+B is $m+n-1$, and degree of freedom is $m+n-2$. Total sample size is mn, and the degree of freedom should be $mn-1$. This means that when we use total as meaning of all data, this total is not meaning A+B. When we use total as meaning of all data and A+B as sum of A and B, each degree of freedom is as follows.

$$\mathrm{df}_{total}=mn-1$$

$$\mathrm{df}_{A+B}=m+n-2$$

df: degree of freedom

From this, we can estimate that there are other variances which correspond to the difference of degree of freedom theoretically. This variance happens by combination of groups. In this meaning, the variance is named interaction. Sometimes, it is called residual as meaning of unexplainable factor. This wording is not correct. $\mathrm{A}\times\mathrm{B}$ is often used as symbol of interaction. The degree of freedom is as follow.

$$\mathrm{df}_{A\times B}=\mathrm{df}_{total}-df_{A+B}=(mn-1)-(m+n-2)=(m-1)(n-1)$$

This means that degree of freedom of interaction is product of degree of freedom of combined elements.

The trick the author put in the explanation is follow

See table 7 carefully again

Table 7

Sum and average

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

This table is unnatural. The differences between neighboring cell are the same among lines, and difference from immediate above cell is the same among column (5-1=4, 6-2 =4, 10-6 =4, 11-7=4, 13-19=4, $\cdots, 7-9=-2, 11-13=-2, 12-14=-2$). This is obvious before confirmation, because we made the dataset by round robin table, and listed it round robin table form. This is the trick of the author. When we list the data randomly in the table the result will be change, and we can detect variance of interaction. He made this unnatural table to make variance of interaction 0. In this text we are discussing method of separation of variance depending on the factors. In several analyses, the data includes various factors. In such case we need to separate effect of each factor including impact of interaction.

IV-2-2-2. Variance of difference

Logic of variance of difference is same as variance of sum though the variance of difference has applicative meaning such as testing of statistical significance of the difference of growth rate of plants.

There are two data groups A:$\{A_1, A_2, \cdots A_m\}$ and B$\{B_1, B_2, \cdots B_n\}$. We consider secondary data of A − B.

Following is $m \times n$ round robin subtraction.

Table 12. Round robin subtraction

	A 1		Ai		A m	Sum
B 1	$A_1 - B_1$		$A_i - B_1$		$A_m - B_1$	$\sum_{i=1}^{m} A_i - B_1$
B j	$A_1 - B_j$		$A_i - B_j$		$A_m - B_j$	$\sum_{i=1}^{m} A_i - B_j$
B n	$A_1 - B_m$		$A_i - n$		$A_m - B_n$	$\sum_{i=1}^{m} A_i - B_n$
Sum	$A_1 - \sum_{j=1}^{n} B_j$		$A_i - \sum_{j=1}^{n} B_j$		$A_m - \sum_{j=1}^{n} B_j$	$n\sum_{i=1}^{m} A_i - \mathrm{m}\sum_{j=1}^{n} B_j$

As an example, A:{1,5,6}, B:{1,5,6,8}

Total sum:108, total average M=108/12=9

Average of A: $M_A = 4$ $SS_A = 14$ variance $\sigma^2{}_A = 7$

Average of $\widehat{B}$: $M_B = 5$ $SS_B = 26$ variance $\sigma^2{}_B = 8.66667$

Here, A: original sub sample population used to make combined data set

B: original sub sample population used to make combined data set

Table 13　sample dataset

Sum and Average

	1	5	6	Sum	Mean
1	0	4	5	9	3
5	−4	0	1	−3	−1
6	−5	−1	0	−6	−2
8	−7	−3	−2	−12	−4
Sum	−16	0	4	−12	−1
Mean	−4	0	1	−1	

Total sum is -12 and average is M=-12/12=-1

$$M = M_A - M_B$$

Table 14. Calculation of variance (SS)

	A_1		A_i		A_m	合計
B_1	$(A_1 - B_1 - M)^2$		$(A_i - B_1 - M)^2$		$(A_m - B_1 - M)^2$	$SS_{\hat{A}} + {me_{B_1}}^{2*}$
B_j	$\left(A_1 - B_j - M\right)^2$		$\left(A_i - B_j - M\right)^2$		$\left(A_m - B_j - M\right)^2$	${SS_{\hat{A}} + me_{B_j}}^2$
B_n	$(A_1 - B_n - M)^2$		$(A_i - B_n - M)^2$		$(A_m - B_n - M)^2$	${SS_{\hat{A}} + me_{B_n}}^2$
Sum	${ne_{A_1}}^2 + SS_{\hat{B}}$		${ne_{A_i}}^2 + SS_{\hat{B}}$		${ne_{A_m}}^2 + SS_{\hat{B}}$	$nSS_{\hat{A}} + mSS_{\hat{B}}$

The formula in the line and column of sum is same as table 10.

The calculation of actual data is as follow.

Table 15.　Calculation of SS by actual data

	1	5	6	Sum
1	1	25	36	62
5	9	1	4	14
6	16	0	1	17
8	36	4	1	41
Sum	62	30	42	134

Sum in table 15 is the same as table 11.

When we see table 14, sum in a line is

$$\left(A_1 - B_j - M\right)^2 + \cdots + \left(A_i - B_j - M\right)^2 + \left(A_m - B_j - M\right)^2$$
$$= \sum_{i=1}^{m} \left((A_i - M_A) - \left(B_j - M_B\right)\right)^2$$

$$= \sum_{i=1}^{m}(A_i - M_A)^2 + 2(B_j - M_j)\sum_{i=1}^{m}(A_i - M_A) + \sum_{i=1}^{m}(B_j - M_B)^2$$

$$= \mathrm{SS}_A + +me_j{}^2$$

$$\because \sum_{i=1}^{m}(A_i - M_A)^2 = SS_A$$

$$\sum_{i=1}^{m}(A_i - M_A = 0)$$

$$\sum_{i=1}^{m}(B_j - M_B)^2 = \sum_{i=1}^{m} e_{B_j}{}^2 = me_{B_j}{}^2$$

This result is sensuously acceptable, when we see the table. From this we can conclude

$$\sigma_{A+B}{}^2 = \frac{(m-1)\sigma_A{}^2 + (n-1)\sigma_B{}^2}{m+n-2}$$

$$\sigma_{A-B}{}^2 = \frac{(m-1)\sigma_A{}^2 + (n-1)\sigma_B{}^2}{m+n-2}$$

Formula 41

The author demonstratively clarifies the difference two variances by index of A − B and A + B, though the tow variance is the same and $\sigma_{A+B}{}^2$ is generally used in the meaning of combined variance. This variance is used in t test for detection of significance of difference.

What we confirmed in this chapter is follows.

1. Total sum of square (SS) is sum of partial sum of square.
2. Total degree of freedom is sum of partial sum of square.
3. $\sigma_{A-B}{}^2 = \sigma_{A+B}{}^2 = \frac{(m-1)\sigma_A{}^2+(n-1)\sigma_B{}^2}{m+n-2}$