37 Incomplete Data

37.1. Introduction

Sometimes during experimentation or simulation, data cannot be obtained as planned because of lost data, equipment problems, or extreme conditions. Sometimes, the SN ratios of a few runs of an orthogonal array cannot be calculated, for various reasons. In the treatment of missing data in the traditional design of experiments, no distinction is made between different types of incomplete data. In this chapter, treatments of such cases are described.

Although it is desirable to collect a complete set of data for analysis, it is not absolutely necessary. In parameter design, it is recommended that control-factor-level intervals be set wide enough so that a significant improvement in quality may be expected. Such determinations can only be made based on existing knowledge or past experiences. In product development, however, there is no existing knowledge to determine the range of control factor levels. To avoid generating undesirable results, engineers tend to conduct research within narrow ranges.

If control-factor-level intervals were set wide, the conditions of some control-factor-level combinations may become so extreme that there would be no product or no measurement: for example, an explosion in a chemical reaction. That's all right, because important information was produced from the result. It must be remembered that the objective of new product development is to produce good knowledge rather than producing a good product. The incomplete data in that case can be analyzed by a method called *sequential approximation.*

Various types and examples of incomplete data follow:

1. All data in one run (or a few runs) of an orthogonal array are missing.
   - 1.1 Samples to be measured are missing.
   - 1.2 Data sheets with the recorded results are misplaced.
   - 1.3 Some experimental runs were discontinued due to budget, time constraints, or job transfer.
   - 1.4 The raw materials or test pieces needed to complete the whole runs of experiments were insufficient.
2. The number of data points in one run are different from others.
   - 2.1 The number of signal factor levels in one run differs from those in the other runs.
   - 2.2 The number of repetitions in one run are different from other runs.
   - 2.3 All results in one noise factor level of one run are missing.
3. Part or all the results of one run is missing due to extreme conditions.
   - 3.1 The chemical reaction apparatus exploded.
   - 3.2 No current flowed in an electric circuit.
   - 3.3 No product was produced.
4. No data are missing, but calculation of the SN ratios results in values of either positive or negative infinity.
   - 4.1 When using the zero-point proportional, reference-point proportional, or linear equation, the error variance (V_e) is greater than the sensitivity (S_β).
   - 4.2 In the nominal-is-best case, the error variance (V_e) is greater than the sensitivity (S_m).
   - 4.3 In classified data with three classes or more, the error variance (V_e) is greater than the sensitivity (S_m).
   - 4.4 In the smaller-is-better case, all results are equal to zero.
   - 4.5 In the larger-is-better case, all results are equal to zero.

Levels of conditions must be within a certain range so that undesirable results are not obtained. Such determinations can only be made based on existing knowledge or past experience. In the case of new product development, however, there is no existing knowledge on which to base determination of the appropriate range of control factor levels within which a product or a system functions. To avoid generating undesirable results, engineers tend to conduct research within narrow ranges.

It is important to realize that the purpose of product development is not to produce good products but to generate useful knowledge. If all results from different experimental runs are close to each other, little knowledge can be gained. To get more knowledge about a new frontier, control-factor-level intervals need to be set wide enough to purposely produce some bad results, such as having explosions or cracked test pieces. Such situations are allowed at the R&D stage. In the case of an L_{18} orthogonal array experiment, there may be as many as several runs with incomplete data, but this is still good enough to draw valuable conclusions.

37.2. Treatment of Incomplete Data

To make analysis and optimization possible, the following treatments are suggested for the cases listed in Section 37.1.

Cases 1.1 to 1.4

Cases 1.1 to 1.4 are referred to as missing data in traditional experimental design books. The Fisher–Yates method is commonly used. The sequential approximation method is recommended.

Case 2.1

An example of case 2.1 is when the number of signal factor level of run 4 is 3 but all other runs are 5. The SN ratio of run 4 is calculated as is and analyzed with the other SN ratios.

Case 2.2

For case 2.2, the number of repetitions under each signal factor are different, as shown in Table 37.1. In the case of the zero-point proportional equation, the SN ratio is calculated as follows:

$$S_T = y_{11}^2 + y_{12}^2 + \cdots + y_{kr_k}^2 \qquad (f_T = r_1 + r_2 + \cdots + r_k) \tag{37.1}$$

$$S_\beta = \frac{(M_1 y_1 + M_2 y_2 + \cdots + M_k y_k)^2}{r_1 M_1^2 + r_2 M_2^2 + \cdots + r_k M_k^2} \qquad (f_\beta = 1) \tag{37.2}$$

$$S_e = S_T - S_\beta \qquad (f_e = f_T - 1) \tag{37.3}$$

$$V_e = \frac{S_e}{f_e} \tag{37.4}$$

$$\eta = 10 \log \frac{[1/(r_1 M_1^2 + r_2 M_2^2 + \cdots + r_k M_k^2)](S_\beta - V_e)}{V_e} \tag{37.5}$$

Table 37.1
Incomplete data

Signal	M_1	M_2	⋯	M_k
Repetition	y_{11} y_{12} ⋮ y_{1r1}	y_{21} y_{22} y_{2r2}	⋯ ⋯ ⋯	y_{k1} y_{k2} y_{krk}
Total	y_1	y_2	⋯	y_k

Table 37.2 is a simple numerical example [1]. For the data in the table:

$$\begin{aligned} S_T &= 0.098^2 + 0.097^2 + \cdots + 0.488^2 \\ &= 2.002033 \qquad (f_T = 4 + 6 + 6 = 16) \end{aligned} \tag{37.6}$$

$$\begin{aligned} S_\beta &= \frac{[(0.1)(0.380) + (0.3)(1.750) + (0.5)(2.955)]^2}{(4)(0.1^2) + (6)(0.3^2) + (6)(0.5^2)} = \frac{2.0405^2}{2.08} \\ &= 2.00175012 \qquad (f = 1) \end{aligned} \tag{37.7}$$

$$\begin{aligned} S_e &= 2.002033 - 2.00175012 \\ &= 0.0028288 \qquad (f = 16 - 1 = 15) \end{aligned} \tag{37.8}$$

$$V_e = \frac{0.00028288}{15} = 0.00001886 \tag{37.9}$$

$$\begin{aligned} \eta &= 10 \log \frac{(1/2.08)(2.00175012 - 0.00001886)}{0.00001886} \\ &= 10 \log 51027.08 \\ &= 47.08 \text{ dB} \end{aligned} \tag{37.10}$$

Case 2.3

An example of case 2.3 is when there are two compounded noise factor levels: the positive and negative extreme conditions. If one of these two extreme conditions is totally missing, do not use the results of another extreme condition to calculate the SN ratio. Instead, treat it as described for type 1.

Cases 3.1 to 3.3

Whether it's an explosion, lack of current, or lack of end product, each problem indicates that the condition is very bad. Use negative infinity as the SN ratio. It is important to distinguish this case from type 1, in which it is not known whether the condition is good or bad.

There are two ways to treat these cases: (1) classify the SN ratios, including positive and/or negative infinity, into several categories and analyze the results by accumulation analysis [2]; and (2) subtract 3 to 5 dB from the smallest SN ratio in the orthogonal array. Assign that number to the SN ratio for the missing run(s). Then follow with sequential approximation.

Table 37.2
Numerical example

Signal	$M_1 = 0.1$	$M_2 = 0.3$	$M_3 = 0.5$
Result	0.098	0.294	0.495
	0.097	0.288	0.493
	0.093	0.288	0.489
	0.092	0.296	0.495
		0.297	0.495
		0.287	0.488
Total	0.380	1.750	2.955

Cases 4.1 to 4.3

Cases 4.1 to 4.3 involve error variance being greater than the sensitivity. Use negative infinity for the SN ratio and treat these cases in the same way as case 3.1.

Case 4.4

Case 4.4 is when all results are equal to zero for a smaller-is-better case. Use positive infinity as the SN ratio and classify the SN ratio into several categories and conduct accumulation analysis. Add 3 to 5 dB to the largest SN ratio in the orthogonal array. Assign that number to the SN ratio of the missing run(s), followed with sequential approximation.

Case 4.5

Case 4.5 is when all results are equal to zero for a larger-is-better case. Use negative infinity, and treat the problem in the same way as case 3.1.

37.3. Sequential Approximation [3]

Sequential approximation is a method used to estimate the data that are supposed to exist. In the experiment for the amount of wear, assume that the data of experiment 3 in orthogonal array L_{12} are missing (Table 37.3). This is the case of missing data; the SN ratio for a smaller-the-better characteristic cannot be calculated.

Sequential approximation is made using the following procedure:

1. Place the average calculated from the existing SN ratio into the missing location; this is called the *zeroth approximation*.

Table 37.3
Missing data in experiment 3

Factor:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	Amount of Wear (μm)		
No.\Column:	1	2	3	4	5	6	7	8	9	10	11	N_1	N_2	η (dB)
1	1	1	1	1	1	1	1	1	1	1	1	39 19	9 10	−27.12
2	1	1	1	1	1	2	2	2	2	2	2	21 20	3 16	−24.42
3	1	1	2	2	2	1	1	1	2	2	2			
4	1	2	1	2	2	1	2	2	1	1	2	42 27	13 24	−29.08
5	1	2	2	1	2	2	1	2	1	2	1	37 35	9 29	−29.44
6	1	2	2	2	1	2	2	1	2	1	1	68 85	38 64	−36.38
7	2	1	2	2	1	1	2	2	1	2	1	21 10	5 2	−21.54
8	2	1	2	1	2	2	2	1	1	1	2	38 23	17 4	−27.55
9	2	1	1	2	2	2	1	2	2	1	1	30 34	30 24	−29.46
10	2	2	2	1	1	1	1	2	2	1	2	72 46	24 40	−33.75
11	2	2	1	2	1	2	1	1	1	2	2	10 6	1 2	−15.47
12	2	2	1	1	2	1	2	1	2	2	1	30 12	8 0	−24.42

2. Conduct response analysis and estimate the missing location using larger effects; call this the *first approximation.*
3. Repeat step 2 until the estimation converages.

The following is the estimation of experiment 3 using sequential approximation:

1. *Zeroth approximation.* The average of 11 pieces of data is −27.15 dB.
2. *ANOVA and estimation of experiment 3.* Put −27.15 dB into experiment 3, and the ANOVA table is used to find larger effects. Tables 37.4 and 37.5 show the ANOVA and supplementary tables.

From Table 37.4, larger effects are *A, C, I,* and *J.* They are used to estimate the condition of experiment 3: $A_1(B_1)C_2(D_2E_2F_1G_1H_1)I_2J_2(K_2)$.

$$\begin{aligned}\hat{\mu}(\text{No. 3}) &= \frac{-173.59 + (-175.81) + (-175.58) + (-142.44)}{6} - (3)\left(\frac{-325.78}{12}\right)\\ &= -29.78 \text{ dB} \qquad (37.11)\end{aligned}$$

The first approximation is −29.78 dB.

Repeat step 2 using −29.78 dB. The following result is obtained:

$$\hat{\mu}(\text{No. 3}) = -30.88 \text{ dB} \qquad (37.12)$$

Using the same procedure, the fifth approximation is obtained as −31.62 dB. This is close to the fourth approximation of −31.53, and calculation is discontinued. The actual result was −34.02; these are fairly close.

Table 37.4
ANOVA table using zeroth approximation

Factor	*f*	*S*	*V*
A	1	38.16	38.16
B	1	10.64	10.64°
C	1	55.64	55.64
D	1	4.84	4.84°
E	1	5.91	5.91°
F	1	0.01	0.01°
G	1	0.08	0.08°
H	1	7.68	7.68°
I	1	53.68	53.68
J	1	139.40	139.40
K	1	9.97	9.97
Total	11	326.02	

°: Pooled as error.

Table 37.5
Supplementary table

Factor	Level Total	Factor	Level Total
A_1	−173.59	G_1	−162.39
A_2	−152.19	G_2	163.39
B_1	−157.24	H_1	−158.09
B_2	−168.54	H_2	−167.69
C_1	−149.97	I_1	−150.20
C_2	−175.81	I_2	−175.58
D_1	−166.70	J_1	−183.34
D_2	−159.08	J_2	−142.44
E_1	−158.68	K_1	−168.36
E_2	−167.10	K_2	−157.42
F_1	−163.06		
F_2	−162.72	Total	−324.78

The calculation above is terminated at the fifth approximation using four factors: *A*, *C*, *I*, and *J*. Since the effect of *E* becomes larger and larger, the approximation could be better if *E* were included in the calculation.

Next, the optimum condition is estimated using the ANOVA and supplementary tables from the fifth approximation (Tables 37.6 and 37.7).

Table 37.6
Supplementary table using fifth approximation

Factor	Level Total	Factor	Level Total
A_1	−178.06	G_1	−166.86
A_2	−152.19	G_2	−163.39
B_1	−161.71	H_1	−162.56
B_2	−168.54	H_2	−167.69
C_1	−149.97	I_1	−150.20
C_2	−180.28	I_2	−180.05
D_1	−166.70	J_1	−183.34
D_2	−163.35	J_2	−146.92
E_1	−158.68	K_1	−168.36
E_2	−171.57	K_2	−161.89
F_1	−167.53		
F_2	−162.72	Total	−330.25

Table 37.7
ANOVA table using fifth approximation

Factor	*f*	*S*	*V*	ρ (%)
A	1	55.77	55.77	15.6
B	1	3.89	3.89°	
C	1	76.56	76.56	21.6
D	1	0.89	0.83°	
E	1	13.83	13.85	3.4
F	1	1.93	1.93°	
G	1	1.00	1.00°	
H	1	2.19	2.19°	
I	1	74.25	74.25	20.9
J	1	110.60	110.60	31.5
K	1	3.49	3.49°	
(*e*)	(6)	(13.33)	(2.22)	(7.0)
Total	11	344.36		100.0

°: Pooled as error.

The optimum condition is $A_2B_1C_1D_2E_1F_2G_2H_1I_1J_2K_2$. Using the effects of *A*, *C*, *E*, *I*, and *J*, we have

$$\begin{aligned}\hat{\mu} &= \overline{A}_2 + \overline{C}_1 + \overline{E}_1 + \overline{I}_1 + \overline{J}_2 - 4\overline{T}\\ &= -25.36 - 25.00 - 26.45 - 25.03 - 24.48 - (4)(-27.52)\\ &= -16.24 \text{ dB} \qquad (37.13)\end{aligned}$$

Sequential approximation can be used when there is more than one missing experiment. But it cannot be used when all data under a certain level of a certain factor are missing.

References

1. Yuin Wu et al., 2000. *Taguchi Methods for Robust Design.* New York: ASME Press.
2. Genichi Taguchi, 1987. *System of Experimental Design.* Livonia, Michigan: Unipub/American Supplier Institute.
3. Genichi Taguchi et al., 1993. *Quality Engineering Series,* Vol. 4. Tokyo: Japanese Standards Association.