Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 3, pp. 321-325, June 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.3.321

<Research Letter>

결측치가 있는 실험계획 데이터의 분석 방법에 관한 소고

변 재 현†

경상대학교 산업시스템공학부

A Study on Analyzing Data from Designed Experiments with Missing Values

Jai-Hyun Byun

Department of Industrial and Systems Engineering, Gyeongsang National University

A data-oriented approach is proposed to analyze experiments designed data with missing values resulting from the invalidity of the design points. The procedure consists of representing available data in a normal probability plot, identifying better design points and the factor levels common to these points, and then deciding on optimal conditions in the experimental region or the direction of improvement. Proposed approach is illustrated with analyses of a fractional factorial design and a robust design data. The procedure is expected to be useful to the practitioners implementing design of experiments to product/process development.

Keywords: Design of Experiments, Missing Data, Screening Design, Fractional Factorial Design, Robust Design, Product Development

1. 서 론

자료수집(Data collection)은 시스템을 분석, 관리, 또는 설계하는 데 있어서 필수적인 활동이다. 수집된 데이터는 과거 데이터, 관찰 데이터, 실험데이터로 구분할 수 있다. 과거 데이터를 이용하여 정보를 추출하는 데이터 마이닝 방법은 비용이 가장 적게 든다는 장점이 있는 대신에 과거와 현재 상황이 다르거나 데이터에 오류가 있을 때, 또는 분석자가 관심 있는 영역의 데이터가 없을 때는 활용 가치가 떨어진다. 관찰 데이터는 관심 있는 문제에 대해 사전에 잘 계획하여 얻은 것이기 때문에 정상상태의 프로세스의 특성에 관한 정보를 잘 제공한다. 관찰 연구는 주로 관리도를 이용하여 프로세스를 모니터하거나, 회귀분석 등의 방법을 이용하여 제품/공정의 변수 간 관계성을 파악하는 것이다. 실험 연구는 실험자가 시스템 또는 프로세스에 개입하여 의도적으로 투입변수(Input variables)의 값을 바꾸었을 때 시스템의 주요 성능 특성이 어떻게 변하는지를 확인하고 최적의 투입변수 조합을 정하는 활동이다. 실험데이터를 효율적이고 효과적으로 얻는 방법이 실험계획법인데, 변수선별, 주요 변수 대상의 최적화, 강건성 확보의 단계를 따른다(Vining, 2013).

실험을 통하여 제품개발이나 공정개선을 도모할 때, 결측치가 발생하는 경우가 자주 발생한다. 결측치는 1) 실험 수행 후 데이터 기록이 없거나, 2) 실험 시료에 문제가 있거나, 3) 실험 시간 또는 자원의 부족, 장비 고장으로 실험을 더는 진행할 수가 없거나, 또는 4) 해당 실험점의 실험조건에서 필요한 형상이나 성능을 확보하지 못하여 분석할 만한 데이터를 얻을 수 없는 경우 등 다양한 이유로 발생한다. 위의 처음 3가지 이유 중 하나로 결측치가 발생하면, 해당 결측치 칸에 값을 추정하여 분석할 수 있다. 반복 실험에서는 결측이 발생하면 해당 실험조건의 나머지 데이터의 평균값을 이용하면 된다. 반복이 없는 완전 무작위 블록계획(Randomized complete block design)이나 라틴방격법(Latin squares)에서는 오차제곱합(Error sum of squares)이 최소가 되도록 결측치를 추정한다(Montgomery, 2017). 반복이 없는 요인배치에서는 오차항을 구할 수 없으므로 최고차 교호작용의 값이 0이 되도록 결측치를 추정한다(Goh, 1997). 예를 들어, A, B, C 3개의 인자를 대상으로 반복이 없는 2수준 3인자 요인배치의 어느 실험점에서

본 연구는 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2018R1D1A1B07049764).
† 연락저자 : 변재현 교수, 52828 경남 진주시 진주대로 501 경상대학교 산업시스템공학부, Tel : 055-772-1692, Fax : 055-772-1699, E-mail : jbyun@gnu.ac.kr
2020년 10월 12일 접수; 2020년 11월 16일 수정본 접수; 2020년 11월 17일 게재 확정.

결측이 발생하면, 3인자 교호작용인 ABC가 0이 되도록 결측치를 추정하는 것이다. 반복이 없는 부분요인배치법을 이용할 때는 결측치 추정을 위하여 어떤 요인효과를 희생할지 결정하기가 쉽지 않다. 만일 결측치가 해당 실험점의 실험조건 때문에 발생한 것이 아니고, 추가 실험을 할 수 있도록 시간과 자원이 충분히 있다면, 가장 좋은 방법은 결측치가 발생한 실험점에서 실험을 다시 수행하여 데이터를 얻는 것이다. 물론 이때에는 기존 실험과 추가 실험의 환경에 차이가 발생할 수 있으므로 블록 효과의 발생 여부를 판단하기 위하여 애초 실험에서 시행한 다른 실험점에서도 실험을 수행하여 실험환경에 차이가 발생하는지를 확인해야 할 것이다. 문제는 결측치가 발생한 실험점 자체가 문제가 있어서 이 실험점의 조건에서 필요한 데이터를 얻을 수 없는 경우에 활용할 방법이 아직 없다는 것이다. 본 연구에서는 이러한 경우에 활용할 수 있는 방법을 제안하고자 한다.

제 2장에는 결측치가 어떤 오류나 실험상황에서 발생하는 것이 아니고 해당 실험점에서 데이터를 얻을 수 없을 때, 이것을 처리하는 절차를 제시한다. 제 3장과 제 4장에서는 각각 부분요인배치법을 이용하는 경우와 강건설계 실험에서 결측치가 생길 때에 좋은 실험점들을 파악하여 이들에 공통인 인자의 수준을 구별하고 주어진 데이터를 이용하여 근사적인 최적조건을 파악하는 방법을 예시한다. 결론과 토의는 제 5장에 기술한다.

2. 데이터 기반 분석 방법 제안

결측이 기록 오류 또는 실험장비 활용 또는 시간 부족이 아니라, 제품이 원하는 기본 형상이나 성능을 갖추지 못하여 생기는 경우 등 실험점 자체의 문제로 인하여 발생할 때 활용할 수 있는 방법을 모색하는 것이 필요하다. 예를 들어, Park *et al*.(2019)은 조영제용 나노입자를 합성하는 강건설계 실험 사례를 제시했는데, 18개의 실험점 중에서 4개의 실험점에서 입자가 형성되지 않는 결측치 문제를 언급하고 있다. 특히, 제품이나 공정 개발을 위한 실험 연구 초기의 선별실험을 위한 부분요인배치법을 쓰거나 강건설계를 위한 실험에서는 이러한 경우가 자주 발생할 수 있다.

Box *et al*.(2005)은 베어링 마모실험 사례를 분석하면서 일반적인 통계적 분석으로는 어떤 요인효과도 유의하지 않지만, 데이터 자체를 관찰하여 마모율이 작게 나타나는 두 개의 실험점을 파악하고 이들 실험점에 공통으로 나타나는 인자 수준을 파악하는 방법을 이용하여 최적 조건을 찾는 방법을 제시했다. 본 논문에서는 이러한 아이디어를 이용하여 실험연구자들이 결측치가 발생하는 상황에서 쉽게 활용할 수 있는 방법을 제시하고자 한다. 기본적인 절차는 다음과 같다 : 1) 결측치를 제외한 반응변수 값을 정규확률그림에 나타내고, 2) 이 중에서 다른 점들과 달리 좋은 반응 값을 갖는 실험점들을 구별하여, 이들에 공통으로 나타나는 인자 수준을 파악하며, 3) 이 수준조합을 최적조건으로 선정하거나 추후 반응변수를 개선하기 위한 방향을 정한다. 반응변수를 정규확률그림으로 나타내어 좋은 실험점을 찾는 방법은 Box *et al*.(2005)이 베어링 마모실험 사례에서 이용했다. 정규확률그림을 이용하면, 세로축의 50% 인근에 있는 데이터들의 직선적 경향에서 벗어난 것들을 묶음으로 구분할 수 있으므로 이 선에서 벗어난 것들을 알아내어 좋은 반응 값들을 파악할 수 있다.

3. 부분요인배치 실험의 결측 데이터 문제

자동차 도장 공정에서 페인트의 광택(glossiness)과 내마모성(abrasive resistance)을 높이기 위하여 8개의 인자를 대상으로 2_{IV}^{8-4} 부분요인배치 실험을 수행한 설계행렬과 실험데이터를 <Table 1>에 나타내었다(Box *et al*., 2005; p. 265).

이 데이터를 분석한 결과, 광택 y_1에 유의한 효과는 A와 B이고, 광택을 높이기 위한 조건은 A+B+이다. 내마모성 y_2 데이터를 분석하면, A, B, F 인자가 유의하며 내마모성을 가장 높이는 조건은 A−B−F+ 이다. 예시를 위하여 처음 2개의 실험조건에서 광택을 측정할 수 없을 정도로 페인트가 제대로 도포가 되지 않아서, 1번과 2번 데이터가 없다고 가정하자. 나머지 데이터를 정규확률그림에 나타내면, 4, 8, 12, 16번 실험점에서 y_1의 값이 다른 것들에 비해 뚜렷하게 크게 나오는 것을 알 수 있다(<Figure 1> 참조). 이들 실험점에서 공통으로 나타나는 인자 수준을 보면 A와 B만 '+' 수준이다. 다른 인자들의 경우에는 −와 + 수준이 섞여 있다. 이는 원 데이터를 분석한 결과와 동일하다. 내마모성 y_2는 전체의 절반인 7개가 정규확률 직선 근처에 몰려 있고, 나머지 7개의 실험점 중에 5, 7, 9, 13번의 내마모성 값이 비교적 크게 나타난다(<Figure 2> 참조). 이 실험점들에 공통으로 나타나는 것은 A가 '−' 수준이다. 이 결과는 원 데이터 분석 결과에서 나타난 B와 F가 영향을 미친다는 것은 반영하지 못하고 있다. 두 개의 반응변수별로 나타나는 차이는 광택 특성에 영향을 미치는 A와 B의 주효과는 다른 요인효과에 비하여 애초부터 크게 나타났고, 내마모성에 영향을 미치는 주효과 A, B, F는 다른 요인효과에 비해 월등하게 크지는 않은 것에 기인한다고 볼 수 있다.

Table 1. 2_{IV}^{8-4} Fractional Factorial Paint Data

No.	Factors								Responses	
	A	B	C	D	E	F	G	H	y_1	y_2
1	−	−	−	−	−	−	−	−	53	6.3
2	+	−	−	−	+	+	+	−	60	6.1
3	−	+	−	−	+	+	−	+	68	5.5
4	+	+	−	−	−	−	+	+	78	2.1
5	−	−	+	−	+	−	+	+	48	6.9
6	+	−	+	−	−	+	−	+	67	5.1
7	−	+	+	−	−	+	+	−	55	6.4
8	+	+	+	−	+	−	−	−	78	2.5
9	−	−	−	+	−	+	+	+	49	8.2
10	+	−	−	+	+	−	−	+	68	3.1
11	−	+	−	+	+	−	+	−	61	4.3
12	+	+	−	+	−	+	−	−	81	3.2
13	−	−	+	+	+	+	−	−	52	7.1
14	+	−	+	+	−	−	+	−	70	3.4
15	−	+	+	+	−	−	−	+	65	3.0
16	+	+	+	+	+	+	+	+	82	2.8

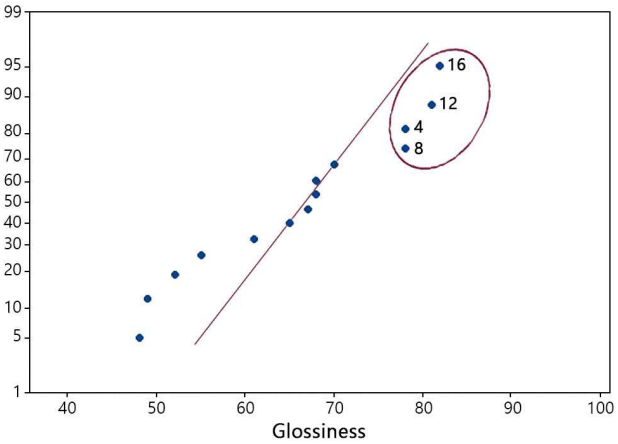

Figure 1. Normal Plot of Glossiness

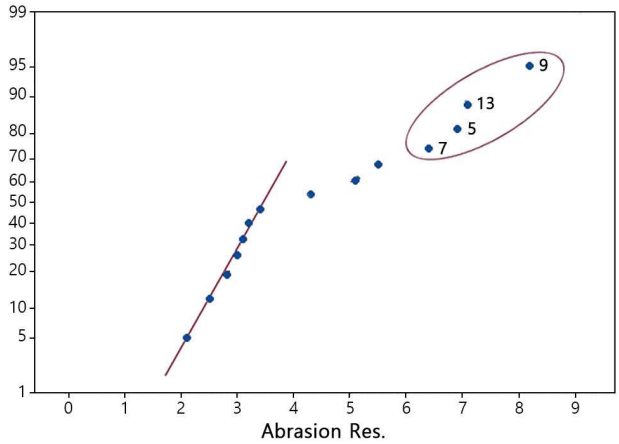

Figure 2. Normal Plot of Abrasion Resistance

Table 2. Factor Levels for the Better Design Points

No.	Factors								Glossiness
	A	B	C	D	E	F	G	H	
4	+	+	−	−	−	−	+	+	78
8	+	+	+	−	+	−	−	−	78
12	+	+	−	+	−	+	−	−	81
16	+	+	+	+	+	+	+	+	82

Table 3. Product Design Variables and Their Levels

Product Design Variables	−	+
Paper Supplier(PS)	UK	USA
Contents of Solids(SC)	28%	30%
Cylinder(CY)	Fine	Coarse
Oven Temperature(OT)	180℃	215℃
Line Speed(LS)	Low	High

4. 강건설계 실험의 결측치 문제

3장에서는 기존 실험데이터의 일부 실험점에서 결측치가 발생한다고 가정을 하고 분석을 했는데, 이번 장에서는 실제로 결측치가 나타난 문제를 보기로 한다. Box *et al.*(2005)은 의약품 패키지 강건성 실험에 관한 실험데이터에 결측치가 있는 사례를 제시하였다. 설계변수는 5개, 잡음변수는 봉인온도(sealing temperature; ST)와 머무름시간(dwell time; SD) 2개를 고려했다. 설계변수와 수준은 <Table 3>에 나타내었다. 5개의 설계변수를 대상으로 2_{III}^{5-2} 부분요인배치법을 이용하고, 이들 각각의 실험조건에서 잡음은 3×2 요인배치를 이용하여 반영하였다. <Table 4>에 실험의 설계행렬과 밀봉결함(defective seal) 수를 나타내었다(Box *et al.*, 2005, p. 540). 여기서 '*'로 표시된 것은 그 조건에서 데이터를 얻을 수 없다는 것을 나타낸다. 결측치가 나타나지 않는 6개의 설계변수 실험조합에서 구한 데이터를 이용하여 각 실험점별로 다음의 식 (1)을 이용하여 망소특성의 신호-잡음비(Signal-to-Noise ratio; SN비)를 구하였다.

$$SN_i = -10\log\left(\sum_{i=1}^{6} y_{ij}^2/6\right) \quad (1)$$

Table 4. Design Matrix and the Number of Defective Seals-Case 1

Product Design	Product Design Variables					Noise Variables							SN Ratio
	PS	SC	CY	OT	LS	ST	−	0	+	−	0	+	
						SD	−	−	−	+	+	+	
1	+	−	−	+	+		8	4	*	0	*	7	
2	+	−	−	+	−		*	*	*	*	*	*	
3	−	−	+	−	+		3	2	1	1	0	0	-3.98
4	−	−	+	−	−		7	1	1	2	0	4	-10.73
5	+	+	+	−	+		0	0	0	0	0	1	7.78
6	−	+	+	+	+		4	0	0	1	1	1	-5.01
7	−	+	−	+	−		9	1	0	4	5	1	-13.15
8	+	+	−	−	−		2	1	3	1	0	0	-3.98

'*' denotes missing data.

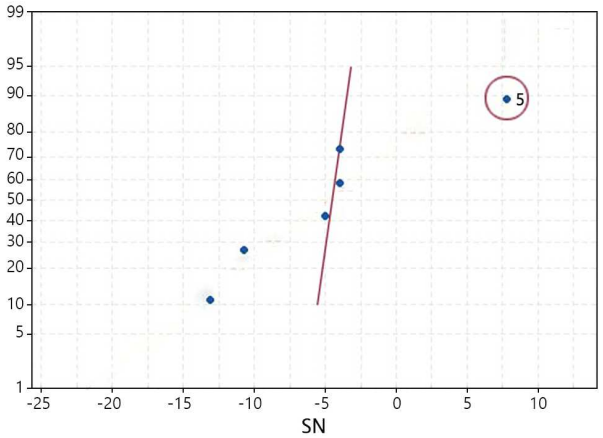

Figure 3. Normal Plot of Defective Seals(Case 1)

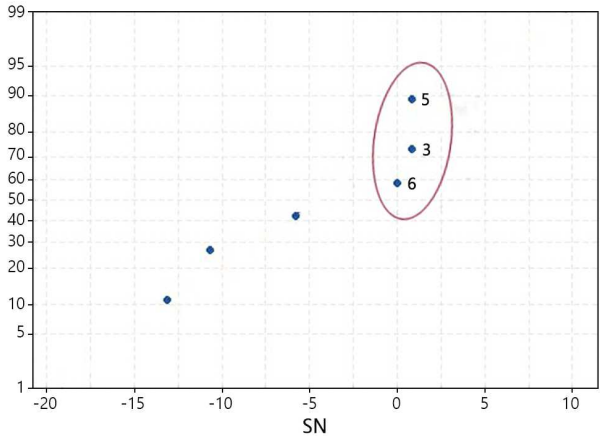

Figure 4. Normal Plot of Defective Seals(Case 2)

위에서 y_{ij}는 i번째 실험점의 j번째 밀봉결함 수이고, SN_i는 i번째 실험점의 SN비이다. <Figure 3>은 1번과 2번 실험점을 제외한 실험점들의 SN비를 정규확률그림으로 나타낸 것이다. 5번째 실험점의 SN비 값이 다른 것들에 비해 뚜렷하게 크게 나타났다. Box *et al.*(2005)은 비록 두 개의 설계변수 실험점에서 결측치가 나타났지만, 5번째 실험점의 설계변수 조합에서 밀봉결함이 거의 없으므로(PS, SC, CY, OT, LS) = (+, +, +, −, +)을 최적조건으로 삼으면 된다고 언급하고 있다.

본 논문에서 제시한 방법의 유용성을 예시하기 위하여, 원래 데이터를 수정하여 의약품 패키지 실험의 데이터가 <Table 5>와 같이 나타났다고 가정해 보자. <Figure 4>는 SN비를 구하여 정규확률그림에 나타낸 것이다. 3, 5, 6번째 실험점에서 SN비 값이 크게 나타났는데, 설계변수의 수준 측면에서 이 세 가지 실험점의 공통점은 (CY, LS) = (+, +)이다. 밀봉결함 수를 최소화하기 위해서는 실린더(CY)는 거친(coarse) 것을 선정하고, 생산 속도(LS)를 높이면(high) 된다. 다른 3개의 인자인 종이공급업자(PS), 고체 함량(SC), 오븐 온도(OT)는 경제성이나 편의성 등 다른 측면을 고려하여 정하면 되는데, 이것은 제품 개발자나 공정을 운영하는 엔지니어에게는 그만큼 인자의 수준 결정에 유연성을 부여하는 것이다. 물론 이러한 최적조건에서 잡음변수들의 수준을 다양하게 변화시켜 가면서 확인실험을 하여 밀봉결함 수가 적게 나오는지를 파악할 필요가 있다.

5. 결론 및 토의

본 논문에서는 결측치가 기록 오류나 장비 문제 또는 시간 부족으로 생긴 것이 아니고 개발하고자 하는 소재, 부품이 원하는 기본 형상이나 성능을 갖추지 못하여 발생하는 등 실험점의 실험조건에 문제가 있을 때, 결측치를 제외한 나머지 데이터 자체를 이용하여 실험 영역 내에서 최적 조건을 선정하거나 반응변수를 개선하기 위한 방향을 설정하기 위한 접근방법을 제안하였다. 특히 제품개발을 위한 변수 선별에 이용하는 부분요인배치법과 강건성 확보를 위한 강건설계 실험에 이 방법이 활용되는 경우를 예를 통하여 살펴보았다.

결측치는 실험을 통하여 제품이나 공정을 개발하는 현장에서 자주 발생한다. 본 논문에서 제안한 방법은 실험연구자나 공정의 운영자가 쉽게 이해하여 적용할 수 있을 것으로 기대한다. 본 논문에서 제안한 아이디어의 적합성에 기반하여 향후 결측치가 발생하는 다양한 상황에 실제로 적용하여 현장 활용성을 높이기 위한 개선된 방법이 나오기를 기대한다.

Table 5. Design Matrix and the Number of Defective Seals-Case 2

Product Design	Product Design Variables					Noise Variables							SN Ratio
	PS	SC	CY	OT	LS	ST	−	0	+	−	0	+	
						SD	−	−	−	+	+	+	
1	+	−	−	+	+		8	4	*	0	*	7	
2	+	−	−	+	−		*	*	*	*	*	*	
3	−	−	+	−	+		0	2	0	1	0	0	0.79
4	−	−	+	−	−		7	1	1	2	0	4	-10.73
5	+	+	+	−	+		0	0	0	2	0	1	0.79
6	−	+	+	+	+		2	0	0	1	0	1	0.00
7	−	+	−	+	−		9	1	0	4	5	1	-13.15
8	+	+	−	−	−		2	1	3	3	0	0	-5.84

참고문헌

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005), *Statistics for Experimenters : Design, Innovation, and Discovery*, 2nd Edition, Wiley, New York.

Goh, T. N. (1997), Use of Dummy Values in Analyzing Incomplete Experimental Design Data, *Quality Engineering*, **10**(2), 397-401.

Montgomery, D. C. (2017), *Design and Analysis of Experiments*, 9th Edition, Wiley, New York.

Park, J., Kim, E., Kwon, E., Jeon, B.-S., Myeong, W., and Byun, J.-H. (2019), A Case Study on Robust Design of Nanoparticles Synthesis for Contrast Agent, *Journal of Korean Institute of Industrial Engineers*, **45**(5), 465-474.

Vining, G. G. (2013), Technical Advice : Scientific Method and Approaches for Collecting Data, *Quality Engineering*, **25**(2), 194-201.

저자소개

변재현 : 서울대학교에서 산업공학 학사, KAIST에서 산업공학 석사, 박사 학위를 취득하였다. 현재 경상대학교 산업시스템공학부 교수이며, 연구분야는 실험계획법, 품질경영, 품질공학이다.