總 說

화학공학에 있어서의 통계학의 활용

김 영 걸

한국과학기술연구소(고분자 연구실) 초빙교수

Applications of Statistical Methods in Chemical Engineering

Young Gul Kim*

Department of Chemical Engineering Northwestern University Evanston, Illinois, U. S. A.

Abstract

A number of statistical methods that are deemed particularly useful in chemical engineering are briefly illustrated. The topics covered are: sampling plan and statistical tests, response surface method, variable screening, and model discrimination. None of the topics is covered in any detail. The objective here is to acquaint the practising chemical engineers with what statistics can do for them and thus stimulate their interest, with the hope that they will explore these topics in depth on their own.

오늘의 강의는 비유컨데 요리강습의 선전과 같다. 수 많은 요리 중에서 몇가지만 추려서 관중에게 맛을 보게 하여 흥미를 일으켜 그 요리를 어떻게 만드는지 배우고싶게 하려는 것이 목적인 것과 같이 이 강의를 통하여 통계학이라는 범위가 넓은 학문분야에서 우리 화공하는 사람들에게 흥미가 있을만한, 몇가지 제목을 간단히 소개하여 거기서 나오는 여러가지 방법을 배우고 활용할 동기를 주고저 하는 바이다.

(I) 저자가 몇달 전 미국의 어떤 석유학공장의 polyethylene 제조 과정에서 일어난 문제 때문에 요청이 와서 몇번 consulting 한 일이 있었다. 제조된 polyethylene의 추후가공을 위하여 erucamide 라는 additive를 가하여야 하는데 이 additive의 농도가 가공업자의 용도에 따라 200 ppm 에서 500 ppm 정도까지 되어야 한다는 것이었다. 이 additive를 섞는 과정은 Fig. 1과 같다.

저자가 이 공장에 처음 consulting 갔을때는 이미 많은 data가 실험에 의하여 수집되어 있어 그것을 어떻게 해석하여 정당한 결론을 내릴 것인가 하는 것이 논의되고 있는 중이었다. 통계학의 기초 지식이 부족한 사람들이 sampling에 필요한 요소를 충분히 고려하지 않았고 data를 몰았기 때문에 불필요한 data가 많았고, 꼭 있어야 할 data가 없어서 결국은 실험을 처음부터

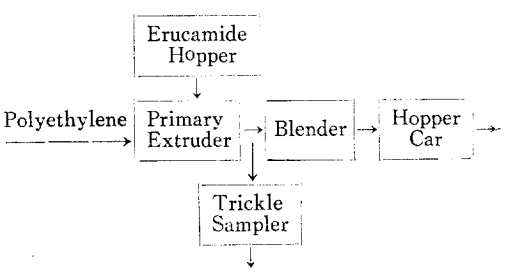

Fig. 1. Additive Blending Process

*At present with Korean Institute of Science and Technology.

다시하여야 할 수밖에 없었다.

예를들면 sample을 prinary extruder, blender, trickle sampler 등 여러 곳에서 취하였으나 분석 방법에 따라서 결과가 달랐고 또 분석하는 사람에 따라 결과에 큰 차이가 있었다. 이러한 때 statistical test를 사용하기 위하여 experimental error variance가 필요한데 같은 sample을 같은 사람이 같은 분석법으로 취급한 반복된 실험치가 없어 이 variance를 estimate 할 도리가 없었다. 즉 replicate가 전연 없는 실험을 한 것이었다.

이 문제의 핵심은 Blender에서 나오는 제품을 hoppercar에 저장하여 놓고 그것의 additive 함량을 확인하는 것이다. 예를들어 어떤 업자가 additive 함량 500±50 ppm 들어 있는 polyethylene을 요구할 때 hopper car에서 sample을 하나만 취하여 분석하여 본 결과 430 ppm이 나왔다고 하자. 물론 sample 하나만은 크게 믿을 수 없는 것이니 이 결과만으로는 그 hopper car의 전량을 불합격이라고 물리칠 수는 없을 것이다.

erucamide의 분포가 불균일(inhomogeneous)하며 화학분석에 실험오차가 따르는 법이니까 개개의 sample은 어떠한 확률분포(probability distribution)을 가지고 있을 것이다. 만약 이 sample의 distribution이 Fig. 2A와 같다고 하면 몇개의 sample을 취하여 그 평균치를 취하면 이 평균치(sample average)의 probability distribution은 Fig. 2B와 같아진다. 즉 분포가 좁아진다. sample의 수가 많을수록 sample average의 분포가 좁아지고 이 sample average를 사용하여 내린 결론의 신뢰성이 높아지게 된다.

그러나 이와같은 신뢰도가 높은 결론을 얻을려면 그만한 대가를 지불하여야 한다. 즉 시간과 경비를 드려 실험을 많이 할수록 더 믿을만한 결론이 나오게 된다. 그러므로 경비(cost)와 혜택(benefit)을 절충하여 sampling plan을 세워야 한다. 이와같은 예에서는 결론의 신뢰성이 직접적으로 공장의 수지면에 반영이 된다. 만약 적은 수의 sample에 의하여 hopper car의 내용물에 관한 판정을 내릴 때 다음과 같은 두 가지의 risk가 항상 동반하는 법이다.

1) 규격에 닷는 제품을 불합격이라 하여 기각하는 risk.

2) 규격에 닷지 않는 제품을 합격이라고 채택하는 risk.

첫번 risk는 좋은 제품을 적하하여 염가로 팔게되어 손해를 보는 경우이고 둘째번 risk는 불합격 상품을 업자에게 팔아 소송문제가 일어나고 손해배상을 지불하여야 할 경우이다. 이러한 두 가지의 risk를 전연 없게 하는 수는 없으니 경영자의 입장에서 경제적 최적화에 의하여 risk의 한계를 정하여야 한다. 위의 예에서 이 risk를 α%, β%라고 각각 정한다면 그것에 따라 sample의 수가 결정된다.

이와같은 것은 간단한 sampling theory로 해결할 수 있는 통계문제중의 하나이고 품질관리(quality control)의 중요한 도구의 하나라고 하겠다.

(Ⅱ) 1971년에 저자가 교편을 잡고 있는 미국 Northwestern 대학의 화공과에 초빙교수로 Byron Riegel 박사가 와 계셨다. 이분은 본래 유기화학자이고 미국

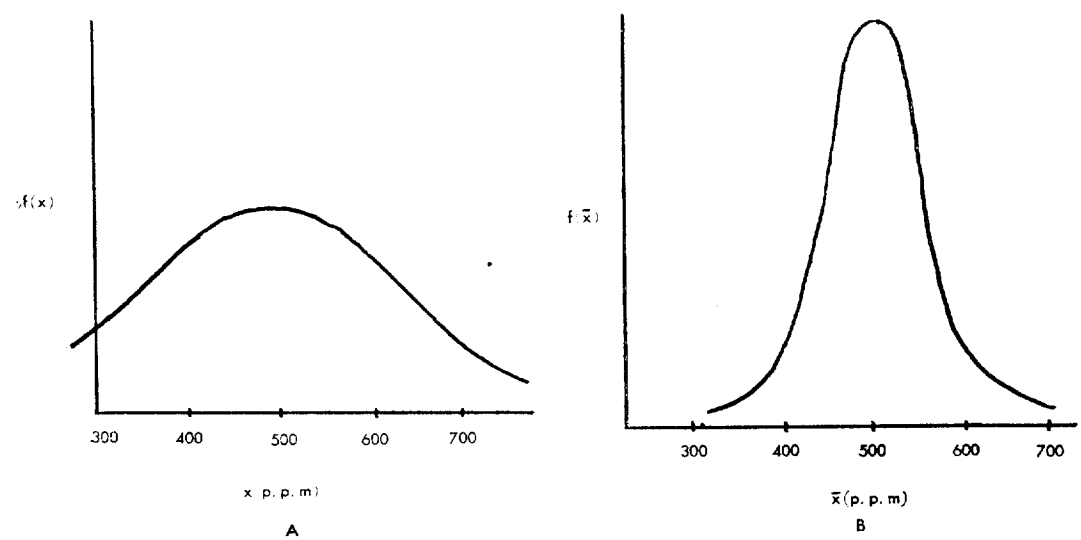

Fig. 2 Sample, x와 sample average, x̄의 probability distribution

의 유수한 제약회사 G. D. Searle Co.의 연구소장으로 오래 일하셨고, 1970년에는 미국화학회(American Chemical Society)의 회장을 역임한 저명한 과학자이다. Riegel 박사와 제약연구 개발 공업화에 관해 여러번 의논할 기회가 있었는데 그분의 경험중에서 제약공정개발에 제일 효율적인 방법으로 생각되는 것이 Response Surface Method 라는 것이었다.

이 방법은 G. E. P. Box 박사가 영국 ICI (Imperial Chemical Industries)에서 통계부를 담당하고 있을 때 개발한 것인데 요점을 추려보면 다음과 같다.

어떤 약품 A를 만드는 과정에서 독립변수로 $x_1, x_2, \cdots\cdots, x_n$이 있다고 하자. 예를들면 x_1은 반응온도, x_2는 pH x_3는 압력, x_4는 반응시간, $x_5, x_6, \cdots\cdots x_n$은 여러 반응물질의 농도일 수 있다. 이것을 수학식으로 표시하자면,

$$y=f(x_1, x_2, \cdots\cdots x_n)$$

(여기서 y는 종속변수인 A의 수익(yield)라고 볼 수 있다).

이 $x_1, x_2 \cdots\cdots x_n$와 y 사이의 함수관계, 즉 mathematical model이 알려져 있으면 그것을 사용하여 최적화(optimization)이 가능하며 그 결과로 여러가지 독립변수 값을 어떻게 선정하여야 수익의 최고치를 얻을 수 있을지 알 수 있을 것이다. 그러나 이러한 mathematical model이 없고 또 시간과 경비가 많이 들거나 그 함수관계가 복잡하여 완전한 model을 얻을 수 없을 경우가 자주 생긴다. 이러한 때에는 실험을 통하여 최적화를 하는 수밖에 없는데 그 방법의 하나가 바로 Response Surface Method 이다.

이 방법을 간단히 설명하자면 다음과 같다. 독립변수 x_1, x_2와 종속변수 y 사이에 어떠한 미지의 함수관계가 있고, 그것을 Fig. 3와 같은 contour map로 표시

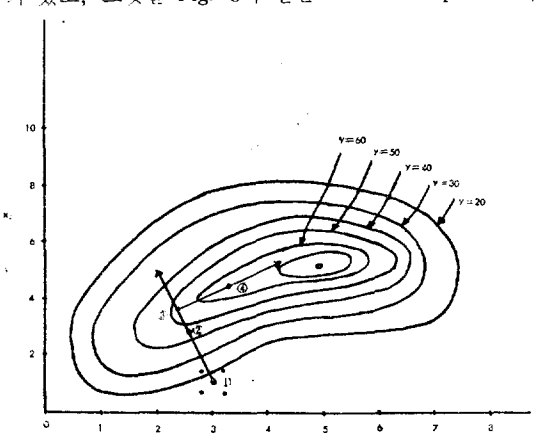

Fig. 3. Contour surface

할 수 있다고 하자. 물론 이와같은 contour map(혹은 contour surface)의 형체는 실험을 하기 전에는 알려지지 않고 있다.

현재의 starting point(x_{10}, x_{20})에서 어떠한 방법으로 제일 적은 노력을 드려 maximum point$(x_{1,\max}, x_{2,\max})$를 찾느냐 하는 것이 문제가 된다. Response Surface Method에 의하면 시발점(starting point)에서 실험에 의하여 contour surface의 gradient를 구하여 점차적으로 maximum point에 접근하게 된다. 구체적인 방법으로는 Fig. 3에서 시발점 ①을 중심으로 하여 two-level factorial design에 따라서 실험을 하고 그 결과로 gradient를 구한다. 이 gradient가 시발점 근처에서 y의 증가율이 제일 큰 방향을 표시한다. 수치예로서 제 1표에 있는 실험결과를 얻었다고 하자.

Table 1. Factorial experiment 결과

x_1	x_2	y
2.8	0.6	11.0
3.2	0.6	8.5
2.8	1.4	23.0
3.2	1.4	20.0

gradient를 계산하면 시발점 $(x_{10}=3.0,\ x_{20}=1.0)$근처에서 y가 제일 많이 증가하는 방향은 Fig. 3 ①에서 화살로 표시된 방향이다. 이러한 실험을 반복하여 ②, ③, ④로 진행하여 최고점에 접근한 다음 3 level factorial design이라는 실험계획을 사용하여 실험을 하고 그 결과로 二차식(second-order equation)을 구하여 그 식을 편미분하여 maximum을 얻을 수 있다.

설명을 쉽게 하기 위하여 독립변수 두개만 있는 간단한 예를 들었으나 이 방법의 참된 장점은 문제의 차원수(dimensionality)가 클수록 더 현저하여 진다.

(Ⅲ) 응용통계의 셋째 예로서 변수심사(variable screening)을 들 수가 있다. 이 방법은 특히 복잡한 공업적 과정에서 문제를 규정하는 단계에 많이 쓰일 수 있는 방법이다. 일의 순서로서 보면 이것이 Response Surface Method 보다 앞서 가야할 단계이다.

저자가 몇 해 전에 지도한 연구문제 가운데 Agglomeration process에 관한 연구가 있었다. 분말(fine powder)를 수증기로 처리하여 적은 덩어리로 엉키게 하여 미립자(particles)로 만드는 과정이 식품공학에서 가끔 일어나는 문제인데 이 과정을 연구하는 제一단계로 미립자(particle)의 size distribution, 혹은 평균크기(mean size)에 영향을 미치는 독립변수가 어떠한 것들인지 알아야 하였었다. 과거의 경험이라든가 유사한 과정에서 생기는 일로 미루어 보아 steam pressure,

moisture content(분말의) 등등 일곱개의 독립변수를 고려의 대상으로 할 수 있었다. 즉,

$$y=f(x_1, x_2, \cdots\cdots x_7)$$

물론 일곱 개가 모두 중요한 변수라는 것은 아니다. 그러나 실험을 하지 않고는 어떤 것이 중요하고 어떤 것이 무시할 수 있는지 알 수 없었다. 이것을 알기 위하여 experimental design 방법 중에서 factorial design을 사용하여 실험을 하여 보기로 하였다. factorial design 중에서 제일 간단한 2-level factorial design인데 이것을 쓸 때 일곱 독립변수 $x_1, x_2, \cdots\cdots x_7$의 값을 상(+), 하(−) 두 가지로 정하여 그 combination을 모두 실험에 사용하게 된다. 설명을 간편하게 하기 위하여 독립변수가 두개만이라고 가정할 때 예를들어 steam pressure를 5 psi(−), 10 psi(+)로 그리고 steam velocity를 2 ft/sec(−), 3 ft/sec(+)로 각각 정하면 다음과 같은 4개의 combination이 생긴다(Table 2).

Table 2. 2-Level factorial design

Steam Pressure (psi)	Steam Velocity (ft/sec)
5 (−)	2 (−)
10 (+)	2 (−)
5 (−)	3 (+)
10 (+)	3 (+)

변수가 2개일 때 2-level factorial design에 의한 실험수는 $2^2=4$, 즉 실험을 최소로 네번하여야 하며 변수가 7개 있는 문제에서는 $2^7=128$번의 실험이 필요하다는 결론이다. 이렇게 많은 실험은 실제로 하기가 곤란하며 또 필요 이상의 information을 얻는 결과가 생김으로 변수의 수가 많을 때에는 complete 2-level factorial design의 일부만을 사용하는 것이 보통이며, 이것을 fractional factorial design 이라고 한다. 7개의 독립변수 사이에 interaction의 가능성이 얼마나 많은가에 따라서 2^{7-r}로 실험수를 주릴 수 있는데 $r=1$이면 64개의 실험으로, 또 $r=3$이면 16개의 실험으로 variable screening을 하게 되는 것이며 물론 실험수가 적을수록 실험에서 얻을 수 있는 information이 줄어드는 것이다.

실험결과를 가지고 7개의 독립변수(통계용어로 factor 라고 함)가 종속변수(response 라고 함)에 미치는 영향을 main effect, interaction effect로 나누어 그 수치를 평가하며 결론을 내릴 수 있게 하는 방법을 analysis of variance라고 한다.

결론으로는 여러가지 main effect, interaction effect 중에서 어느 것이 중요하며 어느 것이 무시할 수 있는지 알려지게 되며, 여기서 중요하다고 알려진 변수만을 대상으로 하여 더 자세한 연구 program을 세울 수 있게 된다.

(Ⅳ) 통계학 용도의 네번째 예로 model discrimination을 들어 보기로 하자. 제일 간단한 예로 독립변수 x와 종속변수 y 사이의 관계를 조사하기 위하여 실험을 하고 그 결과를 $y=a+bx$라는 mathematical model로 표현하여 본다고 하자. 최소자승법(method of least squares, 또는 regression analysis)에 의하여 실험결과를 처리하면 물론 a와 b를 구할 수 있다. 그렇다고 해서 반드시 x와 y의 관계가 직선관계라고 결론을 내릴 수는 없다. 최소자승법은 만약 실험 결과가 직선 관계로 표시될 수 있을 경우에 그 직선의 방정식을 우리에게 줄 수 있으나 그 관계가 과연 직선인지 아닌지는 별도로 확인을 하여야 한다.

극단의 예를 들자면 Fig. 4와 같은 실험결과를 가지고 최소자승법을 적용시켜 직선을 구하면 직선의 방정식을 얻을 수는 있으며, 그 계수 a와 b도 계산할 수 있다.

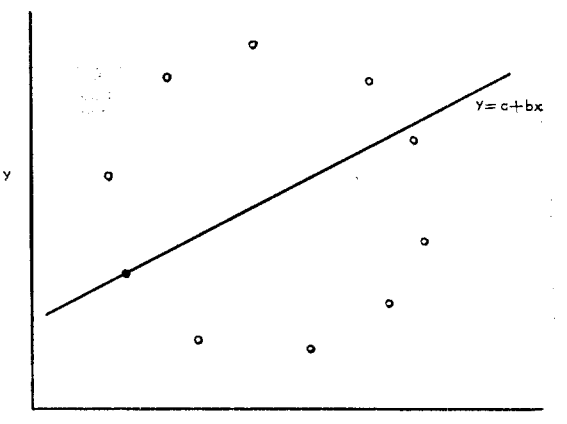

Fig. 4. Straight line drawn through data—poor fit

물론 x와 y의 관계가 직선이 아니라는 것은 graph를 보아 명백한 사실이며 여기에 직선을 적용시키려는 사람은 없을 것이다. 그러나 이보다 덜 뚜렷한 경우에는 관찰만으로는 직선관계 여부를 판단하기 어려울 때가 많다.

Fig5.에 표시된 실험결과를 가지고 x와 y 사이의 방정식을 구할려 할 때 우선 직선을 적용시켜 보면 $y=a+bx$ 라는 一차식이 나온다. 다음의 단계로는 이 식이 과연 적합한지 결정을 하여야 한다.

이것을 위하여 residual sum of squares(SSR)라는 양을 다음과 같은 공식에 의하여 계산한다.

$$SSR=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$=\sum_{i=1}^{n}\{y_i-(a+bx_i)\}^2$$

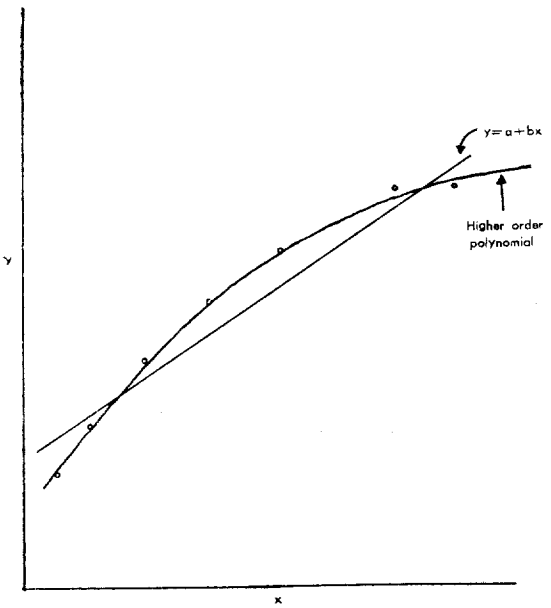

Fig. 5. Data for linear regression

이 SSR는 一차식을 가지고 추정(estimate)한 y 값($\hat{y}_i$)과 실험에서 얻은 y값(y_i)의 차이의 측도(measure)이며, 이것이 클수록 이 一차식이 부적당하다는 말이 된다. 물론 SSR는 실험치의 수가 많을수록 커지는 양이니 그대로 사용할 수 없고 이것을 normalize 하여야 한다. 이 normalization factor로 $(n-2)$를 쓰는데 n은 실험치의 총수이고 "2"는 一차식에 parameter가 둘(즉 a와 b로 추정되는 두 parameter)이라는 데에서 생긴다. 이 normalization factor를 통계용어로 자유도(degree of freedom)이라고 한다. 이와같이 얻은 양 $(SSR)/(n-2)$이 크면 一차식을 부적당하다하여 기각하게 된다.

물론 크고 적고 하는 것은 상대적인 개념이고 이 $(SSR)/(n-2)$가 얼마나 커야 크다고 볼 수 있는지 객관적으로 결정할 수 있어야 한다. 간단히 말하자면 $(SSR)/(n-2)$를 실험오차 variance의 estimate와 비교하여 보아야 한다. 즉 추정한 y 값과 실험치 y값과의 차이가 실험오차 범위 안에 들어오는지 오차보다 훨신 큰지를 계산하여 판단을 하는 것이다.

구체적으로 이것을 하는 방법은 $(SSR)/(n-2)$와 $\hat{\sigma}^2$(오차 variance의 추정양)의 비율을 계산하여F-Distribution 표를 사용하여 一차식의 채택 여부를 결정한다 F-Distribution의 사용은 실험치가 모두 독립우연변수(independent random variable)이고 실험오차가 정규분포(normal distribution)을 가졌다는 가정하에서 가능한 것이다.

만약 이와같은 해석(analysis)의 결과로 一차식이 부적당하다고 결정이 되었다고 하면 二차식, 三차식 등의 다항식(polynomial)을 실험결과에 마추어(fit) 보게 된다.

$$y=\beta_0+\beta_1 x+\beta_2 x^2+\cdots\cdots\beta_p x^p$$

이 polynomial을 사용할 때 우리가 상식적으로 알 수 있는 것은 p가 클수록 실험치와 polynomial의 추정치가 가까워진다는 것이다. 이것을 극단에 이르게 하면 실험치의 수와 parameter의 수가 같으면 실험치와 polynomial의 추정치가 완전히 일치하게 된다.

유명한 수학자 Gauss가 말하기를 자기에게 6개의 adjustable parameter를 주면 코끼리를 그릴 수 있고(즉 코끼리의 형체를 묘시하는 방정식을 구할 수 있고), parameter를 하나만 더 주면 그 코끼리로 하여금 꼬리를 흔들게 할 수 있다고 한것 같이 다항식의 항수가 커지면 어떤 data라도 fit 할 수 있는 것이다.

그러므로 최고차 polynomial이 제일 좋다고 할 수는 없다. 예를들어 二차식과 三차식을 비교할 때 피상적으로는 三차식이 더 좋은 fit를 준다고 볼 수 있겠으나 과연 이것이 실질적으로 그런지 물어볼 필요가 있다. 이러한 질문에 대답을 구하기 위하여 쓰는 방법은 그 원리가 조금전에 다루어 본 一차식여부 결정법과 비슷하다. 우선 polynomial의 추정치와 실험치의 차로 residual sum of squares(SSR)를 다음과 같이 구한다.

$$SSR\ (2)=\sum_{i=1}^{n}\{y_i-(b_0+b_1x_i+b_2x_i^2)\}^2$$

$$SSR\ (3)=\sum_{i=1}^{n}\{y_i-(b_0'+b_1'x_i+b_2'x_i^2+b_3'x_i^3)\}^3$$

b_0, b_1, b_2는 실험치에 二차 polynomial을 fit 하여 얻은 parameter의 estimate이고 b_0', b_1', b_2', b_3'는 三차 polynomial을 fit하여 얻은 estimate이다. 실험 design matrix가 orthogonal이 아닌 경우에는 b_0, b_1, b_2와 b_0', b_1', b_2'는 다르다.

$[SSR(2)-SSR(3)]$는 三차식을 쓸때 二차식보다 얼마나 더 좋은 fit를 얻을 수 있는가를 나타내는 측도이고 이 양이 클수록 三차식을 쓰는 것이 합당하다는 것이다. 전의 예와 마찬가지로 여기서 크기 적기는 역시 상대적인 개념이고, 비교의 대상은 실험 오차 variance의 estimate가 쓰이게 된다. 다시 말하여 三차식을 씀으로 二차식 쓸 때에 비하여 오는 improvement가 실험 오차에 비하여 클 때 실질적인 improvement라고 인정한다는 것이다.

지금까지 다룬 model discrimination은 비교적 간단한 문제이며 linear regression analysis와 analysis of variance를 결합한 방법이라고 볼 수 있다. 이 방법을

쓸 때 실험은 이미 끝나 data가 모두 수집되어 있으며 이 data를 어떻게 처리하느냐 하는 과제가 남아 있는 것이다.

이와 반면에 어떤 복잡한 process가 있어 그것의 mechanism 혹은 model을 알고저 하는데 model이 nonlinear이며 가능한 model의 수가 많고 실험에 경비와 노력이 크게 드는 경우가 생긴다. 이런 문제를 당면할 때는 실험 하나 하나를 잘 계획하여 불필요한 실험은 하지 않고 실험 data에서 얻을 수 있는 최대한의 information을 짜내도록 하여야 한다.

이에 관하여 간단히 설명하자면 첫째로 불필요한 실험은 다음과 같은 경우에 생긴다. A=B=C와 같은 연속화학반응의 mechanism으로

(1) $A \longrightarrow B \longrightarrow C$

(2) $A \longrightarrow B \rightleftharpoons C$

의 두개가 있을 때 중간체 B의 농도와 반응시간 사이의 관계는 Fig. 6과 같다.

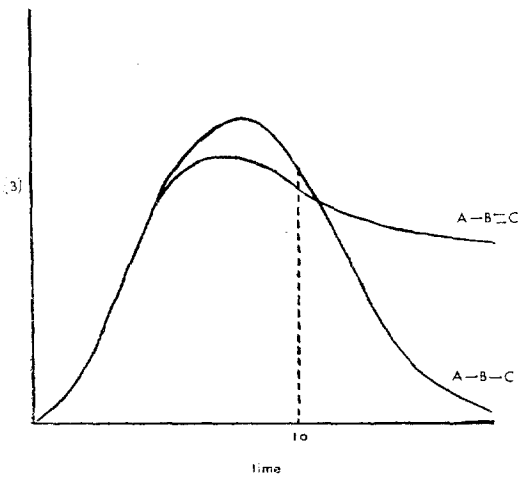

Fig. 6. Two models for a reaction involving chemical species A, B and C

이 두 mechanism 중 어떤 것이 맞는 것인가를 실험으로 결정하려면 반응시간을 t_0 이상 주어야 하며 t_0 이하의 반응시간을 써서 실험하면 두 mechanism을 구별할 수 없다. 이와같이 독립변수가 하나뿐이고 model이 간단한 경우에는 model discrimination을 위한 experimental design이 간단하지만 model 수가 많고 독립변수가 많은 복잡한 nonlinear model을 취급할 때는 experimental design이 매우 중요한 것이다.

저자가 지난 몇해동안 하여본 일중에 iron oxide를 촉매로 사용하는 water-gas shift reaction, 의

$$H_2O+CO \longrightarrow H_2+CO_2$$

kinetic model discrimination에 관한 연구 문제가

Table 3. Reaction Models

Langmuir-Hinshelwood Model

$$r=-\frac{kK_{CO}K_{H2O}[(CO)(H_2O)-(CO_2)(H_2)/K]}{[1+K_{CO}(CO)+K_{H2O}(H_2O)+K_{CO2}(CO_2)+K_{H2}(H_2)]^2}$$

Eley-Rideal Model

$$r=-\frac{kK_{H2O}[(CO)(H_2O)-(CO_2)(H_2)/K]}{1+K_{H2O}(H_2O)+K_{CO2}(CO_2)+K_{H2}(H_2)}$$

Oxidation-Reduction Model

$$r=-\frac{k_1k_2[(CO)(H_2O)-(CO_2)(H_2)/K]}{k_1(CO)+k_2(H_2O)+k_{-1}(CO_2)+k_{-2}(H_2)}$$

$$r=-\frac{k_1K'[(CO)(H_2O)-(CO_2)(H_2)/K]}{K'(H_2O)+(CO_2)}$$

Hulburt-Vasan Model

$$r=-\frac{k(H_2O)}{1+K(H_2O)/(H_2)}$$

Kodama Model

$$\frac{k[(CO)(H_2O)-(CO_2)(H_2)/K]}{1+K_1(CO)+K_2(H_2O)+K_3(CO_2)+K_4(H_2)}$$

Empirical Model

$$r=ak(CO)^m(H_2O)^n(CO_2)^p(H_2)^q$$

있었다. 1971년 현재로 여러 학자들이 제창한 model이 21개가 있었는데 그것을 크게 나누어 Table 3과 같이 6개로 분류할 수 있었다.

이와 같은 경우에 실험을 다하여 놓고 data를 해석하여 model discrimination하는 일은 매우 위험하고 비능율적인 것이다. 우리가 채택한 것은 G. E. P. Box와 그 제자들이 개발한 방법인데 넓게 보아 Fig. 7와 같이 3부로 나눌 수 있다.

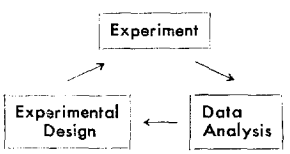

Fig. 7. Iterative Process for model discrimination

즉 실험, 해석, 실험계획이 iterative process가 되는 것이다. 실험을 하나식 하고 그 결과와 그 실험전에 모아놓은 실험결과를 모두 analysis와 design 단계에 사용하여 information을 update하고 점진적으로 옳은 model에 converge 하게 되는 방법이다. 이 방법의 사용을 다음의 간단한 예를 가지고 설명하기로 한다.

$A \longrightarrow B$와 같은 반응에 있어서 차수에 따라 Table 4에 열거한 4개의 model이 고려의 대상이 된다고 하자. 여기서 y는 A의 농도, t는 반응시간, T는 온도 그리고 θ_{i1}과 θ_{i2}는 model i의 parameter (i=1, 2, 3, 4)이다.

Box의 방법에 의하여 model discrimination을 할 때 시작은 factorial design에 따라 광범위의 t와 T를 골

Table 4. Kinetic models

Model 1.	$y=\exp\left\{-\theta_{11}t\ \exp\left(-\frac{\theta_{12}}{T}\right)\right\}$
Model 2.	$y=\frac{1}{1+\theta_{21}t\ \exp\left(-\frac{\theta_{22}}{T}\right)}$
Model 3.	$y=\frac{1}{1+2\theta_{31}t\ \exp\left(-\frac{\theta_{32}}{T}\right)}$
Model 4.	$y=\frac{1}{1+3\theta_{41}t\ \exp\left(-\frac{\theta_{42}}{T}\right)}$

라서 실험을 한다. Fig. 8에 ①, ②, ③, ④로 표시된 것이 시작할 때의 실험 design이다. 이 네개의 실험치를 갖이고 nonlinear least squares 방법으로 $\hat{\theta}_{i1}$과 $\hat{\theta}_{i2}$(θ의 추정치)를 구하고 그것을 사용하여 여러가지 t와 T의 combination에 해당하는 $\hat{y}$ (y의 추정치)를 계산할 수 있고 experimental design 단계에 들어가 $\hat{y}$가 제일 차이가 많이 나는 t와 T를 선택하여 실험을 하나 더 한다.

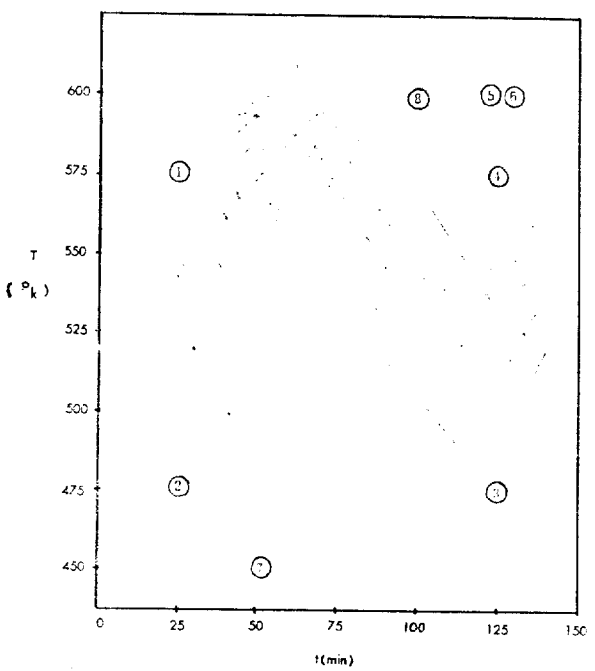

Fig. 8. Experimental design points

이렇게 골른점이 Fig. 8의 ⑤로 표시된 점이다. 이 점에서 실험을 하나 더 하면 실험치가 다 합하여 5개가 되여 이 5개의 data를 모두 사용하여 θ_{i1}과 θ_{i2}를 다시 추정한다. 이 재추정한 θ 값들은 종전것보다 더 정확하다. 이와같이 update 된 parameter estimate를 써서 design을 하면 다음 실험을 어떻게 하여야(즉 t와 T를 어떻게 정하여야) model discrimination이 제일 효과적인지 계산이 된다. 이것이 Fig 8에 ⑥으로 표시된 점이다. 이와같은 iteration을 계속하면 여러 model 중 제일 좋은 model로 converge 할 수 있다.

이 iterative process를 종결시키는 방법으로는 analysis 단계를 지날때마다 Bayes' Theorem을 사용하여 각 model의 확율(posterior probability)를 계산하여 한 model의 확율이 일정한 수치, 예를들면 95%, 에 도달하면 discrimination이 끝난 것으로 인정한다. Fig. 8에 이같은 iteration이 8개의 data를 얻을 때까지 계속된 결과가 표시되어 있다. iteration을 할 때마다 계산한 posterior probability는 Table 5와 같다. 처음에 factorial design에 따라서 4개의 실험을 한 후 확률을 계산하고 그 후는 실험치가 늘때마다 확률 계산을 한 것이다. 실험을 8개 한 후 둘째 model의 확률이 99%

Table 5. Posterior Probabilities

n	t	T	y	P_1	P_2	P_3	P_4
1	25	575	0.3961				
2	25	475	0.7232				
3	125	475	0.4215				
4	125	575	0.1297	0.0069	0.4290	0.5008	0.0633
5	125	600	0.0984	0.0019	0.5602	0.4291	0.0088
6	125	600	0.0556	0.0018	0.8639	0.1339	0.0004
7	50	450	0.7969	0.0021	0.9736	0.0243	0.0000
8	100	600	0.325	0.0032	0.9956	0.0012	0.0000

n은 실험순차, P_1, P_2, P_3, P_4는 model 1, 2, 3, 4의 posterior probability이다.

이상이 되어 discrimination이 끝난 것으로 보는 것이다.

이상 간단히 우리 화공도가 유익하게 활용할 수 있는 통계방법 몇가지를 말하여 보았다. 시간이 허락하지 않어 자세한 말을 할 수 없는 것이 유감이다. 마지막으로 앞으로 통계방법에 관심을 갖이고 좀더 깊이 공부하고저 하는 분들을 위하여 다음의 두책을 추천하고저 한다.

1. Davies and Goldsmith: Statistical Methods in Research and Production
2. Davies: Design and Analysis of Industrial Experiments.

이 두 책은 화학공업계에서 오래 종사한 통계학자와 화공학자들이 실무에서 얻은 경험을 가지고 실무자들을 위하여 쓴 책이니 우리 화공학도들에게 특히 도움이 많이 될 책들이다.

(이 강의에 쓴 Fig. 와 Table 중 Table 4, 5와 Fig. 7, 8은 다음의 논문에서 얻은 것이다.

Box and Hill: "Discrimination among Mechanistic Models" Technometrics 9, 57, (1967)