

Kim Sol

# Error Analysis for Experimental Physicists

THE 1<sup>st</sup> EDITION

*Seoul National University Department of Physics*



# Contents

<b>Contents</b>	<b>i</b>
<b>1 오차 해석학</b>	<b>1</b>
1.1 실험 통계학의 필요성 . . . . .	1
1.2 실험 통계학의 범위 . . . . .	2
1.3 실험 통계학의 목적 . . . . .	3
<b>2 오차와 불확실도</b>	<b>5</b>
2.1 오차의 정의 . . . . .	5
2.2 불확실도의 정의 . . . . .	6
2.3 가중치 . . . . .	8
2.4 최적치 . . . . .	9
<b>3 유의 수준</b>	<b>11</b>
3.1 유의수준과 신뢰구간 . . . . .	11
3.2 유의 수준의 $\sigma$ 표기 . . . . .	12
<b>4 실험 검정</b>	<b>17</b>
4.1 아르키메데스의 왕관 . . . . .	17
4.2 p-value를 이용한 유의성 검정 . . . . .	18
4.3 Rejection of Data . . . . .	22
<b>5 불확실도의 원인</b>	<b>25</b>
5.1 정밀도와 정확도 . . . . .	25
5.2 불확실도의 종류 - 반복 측정에 의한 제거 가능성 . . . . .	26
5.3 불확실도의 종류 - 발생 원인에 따른 분류 . . . . .	27
5.4 불확실도에 대처하는 방법 . . . . .	31
<b>6 불확실도의 검출</b>	<b>35</b>
6.1 기기 눈금의 보간 (Interpolation) . . . . .	35
6.2 변인 통제 (Control of Variables) . . . . .	41
6.3 반복측정을 통한 불확실도의 검출 : t-분포와 선형 회귀 . . . . .	45
6.4 중심 극한 정리와 Student t-분포, 반복 측정으로부터 얻어지는 우연오차 . . . . .	46
<b>7 회귀</b>	<b>49</b>
7.1 회귀 분석 . . . . .	49
7.2 최소 자승법 . . . . .	49
7.3 각종 회귀 모델 . . . . .	54
7.3.1 선형 회귀 . . . . .	54
7.3.2 극값을 가지는 분포의 회귀 . . . . .	60

<b>8 불확실도의 전파</b>	<b>63</b>
8.1 다변수함수 관계를 가진 물리량에서 오차의 전파	63
8.2 회귀 계수의 오차	66
8.3 유효 숫자	69
<b>9 그레프의 Plot</b>	<b>73</b>
9.1 Plot의 형태	74
9.2 Plot의 오차 분석	75
9.2.1 오차 막대 (Error Bar)	75
9.2.2 회귀선의 불확실도	78
9.2.3 유효 데이터의 수	79
9.3 Eye-Balling Estimation	79
9.4 Plot을 이용한 interpolation	82
9.5 Plot 작성 팁 : 들어가야 할 내용	85
9.5.1 타이틀	85
9.5.2 축	86
9.5.3 범례 (legend)	87
9.5.4 회귀선	88
9.5.5 데이터 포인트	89
9.5.6 회귀 모델	89
9.5.7 물리량의 단위	89
<b>10 고급 회귀와 수치 방법</b>	<b>91</b>
10.1 수치 회귀	91
10.2 Gradient Descent	92
10.3 전산 회귀에서의 불확실도 분석	95
10.4 Convolution Fit	97
<b>11 보간법 (Interpolation)</b>	<b>99</b>
11.1 개요	99
11.2 보간 함수의 종류	100
11.2.1 상수 보간 (Nearest-neighbor interpolation)	100
11.2.2 선형 보간 (Linear interpolation)	101
11.2.3 다항 보간 (Polynomial interpolation)	101
11.2.4 3차 스플라인 (Cubic Spline Interpolation)	102
11.2.5 Padé 근사	104
11.3 데이터 샘플링과 불확실도	104
11.4 맷는 말	108

## 1.1 실험 통계학의 필요성

1.1 실험 통계학의 필요성	1
1.2 실험 통계학의 범위	2
1.3 실험 통계학의 목적	3

자연계는 불확실하다. 이는 많은 자연 현상에서 등장하는 물리량들이 정해진 참값을 가지지 않고, 분산을 가진 모집단으로서 존재함을 의미한다. 이러한 분산은 계에 내재된 양자역학적 불확실성과 측정의 분해능, 실험 설계와 이론 간의 차이 등 여러 가지 요인으로부터 발생하는데, 이는 같은 물리량을 반복하여 측정하여도 계속 동일한 값을 얻을 수는 없다는 것을 의미한다. 즉, 원하는 물리량을 얻어내기 위한 어떠한 실험 설계와 측정기기, 그리고 그 근본에 있는 자연 현상조차도 불확실도로부터 벗어날 수 없다.

실험을 통해 측정된 물리량은, 실험계가 가진 앙상블 평균으로서의 기대값 추출 과정의 결과이다. 그러므로 우리는 원하는 물리량의 참값이나 모분산을 직접 관측할 수 없으며, 오로지 실험을 통해 추출된 표본 집단의 Data Set만을 얻을 수 있다. 이 때 참값에 대한 최선의 추정과, 모분산에 대한 추정을 통해 실험 측정값이 얼마나 유의한지 추정하기 위해 표본집단의 통계적 성질을 이용할 수 있다.

모든 측정은 불확실도를 수반하므로, 원하는 물리량을 뽑아내기 위해서는 측정이 얼마나 불확실했는지를 평가하기 위해 모분산의 추정이 필요하며, 이를 토대로 각각의 측정치를 얼마나 신뢰할 수 있는지를 통계적으로 가늠한다. 통계적 유의성을 바탕으로 측정치를 기각할 수도 있다. 반복 측정을 통해 원치 않는 잡음을 배제하고, 실험

설계의 개선을 통해 오차가 적은 실험계가 되게끔 할 수도 있다. 변인 통제와 실험 변수 사이의 관계를 이용해 회귀 분석을 할 수 있으며, 측정치가 이론적으로 제시된 경향과 일치하는지 살필 수 있다. 이러한 과정의 결과는 이론에서 제시되었거나, 물음을 던진 계수나 상수에 대한 실험적 측정값을 정밀 / 정확하게 얻어내는데 사용된다. 더 나아가, 각각의 측정에 대한 유의 수준을 아는 것은 실험 검정을 통해 가설이 옳은지 그른지 판가름하는 데 사용될 근거가 될 수 있다.

이처럼 실험 통계학의 적용은, 가설에서 검증된 이론으로 받아들여지기 위한 통계적 근거를 제공한다. 실험 통계학의 적절한 적용은 오차가 많은 실험계로부터 물리학의 본질을 unveil해 낼 수 있는 강력한 도구인 셈이다.

## 1.2 실험 통계학의 범위

실험 통계학은 오차와 불확실도를 다루는 방법에 대해 논한다. 통계적인 방법으로 얻어진 측정치들의 집합을 분석하는 방법과, 캘리브레이션을 통해 실험계 자체의 신뢰도를 높이는 방법 등이 두루 사용된다. 이를 위해서는 오차의 발생 원인을 파악할 필요가 있으며, 오차의 종류를 발생 원인이나 특성에 따라 분류한다. 실험 설계나 회귀 변수의 적절한 선정을 통해 선형성을 높이고, 오차를 줄이는 방법에 대해서도 논의한다. 뿐만 아니라, 측정치가 어느 정도의 오차를 가지고 있는지 여러 가지 자유도로 추정하는 방법을 기술하게 된다.

본 교재에서는 불확실도의 개념과 이를 다루는 방법에 대해 논하며, 이를 바탕으로 한 실험적 유의성 검정에 대해 익힌다. 실험 설계상의 변인 통제가 측정치에 어떠한 영향을 줄 수 있는지에 대한 논

의를 바탕으로, 반복 측정과 회귀 분석을 통한 불확실도 검출에 대해 익히며, 비선형 경향을 가진 자료에 대한 고급 회귀와 수치적 방법 또한 논의한다. 계측기기의 보간으로부터 얻어지는 불확실도와 회귀 분석으로 얻어지는 통계적 불확실도가 최종적으로 알고자 하는 물리량을 얼마나 불확실하게 (흐릿하게) 하는지 논하기 위해 불확실도가 전파되는 과정을 다룬다. 이를 위한 기본적인 개념의 소개와 통계적 / 수치전산적 기법들을 다루는 방법을 살펴볼 것이다.

### 1.3 실험 통계학의 목적

실험 통계학은 올바른 연구 윤리를 기반으로, 정당한 방법으로 측정된 자료를 통계적 방법으로 분석하여 참값에 가까운 최선의 추정을 얻기 위해 이루어져야 할 일련의 과정을 논한다. 이를 통해 실험치가 가지는 불확실도를 원하는 유의수준에서 얻어내고, 이를 통해 실제 실험에서 양질의 결론을 더 강력한 근거 하에 얻는 데 도움이 될 수 있다.



## 2.1 오차의 정의

2.1 오차의 정의	5
2.2 불확실도의 정의	6
2.3 가중치	8
2.4 최학치	9

참값을 알고 있는 경우, 측정치의 부정확성은 오차를 이용해 논할 수 있다. 오차는 측정치에서 참값을 빼 준 값, 즉 측정치의 편차 (deviation)로서 정의된다.

$$\text{오차 (Error)} e_i \equiv X_i - \mu$$

많은 경우의 실험은 물리량을 측정하는 경우가 많은데, 측정하고자 하는 대상이 되는 물리량의 참값을 모르고 있는 경우가 대다수이다. 이 때에는 측정치가 원래 얼마였어야 하는지에 대한 참값을 모르므로, 측정이 얼마나 잘못되었는지 (빗나갔는지)를 논의하기 위해 오차를 사용할 수 없다. 이 경우에는 오차 대신 불확실도를 이용하여 측정의 정밀성을 논할 수 있다.

오차를 사용할 수 있는 경우는 목표하는 참값이 있는 경우로 한정된다. 가령, 1 [mm] 지름의 철사를 제조하고자 할 경우 생산된 철사의 지름을 레이저나 기계식 / 정전식 게이지로 읽었을 때 1.01 [mm]의 값을 가진다면, 오차는 +0.01 [mm]에 해당함을 알 수 있다. 이러한 오차 신호(error signal)는 기계가 만들어 내는 철사의 직경을 줄여야 한다는 것을 시사하며, 피드백 회로(loop)를 구성함으로서 오차가 0이 되도록, 즉 물리량이 목표치로 수렴하도록 제어할 수 있다. 이는 단순히 기계제조에 국한되지 않으며, 레이저의 파장이나

시료의 온도, 계에 흐르는 전류 등을 일정한 값으로 유지하기 위해 사용되며 PID (Proportional-Integral-Differential) 제어 등 여러가지 기법으로 응용되어 적용된다.

오차의 개념은 Tolerance 개념과 직접적으로 연관된다. 생산의 오차를 고려했을 때 기계부품이 설계 치수로부터 벗어날 수 있는 최대 허용한계를 Tolerance라 하며, 이는 주로 끼워맞춤을 하는 부품에서 중요하게 여겨진다. 병과 뚜껑, 축과 베어링 등 두 기계요소가 얼마나 강하게, 또는 헐겁게 체결되는지의 기준으로 허용치수 한계를 결정하며, 두 부품의 치수가 겹치는 경우 과잉 치수 차이를 Interference라 하며, 두 부품의 치수가 여유있는 차이를 가지는 경우 Clearance가 있다고 한다. 가령, 5.000 [mm] 직경의 구멍에 5.005 [mm] 직경의 막대를 끼운다면 0.005 [mm]의 Interference가 있는 것이며, 4.995 [mm] 직경의 막대를 끼운다면 0.005 [mm]의 Clearance가 있는 셈이다. 이는 기계공학에서 매우 중요한 역할을 갖는 GD&T (Geometric Dimensioning and Tolerancing)에 해당한다.

이처럼 오차(Error)도 중요한 역할을 할 수 있지만, 모르는 물리량을 측정해내는 실험에서는 오차를 계산할 수 없으므로 불확실도(Uncertainty)를 대신 사용해 오차 분석을 진행해야 한다. 이에 따라 앞으로 오차와 불확실도의 용어를 혼용하게 되는데, 대다수의 경우 오차(error)의 terminology로 기술된 경우에도 거의 대부분 불확실도를 의미할 것이다.

## 2.2 불확실도의 정의

참값을 모르는 경우, 그 역할을 대신하기 위해 최확치  $x_{be}$  (Best Estimation)의 개념을 사용한다. 최확치는 값을 얻고자 하는 물리량

의 참값에 대한 최선의 추정치이다. 이상적인 등분산 데이터 (모든 측정이 동등한 불확실도를 가지는 경우)에 대해서는, 단순히 표본의 평균을 최확치로 생각할 수 있다. 그렇지 않은 경우에는 후술할 방법을 통해 최확치를 얻을 수 있다. 정리하자면, 참값에 대한 정보 없이 표본집단의 통계적 성질만을 이용하여 참값을 추정한 결과를 최확치라 한다. 이상적인 경우 이는 데이터의 평균과 큰 차이가 없으나, 각각의 측정이 동등하다고 가정할 수 없을 만큼 분산의 차이가 큰 경우, 혹은 명백히 기각해야 하는 데이터가 포함되어 있는 등의 경우에는 최확치는 평균에서 상당히 많이 달라질 수 있다.

최확치는 참값이 아니다. 데이터를 대표하는 값이나 평균과도 약간의 차이가 있다. 최확치는 그저 참값에 대한 최선의 추정이므로, 이 값이 물리적으로 유의미해지려면 최확치가 얼마나 참값과 가까운지에 대한 강력한 근거가 필요하다. 이를 보장해 주는 척도가 바로 불확실도이다.

불확실도 (Uncertainty ;  $\sigma$ )는 참값이 최확치로부터 가질 수 있는 최대 편차를 의미한다. 가령, 실제 참값이 1.67 [-]인 실험에서 최확치를 1.3 [-]으로 추산하였다면, 불확실도가 0.4 [-]인 경우  $1.3 \pm 0.4$  [-]의 범위 안에 참값인 1.67 [-]이 포함되므로 최확치는 참값으로부터 불확실도보다 멀리 떨어져 있지 않게 되며, 불확실도는 유의미한 값이 된다. 이는 참값에 대한 추정이 최확치에 의해 이루어졌을 때, 실제 참값이 추산된 최확치로부터 얼마나 멀리 벗어날 수 있느냐에 대한 기준으로서의 의미를 가지기도 한다.

즉, 불확실도는 참값이 최확치로부터 어느 정도 범위 이내로 떨어져 있다는 것을 보장하는 구간이며, 참값이 이 구간을 벗어나는 것은 최확치 추산이 실패했다는 것을 의미한다. 따라서 불확실도를 크게 잡으면 (참값에 대한) 추산이 실패할 확률을 0에 가까이 줄일 수 있으나, 0으로 만들 수는 없다. 작은 불확실도를 잡으면 추산이

실패할 확률이 높아진다. 따라서 불확실도의 크기는 최확치의 추산이 얼마나 “튼튼하게” 이루어지느냐에 따라 달라진다. 이는 다음 장에서 논의될 유의수준을 기준으로 삼는다.

한편 불확실도가 크다는 것은 참값이 측정치로부터 멀리 떨어져 있을 수 있음을 의미한다. 동일한 물리량을 측정한 (동일 유의수준) 두 개의 데이터가 서로 다른 불확실도를 가질 때, 참값은 불확실도가 작은 쪽의 측정치에 더 가까울 것이다. 이는 불확실도가 측정치의 모분산의 개념을 한다는 것을 시사하므로 불확실도는 측정의 질을 평가하는 데 사용될 수 있다.

### 2.3 가중치

각각의 데이터 포인트를 얻는 측정마다, 측정 방법에 따라 측정의 불확실도가 달라진다. 따라서 불확실한 측정을 동등하게 반영하여 최확치를 얻을 경우, 참값을 올바르게 추산하지 못할 수 있다. 가령 실험의 오류로 인해 outlier가 발생하였는데도 이를 동일한 가중치로 반영한다면 최확치 추산은 outlier의 편차 방향만큼 참값으로부터 멀어지게 될 것이다. 이에 따라 각각의 측정이 얼마나 믿을만한지를 정량적으로 반영하여, 최확치 산정을 위한 가중치를 부여하여야 한다. 즉, 측정의 불확실도가 클수록 가중치를 덜 부여하여 평균내어야 한다.

이를 반영하는 방법은 다음 절에서 기술하는 것과 같이, 매 측정의 분산이 다른 (이분산) 데이터 셋의 경우 각각의 가중치는 분산의 역수 ( $1/\sigma_i^2$ )에 비례한다.

## 2.4 최적치

위의 내용을 바탕으로 최적치를 추산하기 위한 단계로, 지수제곱편차를 정의한다. 각각의 측정치  $X_i$ 를 얻을 때 사용된 측정의 분산이  $\sigma_i^2$ 일 때, 그 측정치가 참값  $\mu$ 에서 편차  $\delta_i \equiv X_i - \mu$ 만큼 떨어져 검출될 확률은  $\exp(-\delta_i^2/2\sigma_i^2)$ 에 비례한다. 이를 모두 곱한 값인 지수제곱편차  $E(\mu) = \prod_i e^{-\delta_i^2/2\sigma_i^2}$ 로 정의된다. 이를 최대로 만드는 값이 바로 참값에 대한 최선의 추정, 즉 통계적 최적치(best estimation)이다.

여기서 알 수 있는 점은, 불확실도가 3배인 측정에서는 참값으로부터의 편차가 3배가 된 위치에서도 값이 얻어질 확률이 동일하다는 점이다. 이는 반대로 이야기하자면, 해당 데이터 포인트로부터 최적치가 멀리 떨어져도 지수제곱편차를 크게 낮추지 않음을 의미하며, 이는 더 큰 편차가 허용되고, 더 낮은 기여도를 가진다는 것을 의미한다.

계산을 위해  $E$ 를  $\mu$ 로 편미분하면  $\frac{\partial E}{\partial \mu} = \left( \sum_i \frac{\delta_i}{\sigma_i^2} \right) E = 0$ 을 얻는다. 이로부터 최적치는 다음과 같은 식으로 얻어진다.

$$\sum_i \frac{X_i - X_{be}}{\sigma_i^2} = 0 ; \quad \sum_i \frac{X_i}{\sigma_i^2} = \sum_i \frac{1}{\sigma_i^2} \cdot X_{be}$$

$$\therefore X_{be} = \frac{\sum \frac{X_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

(이분산 추출된 데이터 표본의 경우)

이것이 최적치 계산에서 가중치 ( $w_i = 1/\sigma_i^2$ )가 필요한 이유가 된다. 분산이 큰 측정일수록 최적치에 통계적으로 유의미한 기여를

하기 어려우며 낮은 가중치를 가지는 것이다.

이 때, 실험이 충분히 정밀해 계측기기의 분산을 무시할 수 있는 경우, 각각의 측정은 등분산 표본을 추출한 것으로 이해할 수 있다. 이 경우 최확치는 단순히 측정치들의 평균으로 얻어진다.

최확치의 산정에 가중치가 반영된다는 것으로부터 몇몇 부정밀한 / 신뢰하기 어려운 실험으로부터 얻어진 자료는 기각을 검토할 수 있음을 확인할 수 있다. 분산이 매우 큰 실험에서 얻어진 측정치라면, 분산의 역수로 얻어지는 가중치가 매우 작을 것이므로 사실상 최확치 산정에 반영되지 않으며, 이는 데이터를 기각 (rejection)하는 것과 거의 같다. 이처럼 불확실도 해석을 통해 실험에 내재된 여러 가지 오차 요인을 걷어내고, 최확치를 참값에 수렴시키기 위한 측정과 분석을 진행할 수 있다.

# 3

## 유의 수준

### 3.1 유의수준과 신뢰구간

3.1 유의수준과 신뢰구간 . . . . . 11

3.2 유의 수준의  $\sigma$  표기 . . . . . 12

앞서 살펴본 바와 같이 불확실도의 크기는 최확치가 참값에서 벗어나는 것을 얼마나 엄격하게 유의할 것인지에 따라 달라진다. 최확치에서 불확실도를 가감한 구간을 신뢰구간이라 하며, 신뢰 구간에 참값이 포함되지 않을 확률을 유의확률이라 한다. 이에 따라, 유의확률을 작게 선정할 경우 참값이 신뢰구간에서 벗어날 가능성 이 적어지나, 이 경우 신뢰구간을 지나치게 크게 잡아야 할 수도 있다. 너무 큰 신뢰구간은 측정된 물리량의 유효 숫자를 줄이므로, 적절한 유의 수준을 선정하고, 실험의 정밀도를 개선하여 좁은 신뢰구간 하에서도 낮은 유의확률을 가질 수 있도록 하는 것이 실험 설계에서 기본적으로 지향해야 할 방향이다.

이에 따라, 불확실도의 정의를 다시 기술하면 다음과 같다 :

$\sigma : \mu \in [X_{be} - \sigma, X_{be} + \sigma]$  일 확률이  $\alpha$ 가 되는 값

$\alpha$  : 유의 수준 (significance level)

신뢰 구간 :  $[X_{be} - \sigma, X_{be} + \sigma]$

유의수준이 달라지면 불확실도의 크기가 다르게 추산되므로, 오차의 전파나 덧셈/뺄셈 등을 고려할 때 반드시 유의수준을 통일해주어야 하며, 최소자승법 등의 통계적인 방법으로 불확실도를 추산할 경우 유의수준은  $1\sigma$ 가 되는 것이 일반이다. 반면, 육안 측정 시에는 대

체로 알파=0.05 ( $2\sigma$ 에 해당) 의 유의 수준을 가지는 것이 편리하다. 이는 눈을 이용한 측정 시  $1\sigma$ 의 신뢰수준은 너무 좁은 신뢰구간에 대응되어, 참값이 신뢰구간 안에 들어올 것이라고 예측하기 불편함을 느끼게 되기 때문인데, 통계적으로 육안 관측을 통해 신뢰구간을 추정할 경우에는 약  $2\sigma$ 의 신뢰 수준(유의확률 5%)을 제시하는 것이 편안한 크기의 신뢰구간을 얻는 것으로 알려져 있다. 이에 대한 자세한 내용은 6장의 아날로그 기기의 눈금 보간(Interpolation)에서 다룬다.

### 3.2 유의 수준의 $\sigma$ 표기

이처럼 유의수준의 크기를 논하고, 실험 자료간 이를 통일하여 처리하는 것은 실험의 근거를 튼튼히 하기 위해 상당한 중요성을 가진다. 유의 수준을 나타내는 데에는 크게 두 가지 방법이 널리 사용되는데, 첫번째는 유의 확률을 직접 나타내는 것이며, 두 번째는  $\sigma$  표기를 이용한 신뢰수준을 나타내는 것이다.

일반적으로 많이 사용되는 유의 수준은  $1\sigma$ ,  $2\sigma$  정도의 크기를 가진다. 학술적인 용도로는 불확실도를 작게 잡을 수 있으며 통계적으로 쉽게 기술되는  $1\sigma$ 의 유의수준이 자주 사용되며, 육안으로 관측하는 등 직관적인 규모의 불확실도가 필요한 경우(ex. plot 시 오차 막대의 visualization) 등에는  $2\sigma$ 의 유의수준이 종종 사용된다.

$1\sigma$ 는 약 63.2% ( $=1-1/e$ ),  $2\sigma$ 는 약 95%의 확률로 참값이 신뢰 구간 안에 포함됨을 의미한다. 유의수준을 낮출수록 (=더 높은  $n$ 을 가지는  $n\sigma$ 의 신뢰수준) 추산이 옳은 결론으로 이어질 가능성이 높아 (튼튼한 추산), 잘못된 결론 (참값이 신뢰구간 밖에 있는 경우)을 얻을 확률이 낮아진다.  $1\sigma$ 는 추산이 틀릴 확률이 36.8% ( $=1/e$ )

나 되지만,  $2\sigma$ 의 유의수준은 추산이 틀릴 확률이 5%에 불과하다. 다만 참값이 신뢰구간을 벗어날 확률을 줄이더라도, 신뢰구간이 너무 넓다면 실제 참값이 어느 정도 값을 가질지 정밀하게 예측하는 것이 불가능하므로, 좁은 신뢰구간을 잡아 첨예하고 정밀하게 참값을 추산하되 유의할 확률을 높이는 것과, 참값이 신뢰구간을 벗어나는 것으로부터 안심할 수 있을 정도로 (낮은 유의확률) 충분한 신뢰구간을 제시하되 참값이 가질 수 있는 오차 범위를 크게 허용하는 방법 사이의 Trade-off가 존재한다. 수학적인 계산에서는 중심 극한 정리 (CLT ; Central Limit Theorem)에 따라 측정을 통해 추출된 표본 집단이 정규 분포를 따를 것을 상정하므로,  $1\sigma$ 의 신뢰수준을 잡고 불확실도의 전파 연산을 수행하는 것이 편안하나, 이는 상기한 바와 같이 육안으로 보기에는 불편할 수 있다. 예를 들어, 선형 회귀 분석에서 회귀선으로부터 편차가 큰 데이터의 오차 막대를 그릴 때  $1\sigma$ 의 신뢰 수준을 사용한다면 데이터 포인트 중 37%는 오차막대 범위 안쪽으로 회귀선이 지나지 않을 것이다. 그러나  $2\sigma$ 의 신뢰수준을 적용하면 오차막대가 약 두 배 증가하므로 회귀선이 지나지 않는 오차막대를 가지는 데이터 포인트는 고작 전체의 5%에 불과하게 된다. 이는 오차막대의 크기가 데이터 포인트들의 편차의 상식적인 규모와 거의 일치하게 됨을 의미한다.

신뢰수준의  $\sigma$  표기는 표본 집단을 정규분포로 상정하는 중심 극한원리의 결론과 결을 같이 한다. 정규 분포에서 대칭 신뢰구간을 잡았을 때, 참값이 신뢰구간 바깥에 존재할 확률은 양 쪽 끝단 Tail 면적 두 개의 합이며, 이는 유의확률 알파와 같다. 이 Tail이 잘리는 위치가 정규 분포의 표준편차 (standard deviation)의  $n$ 배가 되는 위치일 때의 유의확률을,  $n\sigma$ 의 신뢰수준을 가진다고 표기한다. 보다 직관적으로 표현하면,  $n\sigma$ 의 신뢰수준이 허용하는 유의확률은 표준 편차  $\sigma$ 를 가지는 정규 분포의 밑넓이를  $-n\sigma$ 부터  $n\sigma$ 까지 적분한 값을 1에서 빼준 것과 같다. 이의 역함수를 취하면, 원하는 유의 확률을  $\sigma$

표기를 이용해 신뢰 수준으로 표기할 수 있다.

$$1 - \alpha = \int_{-n\sigma}^{n\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{X^2}{2\sigma^2}} dX = -\sqrt{\frac{2}{\pi}} \int_0^n e^{-X^2/2} dX$$

$\sigma$  수준 표기는 관측 결과의 신빙성을 보장하기 위해서 널리 통용된다. 유명한 예시로는 Higgs 보존, 혹은 중력파의 검출 등의 신뢰 수준을  $\sigma$  표시로 제시하는 것을 쉽게 들어 볼 수 있다. 이는 때때로 99.999... %의 신뢰 수준이라는 설명과 함께 제공되는데, 이는 유의 확률을 1에서 빼준 값과 같다.

신뢰수준의  $\sigma$  값을 두 배 늘리면 불확실도 범위는 대략 두 배 증가하지만, 실제 실험계는 정규분포가 아닌 t-분포를 따르므로 표준 편차의 n배에 해당하는 적분 구간이 정규 분포의 경우로 계산된  $\sigma$  표기에서의 예와는 다르다! 그러나 불확실도의 유효 숫자는 후술할 이유에 따라, 2자리를 넘겨 표기하는 경우가 거의 없으므로 평범한 사용례에서는  $\sigma$  값을 n배 늘리면 신뢰구간 역시 약 n배 넓어지는 것으로 간주할 수 있다. 이를 이용해 오차의 전파를 사용하여 최종적인 값의 불확실도를 추산해야 하는데 측정치들의 신뢰수준이 다른 경우에는 어느 정도의 비례 관계를 이용하여 편리하게 유의수준을 통일해 줄 수 있다.

예를 들어, 광전 효과 실험에서 플랑크 상수를 얻기 위해 주파수  $f$ 를 바꾸어 가며 저지 전압  $V$ 를 측정한다고 하자. 이 때, 조작변인은 광원의 주파수  $f$ 가 되며, 종속변인은 저지전압  $V$ 이고, 그 기울기는  $h/e$ 에 해당한다. 선형 회귀의 기울기를  $B$ 라 하면, 플랑크 상수는  $B$ 에  $e$ 를 곱해 얻을 수 있다. 이때,  $e$ 를 조금 부정확한 실험으로 측정하여,  $(1.6 \pm 0.3)e - 19[C] (2\sigma)$ 를 얻고, 선형 회귀를 최소자승법을 이용하여  $B = (4.1 \pm 0.2)e - 15 [V \cdot s] (1\sigma)$ 를 얻었다면, 이 두 측정치를 곱할 때, 불확실도의 신뢰 수준을 동일하게 맞추어 줄 필요가 있다. 둘 모두  $2\sigma$ 의 신뢰수준으로 통일한다면,  $B = (4.1 \pm 0.4)e - 15 [V \cdot s]$   $(2\sigma)$ 로 높여 준 후 오차의 전파 관계를 이용하여 두 값을 섞어 줄 수

있다. 물론 어디까지나 이 방법은 간략한 추산 과정이며, 정교한 측정에서는 불확실도의 크기가 검정 결과에 치명적인 영향을 미칠 수도 있다. 이 경우에는 정확한 계산 과정을 거쳐 여분의 유효 숫자를 두고 불확실도를 처리하는 것이 권장된다. 물론 최종 표기에는 2자리를 넘는 유효숫자를 가지도록 불확실도를 기입하지는 않는다.



# 4

## 실험 검정

### 4.1 아르키메데스의 왕관

4.1 아르키메데스의 왕관 . . . . .	17
4.2 p-value를 이용한 유의성 검정 . . . . .	18
4.3 Rejection of Data . . . . .	22

실험 검정의 역사는 아르키메데스 시대로 거슬러 올라간다. 어떠한 명제 혹은 가설을 실험을 통해 참인지 거짓인지 검정하고, 그 검정이 잘못될 확률을 추산하는 것은 과학 이론의 발전을 위해 필수적이다. 이러한 실험 검정의 가장 쉬운 예시는, 왕관이 은이 섞인 금 합금인지, 순금인지를 밀도 측정 실험을 통해 분간하려 하였다는 아르키메데스의 왕관 이야기로부터 찾을 수 있다.

금은 은보다 무겁다. 따라서 보다 가벼운 원소인 은과 합금하게 될 경우 밀도가 낮아지게 된다. 그러나 아보가드로 수 규모의 원자가 있는 금속에서, 은 원자가 단 하나도 존재하지 않는 순금이 있을 수 있는가? 혹은, 순금인데도 질량 측정은 정확하나 부피가 실험 오차에 의해 실제보다 크게 측정되어 은이 섞였다는 잘못된 결론을 얻을 가능성은 없는가?

모든 실험은 오차를 포함하고 있기 때문에, 실험을 통해 가설을 판별할 경우 반드시 유의성 검정이 필요하며, 그 기준은 실험의 불확실도로부터 얻어진다.

## 4.2 p-value를 이용한 유의성 검정

가장 널리 쓰이는, 대표적인 실험 검정의 기준은 p-value를 통한 판정법이다. P-value (probability value)는 앞서 살펴본 유의화률과 같은데, 이는 귀무 가설이 옳다고 가정할 때 어떤 통계량이 실험 결과보다 귀무 가설의 구간 밖으로 더 멀리 떨어져서 발견될 확률을 의미한다. 이는 귀무가설이 틀리거나, 대립가설이 맞을 확률이 아니며, 실험 결과가 귀무가설을 기각할 충분한 통계적 능력이 있는지만을 검정할 뿐이다. 이에 따라 계산된 p-value가 검정 기준이 되는 유의수준보다 큰지, 작은지에 따라 가설을 기각할 수 있는지 검정하게 된다.

검정을 위해서는 귀무가설  $H_0$ (null hypothesis)과 대립가설  $H_1$ (alternative hypothesis)을 설정해 주어야 한다. 대립 가설은 실험자가 검증하고자 하는 가설이며, 예를 들면 아르키메데스의 왕관 밀도가 금의 밀도보다 낮다는 것을 대립 가설로 잡을 수 있다. 대립 가설이 참이라면, 왕관은 금이 아닌 다른 물질이 섞인 합금임을 주장할 통계적 근거가 생긴다. 이를 검정하기 위해, 대립 가설과 대립하는 귀무가설을 설정하고, 실험을 통해 귀무가설이 기각된다면 대립가설이 채택되는 방식으로 검정을 진행할 수 있다. 이 경우 귀무가설은 아르키메데스의 왕관 밀도가 금의 밀도와 같거나 그보다 높다는 것이 된다.

검정에는 대소 관계를 검정하는 단측 검정과, 값이 같은지 다른지를 논하는 양측 검정이 있다. 이 두 경우, p-value의 정의가 다음과 같이 달라진다 :

“검정 통계량  $t$  (=실험치)가 분포  $T$ 를 따를 때, 귀무가설  $H_0$

하에서 검정통계량보다 멀리 있는 값이 얻어질 확률”

단측 검정의 p-value (left-tail) :  $p = Pr(T \geq t | H_0)$

단측 검정의 p-value (right-tail) :  $p = Pr(T \leq t | H_0)$

양측 검정의 p-value :  $p = Pr(|T| \geq |t| | H_0)$

P-value를 이용한 실험 검정에서 통용되는 Convention은 유의 수준을 0.05로 정한다. 이에 따라 p-value가 0.05보다 작으면 통계적으로 유의한 결과를 얻음을 의미한다. P-value가 작을수록, 실험을 통한 결과가 귀무가설이 맞다고 상정한 채 발생할 확률이 작다는 뜻이며, 이는 실험 결과가 귀무가설을 기각할 충분한 근거가 있다는 뜻이다. (귀무가설이 맞다면, 이렇게 멀리 떨어진 실험 결과가 얻어지기 매우 어렵기 때문이다.)

최근 학계에서는, p-value의 검정기준을 0.05에서 0.005로 높여는 움직임이 존재한다. 이는 대립가설을 채택하기 위해 귀무가설을 기각하는 데 필요한 기준을 높이는 것이며, 더 정확하고 정밀한 실험이 가능해야 귀무가설을 기각할 수 있도록 요구하는 것이다.

통계학에서의 검정의 예시로, 신발 공장에서 만들어진 신발의 크기가 260 [cm]의 평균을 가지며, 3 [cm]의 표준 편차를 가진다고 하자. 이 때, 270 [cm]의 크기로 만들어진 신발을 불량품으로 판정할 수 있는지 검정한다면, 대립 가설은 ‘신발의 크기는 260 [cm]와 다르다 (그러므로 불량품이다)’로 선정할 수 있으며, 귀무 가설은 ‘신발의 크기는 260 [cm]와 같다’로 선정하여 양측 검정을 시행할 수 있다. 양측 검정 시 검정 변수가 따라야 하는 분포는 평균이 260 [cm], 표준 편차가 3 [cm]인 정규 분포인데, 이 때 p-value는 이 정규 분포에서 편차인 10 [cm]보다 큰 값이 얻어지거나, -10 [cm]보다 작은 값이 얻어질 확률을 의미한다. 편차의 크기가 표준편차의 3.33배

가량이므로, 표준 정규 분포표에서 약  $p=0.0008$ 임을 얻는다. (이는 약  $3.33\sigma$ 의 유의확률과 같다.) 이 값이 검정기준 0.05보다 작으므로, 270 [cm]의 신발은 만들어져야 하는 신발의 크기와 같다는 가설을 기각할 충분한 통계적 근거가 존재한다. 이는 귀무 가설이 기각되고, 대립 가설이 채택되었으므로 신발이 불량품이라고 할 근거가 있다는 것을 의미한다.

통계를 이용한 분석과 실험값의 불확실도를 이용한 분석에는 차이가 있다. 통계에서는 한 값이 통계적으로 얻어진 평균으로부터 통계적으로 얻어진 표준편차에 비해 충분히 큰 편차로 다른지를 검정하는데, 실험에서는 반복 측정 이외의 방법으로도 실험 측정으로 얻어진 하나의 최학치를 정할 수 있으며, 이 값이 기존 이론이나 다른 측정과 같은지, 다른지, 혹은 크거나 작은지를 판별하여야 하므로 평균과 표준편차를 이용하는 위 예시와는 다르게, 최학치와 불확실도를 이용해 검정을 수행하여야 한다.

실험 결과의 분석에서는 불확실도를 이용해 p-value test를 수행할 수 있다. 이 경우 p-value를 계산하기 위해 상정하는 분포는  $1\sigma$ 의 신뢰수준으로 얻어진 실험값의 불확실도를 표준 편차로 가지는 정규분포를 따르는 것으로 상정할 수 있다. 통계학에서는 실제로 여러 번 추출된 표본집단의 분산을 알 수 있지만, 실험에서는 측정 방법이나 실험 설계에 따라 반복 측정에 의한 불확실도 추산을 할 수도, 회귀를 이용한 불확실도 추산을 할 수도, 측정기기의 눈금을 보간하는 과정에서 불확실도를 얻어낼 수도 있다. 이렇게 여러 방법으로 측정된 불확실도가 오차의 전파 관계를 이용해 하나로 합쳐져 검정의 기준으로 사용되기 때문에, 통계학에서의 유의성 검정의 예시에서와는 다르게, 표준 편차 대신 실험으로 얻어진 불확실도를 검정의 기준으로 사용하게 된다.  $1\sigma$  신뢰수준의 정의를 생각해 보면, p-value가 통계학에서 사용되는 예시와 같은 유의수준으로 사용될 수 있음을 알 수 있다.

두 실험값이 같은지 다른지를 논할 때 주의할 점은, 한 실험값과 다른 실험값 모두 불확실도를 가진다는 것이다. 이를 처리하는 좋은 방법은, 두 실험값을 각각의 물리량으로 비교하는 것이 아니라 두 값의 차이를 편차로서 검정변수로 채택하는 것이다. 이 때 새로운 검정변수의 불확실도는 원래의 두 값의 각각의 불확실도로부터 오차의 전파를 이용하여 얻어낼 수 있다. 오차의 전파를 적용하기 위해서는 앞서 언급한 바와 같이 두 물리량 측정의 신뢰수준을 일치시켜 주어야 하며, 오차의 전파에 사용되는 공식은 8장에서 후에 논의하게 될 것이다.

정리하면, 아르키메데스의 왕관 예시에서 사용되는 검정변수는 아르키메데스 왕관의 밀도 - 금의 밀도가 되어야 한다. 이 때 검정통계의 분포는  $1\sigma$  신뢰수준으로 추산한 아르키메데스 왕관의 불확실도와 금의 밀도 측정 불확실도의 합을 표준편차로 가지는, 평균이 0인 정규 분포로 산정한다. 이 때 *p-value*는 실험을 통해 얻어진 아르키메데스 왕관의 밀도와 금의 밀도의 차이보다 더 작은(작다는 것은 절대값이 아니라 실제 값이 작다는 것 - 즉 음의 방향으로 더 멀어진다는 의미이다) 통계량이 생길 확률이 된다. 이 값이 0.05보다 작다면, 아르키메데스 왕관은 금이 아닌 다른 물질이 섞여 있을 충분한 실험통계적 근거가 존재한다는 것이며, 0.05보다 크다면 아르키메데스의 왕관이 순금이 아니라고 주장할 실험적 근거가 부족하다는 것을 의미한다.

이때 0.05 대신 0.005를 검정기준으로 설정한다면, 더 정밀한 실험을 하여 불확실도가 작게끔 측정하거나, 혹은 실제로 왕관 밀도의 최학치가 금의 밀도의 최학치로부터 불확실도 대비 충분히 멀리 떨어진 편차를 가져야 기각이 가능하다. 즉 검정에 보다 엄격한 기준을 부여하는 셈이다.

### 4.3 Rejection of Data

측정치가 상식적인 오차 범위를 한참 벗어난 값을 얻어진 경우, 측정의 오류가 있을 것을 추정해 볼 수 있다. 그러나 명확한 기준 없이 이러한 자료를 오류라고 단정짓는다면, 중요한 peak 등의 물리적 가치를 놓칠 수 있으며, 실험자의 주관에 따라 결과를 조작하는 데 악용될 수도 있다. 이에 따라, 최학치의 최선의 추정을 위하여 데이터를 제거하는 합리적인 기준이 필요하며, 이는 데이터가 잘못 측정되었을 확률, 즉 유의성을 기반으로 검정함으로서 알 수 있다. 측정이 잘못되었을 가능성에 둔감하게 유의하여 모든 데이터를 전부 포함할 경우, outlier들을 모두 반영하게 된다. 이는 첨예한 경향을 잘 반영할 수 있으나 노이즈에 취약하다는 문제를 가진다. 반면 측정이 잘못되었을 가능성을 민감하게 판단하여 트는 데이터를 많이 걸러 낼 경우, 얻어지는 경향은 smoothing되게 된다. 이는 회귀 분석을 위해서는 편리할 수 있으나, 물리적 의미를 가지는 중요한 피크를 놓치는 결과로 이어질 수 있다. 이에 따라, 유의 확률을 상식적인 선에서 조절해 가며 실험치로부터 유의미한 결론을 얻기 위해 데이터를 기각하는 것을 검토해 볼 수 있다.

데이터 기각은 오류가 있거나, 오류가 있을 것으로 의심되는 측정에 의해 실험의 신뢰수준을 해치는 것을 막고 정확한 분석이 가능하게 하기 위해 필요할 수 있다. 데이터 기각이 가능한 경우는 (1) 실험자의 명백한 오류로 실험계가 잘못 구성되었음을 확인한 경우 (이 경우에는 데이터를 억지로 반영하면 안 된다.) 혹은 (2) 통계적인 검출을 통해, 최학치에서 너무 큰 편차를 가지는 자료가 믿을 수 있는 과정으로 검출되었을 확률이 너무 낮은 유의확률을 가지는 경우에 해당한다. 두 번째 과정을 적절한 기준 없이 남용한다면, 실험자 혹은 해석자의 자의적 기준에 따라 데이터를 마사지하는 사태로 이어질 수 있다. 이러한 습관이 쌓이면 심각한 연구 윤리 문제가 될 뿐 아니라, 편향된 (biased) 데이터 추출에 의해 실험계의 참값을 제대로

바라보지 못할 수 있으므로 잘못된 결론에 도달할 위험에 취약해진다. 데이터 기각을 통계적으로 논하기 전에, 데이터의 기각은 반드시, 실험자가 원하는 값으로 최적화를 Shift시키기 위함이 아니라, 참값의 통계적 유의성을 악화시키는 Outlier들을 배제하는 의도임을 명확히 숙지하고 있어야 한다.

데이터 기각에 쓰이는 기준 중 하나로는, Chauvenet Criterion 이 있다. Chauvenet 기준은  $N$ 개의 데이터 셋  $\{X_1, \dots, X_N\}$ 에 대해  $i$ 번째 자료  $X_i$ 가 outlier인지 의심할 때, 나머지의 평균  $\bar{X}$ 로부터  $X_i$  가 얼마나 멀리 떨어져 있는지를 검토한다. 측정 자료의 분포가 정규분포일 것으로 상정하고, 나머지 측정치들의 평균과 표준편차를 이용한 정규 분포로부터  $X_i$ 가 검출될 확률  $P$ 를 얻고, 이에 측정 횟수  $N$ 을 곱한 값이 0.5보다 작은 경우 (데이터가 얻어질 확률이  $1/N$ 은 되어야 통계적으로 유의미한 (허용할 수 있는) 편차임을 의미하므로, 이보다 발생 확률이 절반이 되는 정도로 큰 편차는 통계적으로 유의미하지 않은 자료로 의심하는 것이 합당하다) 해당 데이터를 기각하는 것이다.

데이터 기각은 선형 회귀에도 적용될 수 있는데, 대부분의 자연계는 선형성을 가지는 Golden Rule 구간이 제한적으로 존재한다. 이는 데이터 셋의 경향이 선형적인 구간과, 비선형적인 구간이 섞여 있는 경우가 많다는 것을 의미하는데, (이는 주로 이론오차에 의한 선형화 근사 이론으로부터의 Gradual한 편차로부터 기인한다.) 이때 어느 구간까지가 선형성이 보장되는 구간인지를 알기 위해 데이터의 기각 기준을 검토할 수 있다.

선형성을 논하는 척도는 제곱상관계수  $R^2$ 이며, 이 값이 1에 가까울수록 선형임을 의미하고 잘 구현된 실험계에서는 못 해도 0.99 이상의 값을 가진다. 이때, 비선형 구간으로 진입할수록 데이터 포인트들이 가지는 편차는 한 방향으로 치우쳐 Gradual Shift를 보이게

되는데, 이들을 반영하면 선형성의 척도인 제곱상관계수가 조금씩 감소하게 된다. 이 감소량이 원하는 유의 수준에 해당하는 값을 벗어난다면, 해당 데이터가 발견된 구간을 완전히 배제하고 선형 구간을 잡아야 한다. 이는 회귀선에서 벗어났다고 Data Massage를 하는 것이 아닌, 유의 확률에 따라 어디까지의 데이터를 포함할지의 구간을 변경하는 것에 해당하며 신뢰 구간의 크기가 신뢰 수준에 따라 변화하는 것과 같은 맥락으로 이해해야 한다.

# 5

## 불확실도의 원인

### 5.1 정밀도와 정확도

측정의 신뢰도를 논하기 위해, 정밀한 측정과 정확한 측정을 구분할 필요가 있다. 정확한 측정은, 반복 측정을 통해 얻어진 최적치(참값에 대한 최선의 추정)가 실제 참값과 가까운 경우를 의미한다. 가령, 그림과 같이 표적에 맞은 자국이 모여 있건, 퍼져 있건 그 평균값(기하학적 중심)이 실제 표적의 중심(참값)에 가까운 경우, 이 데이터 셋은 정확한 측정으로 간주한다. 반면, 데이터의 중심이 실제 참값에 얼마나 가까운지와 무관하게, 데이터가 조밀하게 모여 있다면(표본의 분산이 적다면) 정밀한 측정으로 간주한다.

5.1 정밀도와 정확도	25
5.2 불확실도의 종류 - 반복 측정에 의한 제거 가능성	26
5.3 불확실도의 종류 - 발생 원인에 따른 분류	27
5.4 불확실도에 대처하는 방법	31

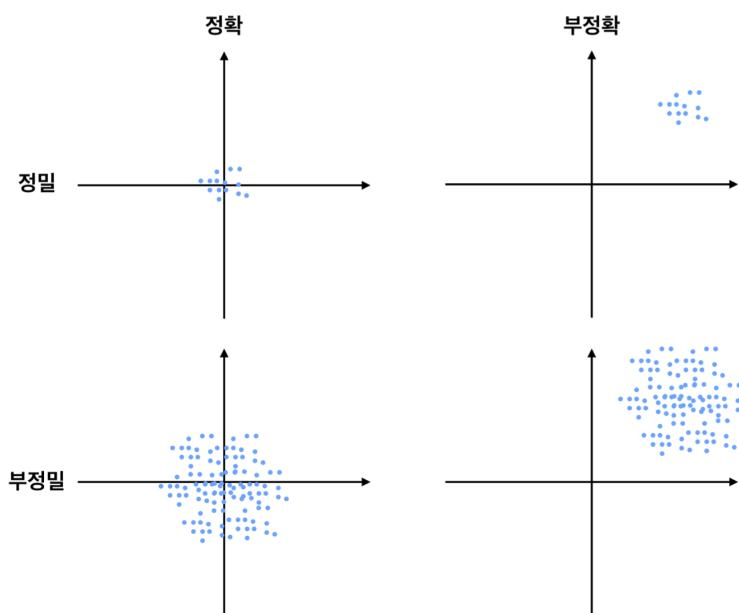


Figure 5.1: 정밀도와 정확도

## 5.2 불확실도의 종류 - 반복 측정에 의한 제거 가능성

실험의 불확실도는 측정을 부정확하고, 부정밀하게 만든다. 데이터가 정밀한지 부정밀한지는 추출된 표본의 분산을 보고 알 수 있으나, 데이터가 정확한지 부정확한지는 참값을 모르면 알 수 없다. 이러한 차이는 반복 측정을 통해 검출할 수 있는데, 정확한 실험이라면 부정밀한 계측 기법을 사용하더라도 측정을 반복하면 Random하게 분포되어 얻어지는 불확실도가 소거되어 참값에 가까운 최적치를 얻을 수 있게 된다. 반면, 부정확한 실험이라면 아무리 반복 측정을 거듭해도 최적치는 참값에 어느 한계 이상 다가갈 수 없게 된다.

이러한 구분을 적용해 오차의 종류를 계통 오차 (Systematic Error)와 우연 오차 (Random Error)로 구분한다. 반복 측정을 통해, 오차가 존재함에도 최적치를 참값에 수렴시킬 수 있는, 즉 제거할 수 있는 오차(오차를 없앤다는 것이 아니며, 오차가 있음에도 참값을 unveil할 수 있다는 의미이다)를 우연오차라 하며, 이는 실험계에서 얻어지는 Random Noise의 크기 척도가 된다. 우연오차의 크기는 반복측정을 통해 검출될 수 있는데, 이때 얻어진 표본의 편차를 기반으로, 표본표준편차와 원하는 유의 수준에 따라 오차의 크기가 결정된다. 즉, 이는 데이터가 참값으로부터 얼마나 퍼져 있는 분포로 추출되는지를 의미한다. 우연오차를 얻어내는 방법은 실험 설계에 따라 1-자유도 반복 측정이 될 수도, 2-자유도 선형 회귀를 통해 얻을 수도 있다. 이는 6장에서 자세히 논의하도록 한다.

반면 반복 측정으로 제거될 수 없는 (검출될 수 없는) 성분의 오차를 계통오차라 한다. 이는 주로 실험계의 Misalignment나 Calibration error 등으로부터 발생하며, 이론이 너무 이상적이거나 단

순하여 실험 결과를 제대로 설명하지 못하는 경우에도 발생한다. 실험자의 실수나 Data Acquisition 상의 한계가 이러한 오차를 발생시킨다. 이런 종류의 오차는 주로 실험 장비 및 측정시료, 실험이 기반으로 하는 이론의 한계 등 실험계 자체의 성질/한계에 의해 발생하므로 계통(Systematic) 오차라 칭한다.

참값을 모르면 계통 오차를 얻어낼 수 없으나, Calibration을 이용해 계통 오차의 크기를 줄일 수 있다. 또한, 같은 실험 방법이나 장비를 이용하여 공칭값을 알고 있는 시료에 대한 측정을 시행하면 계통 오차의 대략적인 크기나 상/하한을 추산할 수 있다. 계통오차는 주로 계측기기의 제조사에서, 시험/교정 과정에서 얻어진 값을 제공하며, 직접 실험계를 구성할 경우 이러한 계통 오차를 최대한 배제할 수 있도록 실험을 설계하는 것이 중요하다.

### 5.3 불확실도의 종류 - 발생 원인에 따른 분류

실험적 측정에서 오차를 일으킬 수 있는 요인에는 여러 가지가 있다. 그 중, 첫 번째로 이론오차를 고려해 볼 수 있다. 이론오차는 계통오차를 유발하는데, 이는 이론이 실험계를 정확하게 반영하지 못하는 경우 나타난다. 예를 들면, 단진자의 주기는 등시성(진폭에 무관)을 가지는 것으로 계산되지만, 실제 진자는 진폭이 커질수록 제곱 항 이상의 보정항이 반영된 주기를 가지게 되어 약간씩 주기가 증가하게 된다. 이를 선형화하여 회귀할 경우, 측정치는 측정 오류로부터 얻어지는 우연 오차 이외에도 회귀식으로부터의 일정한 틀어짐을 가지게 되는데, 이것이 이론 오차의 영향이다.

물리학에서 쓰이는 많은 모델이 근사 모델에 해당한다. 그럼에도 이들은 잘 맞는 경우가 많다. 이때 주의해야 할 점은, 근사가 valid

한 영역 (domain)을 잘 살펴서 모델을 적용해야 한다. 동시 진자 모형은 작은 진폭(가급적 15도 이하의 각진폭)을 가지는 경우에만, 고전역학적 모델은 광속의 1/20 이하로 운동하는 계에 대해서만 적용하는 것이 바람직하다. Model의 Validity를 보장할 수 없는 Domain에서 실험을 진행할 경우, 이론 오차의 영향은 무시할 수 없게 되며 이는 회귀모델로부터의 비선형 편차로 이어지게 된다.

다른 요인으로는, 주변 환경의 요인이 있다. 예를 들어 자석을 이용한 실험을 할 경우, 원치 않는 자기장이 주변의 요인에 인해 잔존하는 경우가 많다. 이는 자기장 측정에서의 bias, 즉 계통 오차로 반영되는데 그 예로 지구 자기장, 주변의 강자성체 등의 요인이 원하는 조작변인에 의한 기여를 가릴 수 있다. 주변 환경에 의한 우연오차 역시 발생할 수 있는데, 주변의 전파나 교류 측정시 전원으로부터 흘러나오는 60 [Hz]의 노이즈 등이 이에 해당한다.

측정 기기에 의해 발생하는 계기 오차 또한 존재한다. 계통오차와 우연오차 둘 모두 발생할 수 있는데, 계통오차의 경우 Calibration이 안 된 경우 발생하며, 우연오차의 경우 계 자체의 불확정성이 의해 발생한다. 예를 들어 전자의 위치를 결정하기 위해 빛을 쏴서 관측을 한다면, momentum transfer가 일어날 것이며 곧바로 위치가 변한다. Commute하지 않는 물리량들 중 하나를 측정해야 하는 실험계에서는 필연적으로 관측 가능한 최소 불확정성이 0이 아니게 되는데, 이는 항상 피할 수 없는 Fluctuation, 즉 우연오차로 이어진다. 계측 기기 또한 실험 시료와의 상호작용으로 물리량을 측정하므로, 어떠한 경우에서도 계를 붕괴시키지 않고 실험 결과를 얻어낼 수 있는 측정은 존재하지 않는다.

이러한 미시적인 요인 이외에도 거시적인 우연 오차가 발생할 수 있다. 모든 측정 과정을 항상 동등하게 재현할 수 없기 때문에 측정기기의 분해능이나 거시 구속조건 하에서 실제 실험계는 많은

자유도를 가지고 다양한 초기 조건의 양상들로 이루어지게 되는데, 이는 우연 오차로 이어진다. 즉, 우연 오차는 검출하여 최확치를 오염시키지 않도록 제거할 수는 있지만, 0으로 만들 수는 없다.

계측기기에 의한 계통 오차 중 주목할 만한 것에는 Backlash에 의한 오차와 시료 오프셋에 의한 오차가 있다. 먼저 전자의 경우, 계측 기기의 Knob이나 기계적 체결부가 있는 경우 주로 발생하는데, 끼워맞춤 과정에서 손잡이와 축, 기어와 기어 사이 등에 기계적인 clearance가 있을 경우 knob을 돌려도 실제 축은 돌아가지 않는 매우 작은 구간이 존재하며, 그 반대의 경우도 가능하다. 이에 따라 실험자가 정해준 조작변인의 값과, 실제 실험계 내부에서 세팅된 조작변인의 값이 일치하지 않을 수 있다. 이 값은 Calibration을 통해 제거할 수 있을 것 같으나, clearance에 의한 불일치가 발생하는 것은 주로 장비 조작을 시작할 때 발생하므로, 스캔 방향을 바꾸면 불일치가 발생하는 시점이 바뀌게 된다. 이는 한 방향으로 측정을 하다가, 중간에 더 촘촘한 측정을 하기 위해 knob을 반대 방향으로 돌리는 순간 계가 가지는 offset이 변하는 일로 이어질 수 있다. 조작 변인의 모든 값이 한 방향으로 틀어진다면 이 영향은 선형화하여 회귀할 경우 절편의 오차로서 배제할 수 있으나, 몇몇 데이터가 반대 방향으로 틀어짐을 가진다면 이는 실험 자료의 선형성을 해치게 된다. 이러한 현상을 실험계가 가지는 backlash라 한다.

유의해야 할 또 다른 계기 오차로는 시료 혹은 측정장비의 오프셋이 있다. 대표적인 예시로, 훌 소자나 스트레인 게이지, 4선 체결식 선형 센서들의 경우 입력이 0일때 도선의 저항 등 여러 가지 요인으로 0이 아닌 출력이 얻어질 수 있으며, 이외에도 신호의 세기가 어느 정도 Threshold 이하일 땐 출력이 없는 경우도 있다. Photodiode를 이용하거나, OP-AMP를 이용한 여러 실험들은 이러한 문제를 고려하여 측정 영역을 잘 설정하여야 한다. 조작변인의 정의역과, 측정치의 치역을 여러 번의 예비 실험을 통해 실험이 잘 되는 구간을 먼저 찾고, 어느 구간에서 얼마나 많은 데이터 포인트를 어떤 간격으

로 얻을 것인지 측정 전에 검토하여야 한다. 이후 Backlash의 반영을 줄이기 위해 한 방향으로 스캔하여 측정을 진행하면 좋다. 만일 범위를 조절하는 정도로 오프셋을 피할 수 없다면, 측정하고자 하는 신호에 일정한 값의 Bias를 걸어 계측기기가 선형으로 작동하는 영역에서 실험을 진행하고 영점 보정을 진행할 수 있다. 이처럼 측정 장비의 비선형성이나 오프셋에 의한 오차가 존재할 수 있으며, 이는 계통 오차에 해당한다.

마지막으로, 실험자의 영향 역시 무시할 수 없다. 실험을 사람이 수행할 경우 완벽한 재현성을 갖추는 것은 사실상 불가능하다. 그러나 이들은 통계를 통해 배제될 수 있으나, 그럼에도 우연오차로 남아 측정의 정밀성을 해친다. 이러한 부분의 개선을 위해서는 컴퓨터를 이용해 전산화되고 디지털화된 DAq (Data Aquisition System)을 사용하면 개선할 수 있다. 그러나 특별한 주의를 기울여야 하는 것은 계통오차를 줄이기 위한 부분인데, 측정자에 의한 계통오차는 주로 Latency에 의해 발생 한다.

인간의 조건 반사는 약 100 [ms] 이하의 전달 시간을 가진다. 이 수치가 대략 어느 정도 스케일 인지 비교하기 위해 길이  $l = 30$  [cm] 인 진자를 생각해 보자.  $T = 400\sim500$  [ms] 정도의 주기를 가지는 진자의 주기를 측정할 때, 조건반사의 전달 시간 문제는 무시할 수 없는 스케일을 가진다. 따라서 이 정도 크기의 진동 주기를 육안으로 측정한다면 어마어마한 계통오차가 발생하게 되므로 줄일 방법을 검토하여야 한다. 1주기 대신 10주기의 측정을 시행할 경우, 응답 지연에 의한 계통오차를 주기당 1/10배 희석할 수 있다. 그러나 이러한 쳐방은 Q인자가 작아 여러 번 진동의 시간 간격으로 평균 주기를 얻기 어려운 실험계에는 적용하기 어려우므로, 주의를 기울여 적용하여야 한다. Q인자가 작을 경우, 여러 주기에 걸쳐 시간 간격을 측정하면 각각의 주기가 유의미한 차이를 가질 수 있으므로 이론 오차를 일으킬 수 있다.

뿐만 아니라 디지털 기기를 사용한 계측에서도, RF Signal을 다룰 경우에는 도선의 길이에 의해서도 무시할 수 없는 크기의 위상 지연이 발생하게 된다. OP-AMP의 경우 Slew Rate에 의해 고주파 특성의 제한이 발생하게 되는데, 이를 주의하기 위해 실험기기의 3dB Frequency를 측정하여 조작변인의 정의역을 주의 깊게 선정하거나, 혹은 사전 보정을 위한 Calibration 실험을 진행해 둘 수 있다.

이처럼 실험을 진행하며 발생할 수 있는 오차는 매우 많으므로 적절한 처방을 통해 오차의 발생 원인을 최대한 배제할 수 있는 실험 설계를 찾아 채택하여야 한다.

## 5.4 불확실도에 대처하는 방법

불확실도의 대처 방법은 오차의 성질과 종류에 따라 달라진다. 우연 오차의 대처는 실험기기를 정밀하게 만들으로서 줄일 수 있을 뿐더러, 평균하면 0이 되는 백색잡음(White Noize)의 경우에는 통계 처리를 이용해 희석할 수 있으므로 비교적 다루기 쉽다. 그러나 계통 오차는 피할 수 없기 때문에, 얻어지는 계통 오차가 원하는 물리량을 오염시키지 않도록 주의할 필요가 있다.

계통 오차의 제거는 근본적으로 캘리브레이션을 통해 이루어질 수 있다. 캘리브레이션의 세 가지 중요한 요소는 영점 조절, 선형성 확보, 감도 교정에 있다. 먼저 영점 조절은 입력 신호가 0일 때 계측 기기의 출력이 0 혹은 기준값이 되도록 개인과 오프셋을 조절하는 것이며, 마찬가지 방법으로 선형적인 입력 신호의 변화에 대한 출력 신호의 크기가 선형적으로 변하게끔 실험 구간과 개인, 오프셋을 조절함으로서 계측기기의 선형성을 얻을 수 있다. 이 때, 선형화된 계측 기기에 대해 크기를 알고 있는 기준 입력 신호를 넣어 주었을 때의

출력을 회귀하여 계측 기기가 입력 전압 변화에 대해 가지는 출력 전압 변화의 기울기, 즉 감도를 알아내고 보정할 수 있다.

캘리브레이션을 마쳐도 백래쉬에 대항하지 못하면 측정치의 선형성을 보장하지 못한다. 측정을 수행하기 전, 항상 예비 실험을 통해 측정기기를 사용하기에 합당한 정의역(Domain)과 치역(Range) - 어느 구간의 입력 신호에 대해 어느 구간의 출력이 얻어지는지 - 을 파악하고, 선형성이 보장되는 구간에서 실험을 한 방향으로 진행하여야 한다. 사전에 예비 실험을 통해 실험 데이터의 경향을 파악하고 있다면, 어느 구간에서 얼마나 많은 자료가 필요한지 알 수 있다. 극값을 포함하는 커브의 피크에서는 적어도 5개 이상의 데이터 포인트를 얻어 국소적으로 Quadratic Fit이 가능하도록 하며, 선형 구간에서는 8개 이상의 데이터 포인트를 얻어 선형 회귀가 가능하도록 한다. 또 변곡점 부근에서는 최소한 세 개 이상의 데이터를 얻을 수 있도록 유의한다. 만약 데이터의 개형을 반영하기 위한 만큼의 충분한 데이터 수를 측정하지 못하였을 경우, 스캔 방향을 거슬러 반대 방향으로 Knob를 돌려 중간 위치를 측정할 경우, 실제 읽히는 값은 조금 다른 방향으로 치우친 조작변인에 대한 실험계의 응답인 셈이므로 치우친 결과를 얻게 된다. 이러한 현상은 가변 저항을 돌려 보며 저항을 읽을 경우, 정확히 같은 눈금으로 되돌아와도 약간 다른 저항값을 가지고 있음으로부터 쉽게 확인할 수 있다.

실험 설계를 적절히 최적화하여 오차에 대처하였다면, 실험 자료를 분석할 때도 유의할 필요가 있다. 그 대표적인 주의 사항 중 하나로, 회귀 변수를 잘 선정할 필요가 있다. 예를 들면, 중력 가속도를 측정하기 위해 단진자를 이용한 실험을 진행한다고 하자. 단진자의 주기는 줄길이의 제곱근에 비례하는데, 줄의 길이 측정은 질량 중심의 위치, 물체가 매달린 매듭의 위치와 pivot 위치의 변화 등 여러 가지 요인에 의해 한 방향으로 치우친 계통 오차를 가진다. (약 전체 길이의 1% 가량) 이에 반해 초시계 혹은 포토인터럽트를 이용한 주기 측정은 매우 정교하게 수행할 수 있다. (물론 위상 오차를 줄이기

위한 주의도 필요하다 - 진자의 주기를 측정하는 지점을 최고점에서 진자가 방향을 바꿀 때가 아닌, 변위가 최소이고 속도가 최대인 지점에서 측정한다면, 정확히 진동의 중심점을 찾아내기 어려울 것이며, 이 때 측정 위치의 오차가 생긴다면 진폭에 따라 매번 진동을 카운트하는 지점의 위상이 변화하게 되고, 주기를 얻는 신호의 Duty ratio가 변하게 되며 이는 불확실한 측정으로 이어진다. 이러한 측방 역시 실험계의 선형성을 높이기 위한 고려에 포함된다.) 이러한 점을 고려할 경우, 데이터의 선형화를 주기와, 길이의 제곱근에 대해 분석하는 것보다 주기의 제곱과, 길이에 대해 분석하는 것이 낫다. 이러한 경우, 계통 오차에 의한 편향이 기울기에 반영되어 측정하고자 하는 물리량을 오염시키는 것을 막기 위해 후자를 택하면, 길이 측정이 가지는 계통 오차는 선형 회귀의 절편에 오차를 주지만, 기울기에는 (균일한 계통오차를 상정하면) 영향을 주지 않고 회귀선의 그래프가 약간 평행이동하는 결과로 이어진다. 반면 전자와 같이 길이의 제곱근을 회귀변수로 사용할 경우에는, 길이에 계통오차가 반영되는 것이 회귀 상에서 선형적으로 응답하지 않아 측정의 선형성을 해치고, 원하는 물리량의 불확실도를 키우며 최적치를 참값에서 더 면값을 얻도록 제한하게 된다. 이러한 현상은 도선의 저항에 민감한 실험 등에서 크게 나타나며, 각별히 주의하여 회귀 변수를 선정하여야 한다.



## 6.1 기기 눈금의 보간 (Interpolation)

아날로그 기기를 이용해 측정치를 얻어낼 때에는, 이산적으로 위치한 눈금들 사이에서 바늘이 가리키는 위치를 찾아 읽어야 한다. 이러한 과정은 이산량을 참조하여 연속량을 얻어내는 과정으로, 보간(interpolation)이라 한다.

비슷한 개념으로, 그래프 처리에서 내삽(interpolation)과 외삽(extrapolation)을 생각해 볼 수 있다. 유한한 수의 데이터 포인트는 정해진 값의 조작 변인에서의 종속 변인의 값만을 제공하지만, 측정하지 않은 (데이터 포인트의 위치들 사이, 혹은 그 범위 바깥에 있는) 조작 변인의 값에 대해 어떤 값의 종속 변인이 얻어질지를 유추하는 과정을 Interpolation(조작 변인이 측정의 정의역 내에 있는 경우) 혹은 Extrapolation(조작 변인이 측정의 정의역 밖에 있는 경우)이라 한다.

이처럼, 아날로그 계측기기를 이용하여 물리량을 측정할 때, 유한한 크기, 유한한 개수의 눈금 사이에 있는 값들을 읽어내기 위해 유판 보간법 (Interpolation)이 사용된다. 보간법은 측정자가 자료를 읽을 때 보장해 줄 수 있는 최선의 최학치 및 신뢰 구간의 추산을 의미하는데, 구체적으로는 ‘참값이 이 정도 신뢰구간 안에는 들어올 것’으로 보장해 줄 수 있는 범위를 실험자의 재량으로 제시하는 것이다. 신뢰구간을 크게 잡으면 참값이 확실히 들어올 것이라고 보장할

6.1 기기 눈금의 보간 (Interpolation) . . . . .	35
6.2 변인 통제 (Control of Variables) . . . . .	41
6.3 반복측정을 통한 불확실도의 검출 : t-분포와 선형 회귀 . . . . .	45
6.4 중심 극한 정리와 Student t-분포, 반복 측정으로부터 얻어지는 우연오차 . . . . .	46

수 있으나, 추산이 불필요할 정도로 부정밀해진다. 이는 실험자가 신뢰 구간을 제안하는 과정에서 불편함을 느끼게 되는데, 이를 바탕으로 더 좁은 신뢰구간을 잡아야 함을 실험자의 재량으로 결정하게 된다. 예를 들어 아래 그림에서 연필의 길이를  $2.9 \pm 0.7 [-]$ 로 제시할 경우 매우 넓은 신뢰구간을 가지므로 참값은 반드시 들어올 것으로 보장할 수 있으나, 불확실도가 너무 크게 잡혔다는 불편함을 느낄 것이다. 반대로 너무 작은 신뢰구간을 잡으면 참값이 신뢰구간 안에 포함될 확률이 너무 작다고 느낄 것이다. 이처럼 실험자/측정자의 재량에 따라, 참값을 포함할 것이라고 확신하되, 충분히 정밀하게 잡은 신뢰구간을 제시하게 되면 통계적으로 참값이 포함될 확률이 약 95%에 해당하며, 이는 유의확률 5%, 신뢰 수준  $2\sigma$ 에 해당한다. 이에 따라, 다음 그림의 예시를 측정하면  $L=2.9 \pm 0.1 [-]$  (Significance level  $2\sigma$ , by Interpolation) 으로 기입할 수 있다.



Figure 6.1: 아날로그 눈금의 보간

불확실도는 대체로 최소 눈금의 크기와 비슷한 규모를 가진다. 그러나 최소 눈금, 최소 눈금의 절반, 최소 눈금의 0.1~0.3배 등 문헌에 따라 다양한 기준이 제시되는데, 이는 눈금의 크기에 따라 다르다. 가령 밀리미터 간격의 자나 작은 각도기를 이용하여 측정할 때는 최소 눈금의 절반도 간신히 읽어 보장할 수 있지만, 센티미터 크기의 간격을 가지는 자를 사용하여 측정한다면 아무리 못 해도 최소 눈금의 0.1배까지는 육안 보간하여 불확실도를 제시할 수 있을 것이다. 이러한 점들은 계측 기기의 특성에 따라 결정되는 것이므로, 명확한 기준을 제시하기보다는 최소 눈금보다 작은 크기의 추산을 위해 보간하되, 불확실도는 실험자의 재량으로 참값이 포함될 확률이 약 95%가 되게끔 최선의 범위를 제시하는 것이 합당하다.

한편 아날로그 기기의 눈금을 읽을 때, 눈금이 실제 조작변인

과 일치하지 않는 경우도 있다. 이러한 오차 요인을 배제하기 위한 실험 설계가 필요한데, 앞서 언급한 바와 같이 여러 기계 요소에는 clearance가 존재하며, 아날로그 기기의 눈금 중 특히 조작변인의 값을 읽을 때에는 backlash가 존재하므로 한 방향으로 읽어 눈금을 보간하여야 한다. 이러한 현상을 주의해야 할 상황의 예시로는, 피크의 최댓값이나 최솟값 등의 위치를 읽어내기 위해 한 방향으로 눈금을 돌렸다가, 극값을 지나쳐 다시 반대 방향으로 돌리게 되면 눈금의 위치가 내부적으로 미세하게 바뀌어 극값을 같은 위치에서 재현할 수 없게 된다. 이러한 일을 피하기 위해서는 예비 실험을 통해 대강의 특별한 위치들을 파악하고, 한 방향으로 적절한 간격으로 (변화가 적은 곳에서는 듬성듬성, 변화가 크거나 극값 근처에서는 촘촘히) 데이터를 수집하여야 한다. 만약 반복된 Backlash의 누적으로 극값에 다시 도달할 수 없다면, 충분히 Knob을 멀리 밀었다가 다시 극값 근처로 한 방향으로 천천히 가져오는 방법이 도움이 될 수 있다.

보간법은 기본적으로 아날로그 계측기기의 불연속 눈금으로부터 연속적인 최확치를 추산하기 위해 육안 관측 시 사용되는 방법인데, 현대적인 물리 실험 설계에서는 아날로그 계측기기를 육안으로 읽는 것보다 디지털 데이터 수집을 사용하는 것이 보다 일반적이다. 이러한 관점에서, 계기의 측정 오차를 읽어내는 원리는 육안 보간법과 어느 정도 유사성이 있으므로, 이 절에서는 디지털 기기와 시계열 데이터의 불확실도를 추산하는 법에 대해서도 간략히 언급한다.

디지털 기기의 불확도 추산은 기본적으로 육안 보간과는 결이 다르다. 두 가지 큰 차이점은 (1) 디지털 기기의 경우 최소 눈금 이하의 값을 읽어낼 수 없다는 점이며, (2) 디지털 기기는 일반적으로 측정의 불확실도 (Noize에 의해 발생)의 크기에 비해 최소 눈금의 간격이 매우 작다는 점이다. 이러한 점을 이용하여, 측정치를 샘플링하면 노이즈의 크기, 즉 불확실도의 크기가 충분한 유효 숫자로 얻을 수 있다. 이는 눈금 사이의 값을 읽는 것이 아닌, 눈금보다 훨씬

더 큰 크기로 발생하는 오차를 통계적 기법을 이용해 얻어내는 것에 가깝다. 이 때문에, 디지털 계기를 이용한 불확도 추산은 아날로그 기기의 육안 보간과는 다소 차이가 있다.

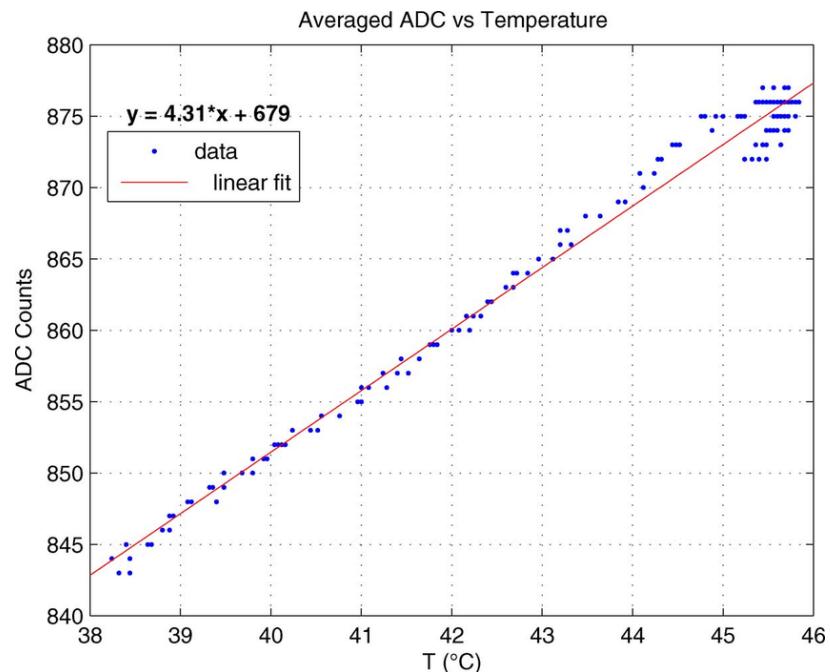


Figure 6.2: 디지털 기기의 노이즈

컴퓨터를 사용해 디지털 측정치를 샘플링할 경우, 측정 자료의 평균값을 바탕으로 통계 처리를 이용해 오차의 크기를 얻을 수 있다. 이는 본질적으로 반복 측정에 의한 우연오차의 검출에 가까우며, 이는 6.4절에서 조금 더 자세히 다루게 된다. 반면 육안으로 디지털 기기의 출력값을 읽는 경우, 가중치를 제대로 반영하여 통계 처리를 내기 힘들다. 이 때 손쉽게 적용할 수 있는 방법은 눈으로 볼 수 있는 데이터의 혼들림을 이용한 방법이다. 바로 데이터의 최저점과 최고점의 값을 읽는 것인데, 이렇게 할 경우 혼들리는 데이터를 모두 포함할 수 있는 범위를 제시해 줄 수 있다. 이 때 불확실도는 이 범위의  $1/2\sqrt{3}$ 으로 산출할 수 있는데, 이는 측정 표본이 균일 분포로 얻어진다고 상정하고, 제곱 편차를 이용해 분산을 구하면 얻어낸 구간의 총 길이의  $1/2\sqrt{3}$ 만큼의 표준 편차를 얻는다. 이는  $1\sigma$  유의수준과 대응되는 불확실도로서 사용될 수 있는 값이다. 오차의 전파에서 사용되는 최대-최소 기법과 결을 같이 하는 방법이다.

혹은 계측기기나 시료에 대해 제조사 제공 불확실도를 확인할

수 있는 경우 이를 고려하여 반영하여야 한다. 이 경우 계통 오차가 기입된 경우도 있으며, 신뢰할 수 있는 유효 숫자 자릿수가 제시되거나, 상대 오차가 절항의 색 피와 같이 기입되어 있는 경우도 있다.

시계열 데이터의 경우 눈금의 흔들림을 읽어낼 수 없는데, 디지털 초시계를 사용하던, 아날로그 초시계를 사용하던 눈금은 측정을 종료하는 순간 멈춘다. 눈금은 흔들리지 않으며, 측정된 값의 우연오차를 얻으려면 또 다른 차원에서 시간 간격을 측정한 눈금의 흔들림을 읽어야 하는데, 이는 본질적으로 반복 측정을 통한 불확실도 검출과 결을 같이 한다. 한편, 이러한 방법으로 얻어진 데이터를 회귀 분석할 경우, 등분산 가정을 해치지 않는 규모로 불확실도가 각 데이터 포인트들에 대해 어느 정도 비슷하게 얻어지는지 확인해 보는 등 주의를 기울여야 한다. 대부분의 경우, 어느 정도 재현성이 높은 실험에 대해서는 회귀 분석을 할 경우 각각의 데이터 포인트를 여러 번 측정하여 분석할 필요는 없으나, 재현성이 낮거나 노이즈가 큰 경우, 노이즈를 줄이고 우실험계 자체의 우연오차를 줄여 회귀분석에 용이하게 하기 위해 시간 평균이나 반복 측정을 한 데이터를 이용해 회귀하는 경우가 있다.

예를 들면, 양초의 연소 시간을 계산하는 경우를 생각해 보자. 양초의 연소는 매우 느리다. 그 긴 시간 동안 연소 속도에 영향을 줄 수 있는 요인은 주위의 바람, 기온, 양초의 굵기 변화 등등으로 매우 많은 오차 요인이 존재한다. 이러한 경우, 양초 연소시간의 측정치는 재현성을 확보하기 어렵다. 즉, 매 측정마다 상당히 다른 값이 얻어질 것이며, 이는 매우 큰 우연오차 혹은 큰 잡음을 가지는 실험계임을 의미한다. 이러한 경우 통계적 추출 과정을 통해 우연오차의 크기를 줄일 수 있다. 표본을  $N$ 번 추출하면 표본집단이 따르는 통계의 분산은 표본분산  $s^2$ 의  $1/N$ 이 되므로, 우연오차, 즉 잡음의 영향을 배제한 채 회귀를 진행할 수 있다. 이 경우, 예를 들어 양초의 길이가 1, 2, 3, 4, 5, 6, 7인 경우에 대해, 각각 같은 길이의 양초를 10번 반복 시험하여 얻은 평균치와 표본오차를 회귀에 사용할 하나의 데이터 포인트의

최확치와 불확실도로 사용한다. 이 때 총 70번의 측정이 이루어지며, 회귀에는 7개의 데이터가 사용되는 것이다. 이러한 과정에서 신뢰수준을 섞어 쓰지 않도록 주의해야 한다. 이처럼, 반복 측정을 회귀와 함께 사용하는 경우에는 까다로운 오차 고려가 필요하므로 특수한 경우가 아니라면 굳이 회귀의 한 데이터 포인트들을 반복 측정할 필요가 없다. (노이즈가 크거나, 조작변인에 의한 종속변인의 변화가 작아 노이즈에 가려지는 경우에 해당)

한편, Random walk의 결과로 얻어지는 위치 분산을 시간에 대해 회귀하는 경우를 생각해 보자. Random Walk는 위치의 기댓값은 항상 원점이지만, 위치의 분산은 시간이 지남에 따라 점점 커지게 된다. 한 입자의 브라운 운동을 Track하여, 시간에 따른 위치 정보를 알아 내고 이를 바탕으로 변위 분산을 계산하면, 시간에 비례하는 값을 얻을 것으로 이론적으로 예측된다. 이 경우 변위 분산은 다소 특별한 방법으로 기술된다. 시간에 따라 계의 분산이 증가하지만, 단 한 번의 측정으로 분산을 얻어낼 방법은 없다. 이 때, 임의성이 충분히 보장된다면 변위의 평균은 원점으로 상정할 수 있으며 이 경우 변위 편차는 변위와 일치하게 된다. 이때 분산을 얻기 위해 편차의 제곱을 시간 평균 내는 것을 검토할 수 있는데, 이 경우 시간  $t$ 일 경우의 분산이  $t^n$ 에 비례하는 power law를 따른다고 하면, 이를 시간적분한 뒤 다시 시간으로 나누면  $t^n/(n+1)$ 에 비례하는 거동을 가지게 되므로, 반복 측정을 통해 각각의 시간에서의 편차 제곱 기댓값을 구하지 않아도, 한 번의 시간 누적 측정으로 기존의 power law를 유지하는 회귀량의 채택이 가능하다. 이 때  $t^n$ 을 조작 변인으로, 변위 제곱의 시간평균을 종속변인으로 잡을 경우 선형성이 보장되어 이론적으로 예측된 random walk의 dispersion에 대한 정보를 회귀선의 기울기로부터 얻을 수 있다. 이는 재현성이 떨어지는 실험을 단 한 번 수행하여 평균 처리를 함으로서 우연오차를 줄일 수 있다는 의의를 가진다. 이 경우 시간  $t$ 동안  $N$ 번의 측정을 하였을 경우, 우연오차의 크기는  $1/\sqrt{N}$ 로 감소하게 된다. 이는 노이즈가 많고 random한 거동을 선형적인 관계식으로 회귀해낼 수 있게 만드는 신기한 기법이다.

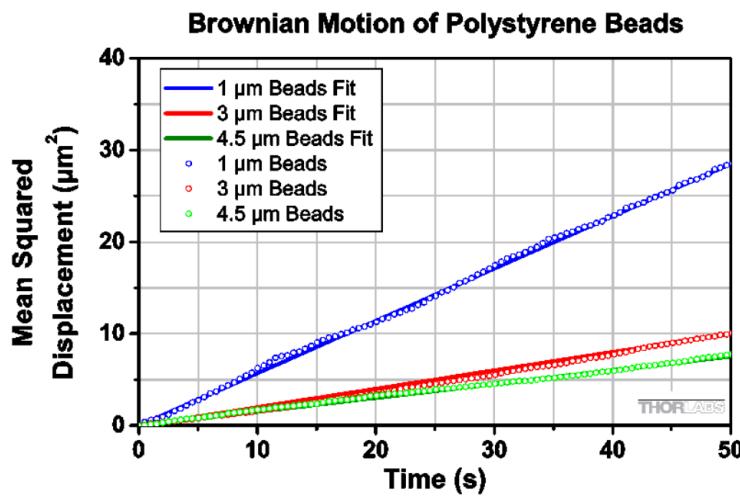


Figure 6.3: Random Walk의 분산 회  
귀

## 6.2 변인 통제 (Control of Variables)

변인 통제 (Control of Variables)는 실험 변인들을 적절히 제어하여 원하는 결론을 도출할 수 있도록 실험을 설계하는 기법을 말한다. 실험 변인은 실험계에 영향을 주거나, 그 결과로서 얻어지는 변인들로, 조작변인, 통제변인, 종속변인으로 나뉜다. 조작변인은 실험계를 제어하기 위해 변화시키는 변인을 의미하며, 종속변인은 조작변인의 변화에 영향을 받아 변화하는 실험계의 응답이다. 통제변인은 조작변인 이외의 종속변인에 영향을 줄 수 있는 변인으로, 종속변인의 변화가 오로지 조작변인의 변화에만 의존하도록, 가능한 한 모든 통제 변인이 실험이 진행되는 동안 일정하게 유지되게 실험을 설계하여야 한다.

실험계는 추출된 데이터들의 모집단이다. 여기서 측정을 통한 자료의 추출을 시행하면, 측정 표본인 데이터 셋이 얻어진다. 이 때, 측정하고자 하는 물리량의 참값은 실험계에 대한 Ensemble 평균으로 기술할 수 있는데, 모집단에 대한 정보를 직접 알아낼 수는

없으므로, 표본에 대한 통계를 최대한 Ensemble 평균이 이루어지는 과정과 유사하게 만들어 주어야 한다.

Ensemble 평균을 얻기 위해서는 많은 자유도를 갖고 결코 동등하게 재현될 수 없는 수많은 미시상태들을 가지는 실험계가, 적어도 일련의 거시적인 구속 조건 / 초기 조건들을 만족시키게 할 수 있는 실험적 제어가 필요하다. 만일 매 추출이 동등하지 않다면, 서로 다른 거시 상태에서 출발하여 얻어진 측정치들이 섞일 것이므로 표본 평균은 Ensemble 평균으로부터 멀어지게 된다. 이 때문에 반복 측정을 통해 실험 데이터의 모집단을 추정해내기 위해서는, 적절한 실험 설계를 통해 모든 측정치가 동일한 초기 조건으로부터 반복 재현되어 얻어졌음을 상정할 수 있도록, 즉 측정치의 표본평균을 Ensemble 평균으로 간주할 수 있을 정도의 정교함이 필요하다.

측정을 통해 얻어지는 값은, 실험계를 구성하는 몇몇 거시 변수에 대한 함수로 기술될 수 있다. 물론 불확실도의 존재로 인해, 초기 조건을 기술하는 실험 변수와 각각의 측정치는 일대일 대응이 이루어지는 수학적 함수가 아니므로 (= 같은 조건에서 여러 가지 측정값이 나올 수 있으므로), 측정을 통해 추산한 실험계의 양상을 평균이 각각의 실험 변수에 대한 다변수함수로 기술된다고 하는 것이 정확하다.

이때 실험계를 조절할 수 있는 변수들은 정의역 (Domain) 상에 놓이며, 이들의 영향을 받아 측정할 수 있는 변수는 치역 (Range)로 Mapping되어 종속 변수가 된다. 이때, 다변수함수를 통계적으로 분석하기 위해서는 한 가지 변인만을 바꾸어 주며 나머지 변인을 일정하게 통제해 주는 것이 중요한데, 이러한 목적으로 실험계에 영향을 주는 (= 정의역에 놓이는) 여러 가지 실험변수 중 원하는 한 가지를 조작 변인으로, 나머지 전체를 통제 변인으로 놓게 된다.

명확한 물리적 상관관계를 얻어내기 위해서는 항상 조작 변인의 수는 0개 혹은 1개여야 한다. 첫 번째 경우는 완벽히 동일한 실험계를 적정 횟수 반복 재현 및 측정하여 얻은 측정치들의 표본을 통계적으로 기술하는 것이며, 반복 측정에 의한 오차 검출을 수행할 수 있다. 두 번째 경우는 조작 변인을 바꾸어 가며 그에 따른 실험계의 응답을 종속 변인을 관측함으로서 읽어내는 것으로, 이들의 변화 경향을 회귀 분석을 통해 파악함으로서 실험계의 응답 특성을 얻을 수 있다. 반복 측정으로는 계통오차가 잔존할 경우 제거하거나 검출할 수 있는 방법이 없으나, 회귀를 수행할 경우 계통오차로 인한 균일한 Shift 등이 있는 경우에도 조작 변인의 변화에 의한 종속 변인의 응답만을 검출해 낼 수 있다. 이러한 이유 때문에, 1자유도 실험 (0개의 조작 변인)인 반복 측정보다 2자유도 실험 (1개의 조작 변인)인 회귀 분석이 실험 설계에서 보다 선호되는 형태이다.

이를 통계 처리가 아닌 실험계 차원에서 구현한 것의 대표적인 예시가 Lock-In Detection이다. 자연에는 여러 가지 주파수의 신호와 잡음이 섞여 있는데, 측정하고자 하는 종속 변인이 여러 가지 주변 환경에 영향을 받는다고 할 때, 정의역에 있는 여러 변인 중 실험자가 원하는 조작 변인만 특정 진동수로 진동시켜 주게 되면, 종속 변인의 측정된 시계열 (temporal) 성분 중 조작 변인을 떨어 준 진동수와 일치하는 성분만 걸러내면 통제 변인의 제어가 어려운 경우에도 조작 변인에 의한 영향만을 걸러낼 수 있다.

한편, 조작 변인과 종속 변인은 인과율로 인해 얹혀 있는 경우가 많다. 가령, 순수하게 물리적 의미가 전혀 없는 수학적인 계를 생각해 볼 때, 단위원의 방정식  $X^2 + Y^2 = 1$ 에서 X를 바꾸면 Y도 변화하지만, Y를 바꾸어도 X가 변화한다. 이 때 X를 조작 변인으로 잡을 수도, Y를 조작 변인으로 잡을 수도 있다. 이는 X와 Y 사이에 인과 관계가 없기 때문인데, 대부분의 물리 현상의 경우 변인들 사이의 인과 관계가 존재하는 경우가 많다. 예를 들면, 시간에 따른 속도를 측정하였다고 할 때, 시간에 따라 속도가 변화하는 것은 측정할 수

있지만, 원하는 속도에 도달할 시간은 측정할 수 없다. 물론 물리 법칙을 통해 예측할 수는 있지만, 실험을 통해 얻어낼 수는 없다. 이는 space-like interval에는 자를 가져다 댈 수 있지만, time-like interval에는 자를 가져다 댈 수 없기 때문이다. 초시계로 시간 간격을 측정할 수는 있지만, 이는 눈금을 읽는 측정과는 본질이 매우 다르며 인과율 역시 성립하지 않음이 자명하다.

비슷한 예시로 길이에 따른 주기를 측정하는 경우, 길이를 바꾸면 진자의 주기가 바뀌지만, 진자의 주기를 원하는 값으로 맞추기 위해 길이가 자동으로 맞추어져 조절되게 할 수는 없다. 이에 따라, 진자의 길이가 원인, 진자의 주기가 결과가 되는 강력한 인과 관계가 성립한다. 이는 물리 법칙이 어떠한 외부 파라미터의 변화에 대한 응답의 개념으로 기술되며, 정보의 전달에는 유한한 시간이 걸리기 때문에 time-like interval로 묶일 수 있는 변화의 전파가 방향성을 띠도록 제한된 것의 결과이다. 이 경우를 다시 살펴보면, 중력이 진자에 먼저 작용하고, 진자가 가진 길이에 따라 다른 회전 관성과 복원 토크를 받는 진동을 일으키는 것이므로 진동의 주기가 결정되기 위해서는 반드시 진자의 길이가 ‘먼저’ 확정되어 진동을 일으키는 데 반영되어야 한다는 것을 확인할 수 있다. 이 경우 주기는 ‘시간’의 개념인데도 종속 변인에 해당한다.

따라서, 물리 실험에서의 변인 통제는 항상 인과 관계를 고려하여 이루어져야 한다. 특히 회귀변수를 설정할 때에는 회귀의 성공률을 높이기 위해 변인 간의 인과 관계가 반드시 고려되어야 한다.

물론 가역적인 인과 관계도 있으며, 선형 저항체에서 전류와 전압의 관계 등이 그 예시이다. 이 경우 제어해 주기 편하거나, 보다 정밀한 제어가 가능하고 Backlash의 발생 우려가 적은 변인을 조작 변인으로 채택할 수 있다.

## 6.3 반복측정을 통한 불확실도의 검출: t-분포와 선형 회귀

앞서 언급한 바와 같이 측정을 통해 우연오차를 검출하는 과정은 조작변인의 수에 따라 반복 측정과 회귀 분석으로 나뉜다. 전자의 경우, 측정된 데이터 셋이 Student t-분포를 따를 것임을 상정하여 모분산의 추정을 수행하는 것이며, 6.4절에서 상세히 다룰 것이다. t 분포는 자유도 (측정 횟수 혹은 데이터의 개수에 의존)에 따라 분산이 달라지는데, 이때 유의 수준을 조절해 주면 신뢰 구간의 길이를 조절할 수 있다. 이에 따라 앞에서 논의된 것과 같이 필요한 유의 수준에 맞게 신뢰구간을 추정하고, 이는 우연 오차를 얻어내는 방법에 해당한다.

후자의 경우, 이론 모델을 선형화하여 회귀 분석을 수행하는 것이 일반적이며, 복잡한 모형의 경우 고급 전산 회귀 기법이 사용될 수 있다. 간단하게 선형화가 가능한 실험의 예시에는 지수적 변화를 가지는 시상수 감쇠 / 붕괴 등의 측정 및 power law를 따르는 경우가 있다. 선형화가 가능한 실험의 분석에 대해서는 7장에서 자세히 다룰 것이며, 이를 위해 기본적으로 종속변인의 등분산 가정 및 정확한 조작 변인 가정 하에 회귀선과 데이터 포인트 간의 편차를 최소화하기 위한 최소자승법이 사용된다. 등분산 가정이 유효하지 않은 등 몇몇 경우에는, 최소자승법 대신 Plot과 오차막대의 크기, 데이터의 분포를 이용해 회귀선의 정보를 육안 보간할 수 있다. 이러한 기법들의 사용 방법과, 사용해야 하는 경우를 오차의 크기 및 데이터의 분포, 실험의 특성 등으로부터 분석하여 다루게 된다.

## 6.4 중심 극한 정리와 Student t-분포, 반복 측정으로부터 얻어지는 우연오차

반복 추출된 데이터 셋은, 추출 횟수가 무한히 많아진다면 정규 분포로 수렴한다는 것이 중심극한정리의 결론이다. 그러나 30회 미만의 측정을 통해 얻어진 데이터 셋의 경우 보다 분산이 큰 t분포를 따를 것으로 기대할 수 있다. 이를 바탕으로 모분산의 추정을 시행할 수 있으며, 이는 실험계의 우연오차에 해당한다.

t 분포는 자유도에 따라 다른 분포를 가진다. 자유도( $\nu$  or df; degree of freedom)는 (추출한 표본의 수) - 1로 정의되며, 한 번의 측정으로는 최학치와 불확실도를 모두 알 수 없으므로 최소 2번 이상의 추출이 요구된다고 생각할 수 있다. 자유도가 높아질수록 t 분포는 정규 분포에 수렴하며, 점점 분산이 좁아진다. 이는 더 적은 측정으로는 우연오차 추산이 더 부정밀할 수 밖에 없음을 반영한다. 정량적으로는, t 분포표로부터 꼬리 확률을 얻어낼 수 있다.

이를 바탕으로 모분산을 추정하기 위해서는, 통계량을 표준화 할 필요가 있다. 참값(모평균)이  $\mu$ 인 분포에서 최학치가  $X_{be}$  일 때, 표본표준편차  $s$ 를  $\sqrt{N}$ 으로 나눈 값이 표준편차로 사용된다. 이 때, 표준화된 통계량  $T = \frac{X_{be} - \mu}{s/\sqrt{N}}$ 은 t-분포를 따르게 된다. 위 표에서 꼬리 확률  $q$ 에 상응하는 t인자는 통계량이 t보다 크거나 같을 확률이  $q$ 가 되는 위치 t를 의미한다. 이에 따라 양측 신뢰구간을 얻기 위해서는 확률  $q = \alpha/2$ 로 놓아야 하며, 이 때의 t인자가 신뢰 구간의 기준이 된다. 따라서 주어진 유의 수준 하에 통계량이  $[-t_{\alpha/2,\nu}, t_{\alpha/2,\nu}]$ 의 구간 안에 들어올 것으로 추산할 수 있다. 이를 다시 쓰면 다음 관계를 얻는다.

$$X_{be} - s \frac{t_{\alpha/2,\nu}}{\sqrt{N}} < \mu < X_{be} + s \frac{t_{\alpha/2,\nu}}{\sqrt{N}}$$

이에 따라 반복 측정을 통해 추산된 유의 수준 알파의 불확실도는

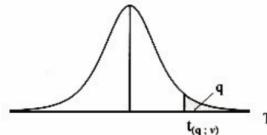
다음과 같다.

$$\sigma = s \frac{t_{\alpha/2; v}}{\sqrt{N}}$$

t 인자의 성질로부터 역시 유의확률이 증가할수록 불확실도의 크기는 줄어든다는 것을 알 수 있으며, 유의확률을 줄이기 위해서는 불확실도의 크기를 키워 신뢰 구간을 넓혀야 함을 확인할 수 있다.

[표 A-2] t-분포표

$$P\{T \geq t_{(q; v)}\} = q$$



자유도 <i>v</i>	꼬리 확률 <i>q</i>									
	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.792	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
$\infty$	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure 6.4: Student t-분포표



## 7.1 회귀 분석

불확실도를 포함하는 측정치로부터 물리 법칙을 얻어내기 위해서는 데이터의 경향을 통계적으로 검출해 내야 한다. 이 과정에서, 얻어진 데이터의 개형 혹은 변화 경향을 가장 잘 반영하는 간단한 이론 모델을 찾아내는 것을 회귀 (regression) 라 한다. 가령, 선형 회귀의 경우에는 선형 모델 ( $Y = A + BX$ )를 이용하여, 회귀 변수  $A, B$ 를 적절히 조절하여 데이터 셋 ( $(X_i, Y_i)$ )에 가장 근접하게 지나가는 회귀선을 찾는 과정에 해당한다. 데이터의 측정이 이상적일 수 없으므로, 모든 회귀는 근사적으로 수행되며, 오차를 최소화하는 방향으로 이루어진다. 이 때 회귀선과 데이터 포인트 사이의 편차가 얼마나 작아졌는지 가늠하기 위한 척도로서 제곱오차가 사용된다. 이를 토대로 근 평균 제곱오차를 최소화하는 회귀변수 (회귀변수는 회귀선을 유일하게 결정하기 위한 모든 파라미터를 의미하며, 선형 회귀의 경우 기울기와 절편을 말한다.) 를 결정하는 과정을 최소 제곱법 혹은 최소 자승법(least square method)에 의한 회귀 분석이라 한다.

## 7.2 최소 자승법

최소자승법은 제곱오차를 최소화하는 과정을 통해 데이터 셋의 경향을 가장 잘 반영하는 회귀 추세선을 얻는 방법을 의미한다. 이

7.1 회귀 분석 . . . . .	49
7.2 최소 자승법 . . . . .	49
7.3 각종 회귀 모델 . . . . .	54
7.3.1 선형 회귀 . . . . .	54
7.3.2 극값을 가지는 분포의 회귀 . . . . .	60

를 기술하려면 먼저 제곱오차에 대한 정의가 필요한데, 이는 주어진 데이터 포인트의 조작 변인 위치에서, 회귀 모형(추세선)이 옳다는 전제 하에 계산된 종속변인의 이론치(회귀선 위의 점에서의  $Y$ 값)로부터 측정치(데이터 포인트 상  $Y$ 값) 간 편차를 제곱하여 모두 합한 것이다. 선형 회귀의 경우, 이는 다음 식으로 표현된다.

$$E^2(A, B) = \sum_i (Y_i - A - BX_i)^2$$

제곱 오차를 위와 같이 정의하기 위해서는 두 가지 기본 전제 조건이 요구된다. 첫 번째는, 원하는 조작 변인을 정확하게 불확실도 없이 제어해 낼 수 있다는 것이며 ( $\sigma_x = 0$  : 따라서  $Y$ 방향 편차만을 고려), 두 번째는 각각의 데이터가 가지는 가중치가 동일하다고 가정한 것으로, 가중치가 동일함은  $Y$ 방향 불확도가 일정하여 모든 측정이 동등한 분산을 가지고 이루어졌다는 가정을 의미한다.(등분산 가정) 만약 등분산가정이 성립하지 않거나 조작변인의 오차가 큰 경우에는 변형된 기법을 적용하여야 하며, 이는 다음 절에서 언급하도록 한다.

제곱 오차의 값은 어떤 회귀 계수  $(A, B)$ 를 채택하느냐에 따라 다른 값을 가지며, 이를 최소화하는 회귀 계수를 찾는다면, 이는 데 이터들로부터 가장 편차가 작은 회귀선을 찾음을 의미한다. 이를 토대로 선형 회귀를 수행하면, 다음 두 식을 얻는다 :

$$\frac{\partial E^2}{\partial A} = 2NA - 2 \sum Y_i + 2B \sum X_i = 0$$

$$\frac{\partial E^2}{\partial B} = 2B \sum X_i^2 - 2 \sum X_i Y_i + 2A \sum X_i = 0$$

이를 연립하여  $A$ 와  $B$ 값을 얻으면 다음과 같다.

$$\begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$\Delta \equiv N \sum X_i^2 - (\sum X_i)^2$$

$$\begin{bmatrix} A \\ B \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & N \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} = \frac{1}{\Delta} \begin{bmatrix} \sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i \\ N \sum X_i Y_i - \sum X_i \sum Y_i \end{bmatrix}$$

$$A = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

∴

$$B = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

이러한 일련의 과정을 최소 자승법 (Least Square Method) 이라 한다.

이에 따라 얻어진 회귀선은  $Y = A + BX$ 의 형태로 구할 수 있다. 그렇다면 이 회귀선은 데이터를 얼마나 잘 반영하는가? 이 데이터 셋을 나타내는 회귀선은 유일한가? 이 질문들에 대한 답은 차례로 균평균제곱편차 (RMSE) / 상관계수 ( $R^2$ ), 그리고 회귀선의 불확실도 추산에 있다. 후자는 8장에서 자세히 다루고자 한다.

전자의 경우, 데이터 포인트가 얼마나 선형적인지 (혹은 얼마나 회귀 모형을 잘 따르는지)를 기술한다. 데이터 포인트의 분포가 선형적일수록 데이터 포인트가 회귀선으로부터 떨어진 편차가 작아질 것이며, 이들의 균평균제곱편차는 감소하고 상관계수는 선형성이 높을수록 (회귀선이 잘 맞을수록 = 정확할수록) 1에 수렴한다. (0인 경우 완전히 비선형) 이를 정의하기 위해 다음 관계가 성립한다.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{(S_{xy})^2}{S_{xx} S_{yy}}$$

이 때, 공분산  $S_{(AB)} \equiv \sum_i (A_i - \bar{A})(B_i - \bar{B})$ 을 정의한다. 한편,  $R^2$ 의 성질을 알아보기 위해 제곱합의 분해를 살펴보면 다음과 같이  $SSR, SSE, SST$ 를 각각 정의할 수 있다.

회귀제곱합 (Regression Sum of Square)  $SSR \equiv \sum_i (A + BX_i - \bar{Y})^2$ ; 자유도 1

오차제곱합 (Error Sum of Square)  $SSE \equiv \sum_i (Y_i - A - BX_i)^2$ ; 자유도  $(n-2)$

총제곱합 (Total Sum of Square)  $SST \equiv \sum_i (Y_i - \bar{Y})^2$ ; 자유도  $(n-1)$

$$SST = SSE + SSR$$

$R^2$  값이 선형 데이터 셋에 대해서 1이 됨은 손쉽게 보일 수 있다. 한편, 등분산 가정 하에 얻어진  $R^2$ 값은  $Y$ 의 불확실도, 즉 오차막대의 크기와 밀접한 관계를 가진다.

$$\frac{1}{R^2} - 1 = \frac{SSE}{SSR} = \frac{\sum_i (Y_i - A - BX_i)^2}{SSR} = \frac{(N-2)\sigma_Y^2}{SSR}$$

$$\text{where } \sigma_Y^2 \equiv \frac{\sum_i (Y_i - A - BX_i)^2}{N-2}$$

회귀가 측정 데이터와 얼마나 가까운지를 나타내는 또 다른 값으로는 RMSE (Root Mean Square Error)가 사용된다. 이는  $Y$ 방향 회귀 편차의 제곱합을 데이터의 수  $N$ 으로 나눈 뒤 제곱근을 취한 값으로, 0에 가까울수록 회귀가 정밀함(잘 들어맞음)을 의미한다.

회귀 변수의 선정 시에는 상술한 바와 같이 선형화와 계통 오차 문제를 조심하여야 한다. 계통 오차가 발생하기 쉬운 물리량들은 절편에만 영향을 주고 기울기에 영향을 주지 않도록 회귀 모델을 설계하고, 얻고자 하는 물리량은 가급적 기울기로부터 얻어내도록 하는 것이 불확도를 줄여 정확한 최적치 추산을 하기 위한 방법이다.

구체적인 예시를 들자면, 진자의 길이를 바꾸어 가며 중력 가속도를 측정하는 예시를 다시 기억해 보자. 이 경우에는 실의 길이가 실제 Pivot과 질량중심 간 간격과 다르므로, 이론에서 사용되는 값과 일정한 편차를 가지게 된다. 이러한 계통 오차를 배제하기 위해서는 실의 길이를 선형으로 놓고, 주기를 제곱으로 놓아 회귀하는 것이 선형성에 도움이 된다. 이론으로는 양 변을 제곱하거나, 로그를 취하는 등의 처리를 하고 나서 얻어진 회귀변수들끼리도 선형성이 보장되지만, 실제 실험계에서는 계통 오차의 존재가 어떠한 형태 (선형 /

제곱형 / 지수형 등)의 회귀 변수를 선정하였느냐에 따라 회귀의 선형성을 악화시키거나 개선할 수 있다.

뿐만 아니라, 실험계의 인과 관계에 주의를 기울여야 한다. 회귀 변수의 인과율이 섞이게 되면 회귀의 선형성을 악화시킬 수 있다. 예시로, 실제 실험계에서 조작 변인으로 사용된 물리량을  $X$ , 종속 변인으로 사용된 물리량을  $Y$ 라 하자. 이 때  $Y^2X$ 가  $X^2$ 에 비례하는 경우를 생각할 수 있다. (이러한 예시는 균일한 막대의 길이를 바꾸어 가며 주기를 측정하는 실험에서 유사한 형태를 찾을 수 있다.) 이 경우 조작 변인과 종속 변인이 섞인 채로 새로운 종속 변인으로 사용되는데, 대체로 조작 변인의 무분산 가정이 성립하는 영역에서는 조작 변인의 불확도가 종속 변인에 비해 작을 것이라고 기대할 수 있으며, 이에 따라 조작 변인을 종속 변인에 섞어 주는 것이 종속 변인을 조작 변인에 섞어 주는 것보다 더 바람직하다. 이는 단순히 불확실도의 크기 뿐 아니라, 어떠한 변수를 어떠한 용도로 사용하는지에 따라 선형성이 희석될 수 있음을 시사한다. 가령, 이론을 생각해보면  $YX = XY$ 의 관계 또한 당연히 보장되어야 하지만, 실험에는 불확실도가 존재하기 때문에 종속 변인의 불확실도보다, 조작 변인의 변화에 의한 종속 변인의 변화가 두드러지게 보일 정도로 충분히 큰 기울기를 가져야 한다. 이 경우 회귀 변수를 위와 같이 잡으면, 극단적인 경우이지만 당연히 기울기는 정확히 1이 나올 것이며, 조작 변인과 종속 변인 간의 상관 관계를 전혀 알아낼 수 없다. 이처럼 선형성을 보장하기 위해 이론적인 처리로 회귀변수를 바꾸어 줄 수 있지만, 이 결과로 실제 실험계의 조작 변인과 종속 변인 간의 상관관계가 희석될 수 있기 때문에, 조작 변인과 종속 변인을 섞을 때는 항상 주의하여야 하며, 어쩔 수 없는 경우라면 가급적 조작 변인을 종속 변인에 섞는 방향으로 검토하여야 하며, 선형성이 악화되었거나 기울기에 비해 오차막대가 크게 얻어진다면 가장 먼저 회귀 변수의 선정을 의심해 보아야 한다.

회귀 분석을 통해 얻은 회귀 변수에도 당연히 불확실도가 존

재하는데, 이는 측정치의 불확실도들로부터 물려받은 것이다. 이를 분석해 주어야 기울기 혹은 절편 등의 회귀변수로부터 원하는 물리량을 역산할 때, 물리량의 불확실도 추정을 위해 반드시 필요하다. 선형회귀변수의 오차는 8장에서 논의하고 고급회귀의 경우 10장에서 논의한다.

### 7.3 각종 회귀 모델

위에서 설명된 최소 자승법을 이용해 수행할 수 있는 여러 가지 회귀 분석에 대해 소개한다.

#### 7.3.1 선형 회귀

##### 절편을 갖는 선형 회귀 - 등분산 가정

이 경우 앞서 논의한 예시와 정확히 일치하며, 최소 자승법에 의해 다음 계수를 얻어낼 수 있다:

$$A = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

$$B = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

### 원점을 지나는 선형 회귀

간접 무늬의 위치, 개수에 따른 분석 등 조작변인이 정수인 실험 설계의 경우 정확하게 원점을 지나야 할 것이 이론적으로 강하게 요구되는 경우가 있다. 이 경우  $Y = BX$ 의 선형 회귀모형을 적용할 수 있다. 이 때 제곱오차를 계산하면

$$E^2 = \sum_i (Y_i - BX_i)^2 = \sum Y_i^2 + B^2 \sum X_i^2 - 2B \sum X_i Y_i$$

이고, 이를  $B$ 로 편미분하여 조건을 얻으면 다음과 같이  $B$ 를 얻을 수 있다.

$$\frac{\partial E^2}{\partial B} = 2B \sum X_i^2 - 2 \sum X_i Y_i = 0$$

$$\therefore B = \frac{\sum X_i Y_i}{\sum X_i^2}$$

### 분산이 다른 선형 회귀 - 조작 변인의 불확실도를 무시하는 경우

분산이 다른 여러 개의 데이터를 수집하여 회귀하는 경우, 각각의 데이터 포인트에 가중치를 부여하여야 한다. 분산이 다른 선형 회귀는 조작 변인의 불확실도를 무시하는 경우와 고려해야 하는 경우로 나눌 수 있는데, 전자의 경우 앞서 논의한 제곱 오차의 정의를 각각의 편차가 반영된 가중치를 부여하는 형태로 수정해 주어야 한다. 앞서 가중치가 분산의 역수로 얻어짐을 확인하였으므로, 제곱 오차는 다음 형태로 기술된다.

$$E^2 = \sum_i \frac{(Y_i - A - BX_i)^2}{\sigma_i^2}$$

이를 마찬가지로 회귀계수  $A, B$ 에 대해 편미분하면 다음 연립방정식을 얻는다.

$$\frac{\partial E^2}{\partial A} = 2A \sum \frac{1}{\sigma_i^2} - 2 \sum \frac{Y_i}{\sigma_i^2} + 2B \sum \frac{X_i}{\sigma_i^2} = 0$$

$$\frac{\partial E^2}{\partial B} = 2B \sum \frac{X_i^2}{\sigma_i^2} - 2 \sum \frac{X_i Y_i}{\sigma_i^2} + 2A \sum \frac{X_i}{\sigma_i^2} = 0$$

이에 따른 회귀계수의 최적치는 다음과 같다. (회귀계수의 불확실도 논의는 복잡하지만, 후술될 오차의 전파 기법을 이용하여 추산할 수 있다.)

$$\boxed{\begin{aligned} A &= \frac{\sum X_i^2 / \sigma_i^2 \sum Y_i / \sigma_i^2 - \sum X_i / \sigma_i^2 \sum X_i Y_i / \sigma_i^2}{\sum 1 / \sigma_i^2 \sum X_i^2 / \sigma_i^2 - (\sum X_i / \sigma_i^2)^2} \\ \therefore B &= \frac{\sum 1 / \sigma_i^2 \sum X_i Y_i / \sigma_i^2 - \sum X_i / \sigma_i^2 \sum Y_i / \sigma_i^2}{\sum 1 / \sigma_i^2 \sum X_i^2 / \sigma_i^2 - (\sum X_i / \sigma_i^2)^2} \end{aligned}}$$

상관계수의 정의는 달라지지 않는다. 이러한 회귀 분석은 계산량이 많으므로 Python이나 MATLAB 등의 스크립트 언어를 이용하거나 상용 데이터 분석 소프트웨어를 사용하는 것이 좋다.

### 조작 변수의 불확실도를 허용하는 선형 회귀

이 경우 최소자승법을 사용하는 것보다 Plot을 바탕으로 회귀선을 육안 추정하는 것이 간편하다. 물론 복잡한 계산을 감수한다면, 최소 자승법을 약간 변형하여 통계적인 회귀 분석을 시행하는 것도 가능하다.

이론식을 제시하면,  $Y$ 방향 편차뿐 아니라  $X$ 방향 편차까지 고려하여 제곱 오차를 제시할 수 있는데, 이는 회귀선 직선의 방정식으로부터 데이터 포인트가 떨어진 거리를 이용하여 얻어낼 수 있다. 이를

바탕으로 계산한 제곱 오차는 다음과 같이 바뀐다.(위와 마찬가지로, 상관계수  $R^2$ 의 정의는 바뀌지 않는다.)

$$E^2 \equiv \sum_i \left( \frac{|Y_i - BX_i - A|}{\sqrt{B^2 + 1}} \right)^2 \cdot \frac{1}{\sigma_{X_i}^2 + \sigma_{Y_i}^2} = \sum_i \frac{(Y_i - BX_i - A)^2}{(B^2 + 1)(\sigma_{X_i}^2 + \sigma_{Y_i}^2)}$$

이 때, 회귀계수를 얻기 위해서는 다음 연립방정식을 풀면 된다.

$$\frac{\partial E^2}{\partial A} = 0$$

$$\frac{\partial E^2}{\partial B} = 0$$

### 로그를 이용한 지수함수의 선형화 (Semi-log regression)

지수적 변화 경향을 가지는 변인에 대해서도 선형화를 거쳐 회귀를 수행할 수 있다. 지수적인 감쇠에서 반감기를 얻는 예시와, 지수적으로 발산하는 경우 대표적으로 이러한 방법의 적용이 유리하다.

$Y = A \exp(BX)$  꼴의 회귀 모형을 적용할 경우, 양 변에  $\ln$ 을 취하여  $\ln Y = \ln A + BX$ 의 형태를 얻는다. 이 때, 조작변인은  $X$ , 종속변인을  $\ln Y$ 로 채택하면 선형 회귀를 통해 기울기로부터 지수, 절편으로부터 계수를 얻게 된다. 그러나 이 경우 절편의 불확실도는 기울기에 비해 크며, 계통 오차에 의해 오염되기 쉽기 때문에 계수를 정확히 얻기 위해서는 한번 더 회귀하는 것이 바람직하다. 이 때 종속변인을  $Y$ , 조작변인을  $\exp(BX)$ 로 잡으면 기울기로부터  $Y$ 를 얻을 수 있다.

이 때 유의해야 할 것이, 조작변인을 복잡한 꼴로 잡을 경우 (부정밀한 측정치나 회귀 계수의 결과값을 사용해 구한 경우) 오차의 전파가 누적되어 불확실도가 큰 오염된 값을 가지게 될 수도 있다. 이 경우 최소자승법을 위한 가정인 ‘정확한 조작변인 가정’을 충족하기 어려우므로 위에서 언급된 특별한 방법으로 회귀를 수행하여야 한다. 그러나 이는 번거로운 일이므로, 보다 쉬운 방법으로는  $Y$ 의

우연오차가 작을 경우  $Y$ 를 조작 변인으로 잠시 놓고,  $\exp(BX)$ 를 종속 변인으로 놓아 회귀하는 것도 검토해 볼 수 있다. 한 변인이 다른 변인에 비해 우연오차가 압도적으로 클 경우, 인과율을 뒤집더라도 통계적으로 우연오차가 큰 변인을 종속변인으로 놓고 등분산 선형회귀를 수행하는 것이 더 유의미한 결과를 얻게 된다.

### Power Law를 따르는 물리량의 선형 회귀 (log-log / linearized regression)

제곱형 저항력, 흑체 복사 일률 등 Power law를 따를 것으로 기대되는 경우, 두 가지 회귀가 가능하다. 첫 번째로, power law의 exponent를 얻어내기 위해 log-log 회귀를 수행할 수 있다.  $Y = AX^n$  꼴(power law를 가짐)의 자료에 대해  $\ln Y = \ln A + n \ln X$ 이므로 종속 변인을  $\ln Y$ , 조작 변인을  $\ln X$ 로 하여 선형회귀하면 절편으로부터 비례 계수를, 기울기로부터 exponent를 얻을 수 있다. 그러나 앞의 경우와 마찬가지로 절편은 대체로 계통 오차에 의해 오염되는 경우가 많기 때문에, 선형화된 회귀를 수행하여 비례 계수를 다시 결정해 줄 필요가 있다. 이 경우 조작 변인을  $X^n$ , 종속 변인을  $Y$ 로 놓고 회귀하여 기울기로부터 비례계수  $A$ 를 얻는다. 이 때, 절편이 존재한다면 이는 앞서 수행된 log-log 회귀의 선형성을 해치게 되므로, 종속 변인의 오프셋을 보정한 후 다시 시도할 것을 검토할 필요가 있다. 이러한 회귀 분석은 물리학의 여러 법칙을 검증하고, 여러 중요한 물리 상수들을 실험적으로 결정해내는 핵심 테크닉이 된다.

그러나 슬프게도 자연계에는 한 가지 power term만 가지는 대신 여러 항이 섞여 있는 Polynomial behaviour를 가지는 경우가 많다. 대표적인 예시로 점성 저항과 충동 저항의 기여가 섞여 있는 저항력의 경우와, 복사 및 대류/전도에 의해 물체로부터 빠져나가는 열량을 생각해 볼 수 있다. 자세히 논하자면, 저항력의 경우 점성 저항

은 속도에 비례하는데, 충동 저항은 속력 제곱에 비례하므로, 속도가 낮을 때는 선형성을 얻을 수 있으며, 속도가 높을 때는 속도 제곱에 대한 저항력의 선형성을 얻을 수 있다. 발열체를 빠져나가는 열량은 주변과의 온도차에 거의 비례하는 전도/대류에 의한 열류량과, 온도 네제곱에 비례하는 복사에 의한 열유량의 합으로 이루어지는데, 이 경우 저온 구간에서는 선형적으로, 고온 구간에서는 온도 네제곱에 대한 선형성이 dominant한 경향을 가진다. 이러한 형태로 실험계의 정의역을 조절하여 비선형 다행 관계를 따르는 경우에도 근사적으로 Power Law를 따르는 것처럼 회귀 분석을 수행할 수 있다.

### 이론식을 알고 있는 변수의 비례계수 결정을 위한 선형 회귀

$Y = A + Bf(X)$ 의 형태로 복잡한 이론적 관계를 제시할 수 있는 경우, 이러한 이론적 예측이 실험적인 유의성을 가지는지에 대해 실험을 통해 확인해 볼 수 있다. 이때의 회귀는 조작 변인을  $f(X)$ , 종속 변인을  $Y$ 로 두어 선형 회귀를 수행할 수 있다. 이때 조작 변인의 계산 시에 오차의 전파를 잘 사용하고, 오차의 형태에 따라 적절한 회귀 방법을 상술한 것과 같이 잘 골라 사용하여야 한다. 이때의 기울기로부터 이론치와 실험치 사이의 비례 계수를 얻을 수 있으며, 절편으로부터 계통 오차를 걸러낼 수 있다.

한편  $A = 1$ 이도록 함수를 결정할 경우, 조작 변인은 이론적 예측치, 종속 변인은 실험적 측정치의 의미를 가지므로,  $A = 1, B = 0$ 으로 얻어질 것은 각각 이론이 얼마나 선형적인지에 대한, 그리고 이론이 얼마만큼의 오프셋(편차)을 가지느냐에 대한 검정에 사용할 후 있다.

### 7.3.2 극값을 가지는 분포의 회귀

#### 이차 함수 근사 회귀

이차 함수 회귀 (Quadratic Fit ;  $Y = A + BX + CX^2$ ) 는 극값을 가지는 비교적 대칭적인 분포의 극값 주변 변화 경향을 근사하여 안정성/불안정성을 평가하는 데 주로 사용된다. 퍼텐셜의 테일러 전개 시 2차항의 계수는 복원력 혹은 힘 상수의 개념을 갖기 때문에, 극값 근처에서 일부 구간을 잡아 데이터들의 Quadratic Fit을 수행하면 안정성에 관련된 정보를 얻을 수 있다. 어느 구간의 데이터를 선택해야 할지는  $R^2$ 이 유의미하게 작아지지 않도록 Rejection of Data 기준을 적용하도록 한다.

Quadratic Fit은 최소 자승법을 이용해 다음 행렬의 해로서 구할 수 있다. 이는 제곱 오차를 각 회귀 계수로 편미분해 얻어진 선형방정식들의 연립해이다.

$$\begin{bmatrix} \sum X_i^4 & \sum X_i^3 & \sum X_i^2 \\ \sum X_i^3 & \sum X_i^2 & \sum X_i \\ \sum X_i^2 & \sum X_i & N \end{bmatrix} \begin{bmatrix} C \\ B \\ A \end{bmatrix} = \begin{bmatrix} \sum X_i^2 Y_i \\ \sum X_i Y_i \\ \sum Y_i \end{bmatrix}$$

이 때, 상관계수  $R^2 = 1 - \frac{SSE}{SST}$ 로서 계산할 수 있다.

#### 통계적 분포를 가지는 자료의 회귀 분석

누적된 데이터 수가 많은 경우, 표본 집단이 특별한 분포를 따르는 경우가 있을 수 있다. 이럴 때에는 단순한 회귀 모형 대신 추출의 통계적 성질을 고려하여 특정 분포를 따르는 결과가 얻어질 것을 기

대해 볼 수 있다. 이 경우 Binomial 회귀, Poisson 회귀 등의 분포를 대입하여 전산 회귀를 수행할 수 있다.

### 회귀를 통한 Peak의 분석

물리학에서 Peak의 성질을 알아내는 것은 매우 중요하다. 피크의 높이, 위치, 폭/FWHM, 밑넓이 등을 알아내는 것은 계의 극값 근처 거동 (안정성, 복원력 등의 정보 포함)이나 특이점, 공명 관련 성질을 얻어내는 데 깊게 연관되어 있다. 이러한 경우 곡선의 형태에 따라 Gaussian 또는 Lorentzian 모형을 이용한 회귀가 사용된다. 열역학적 분포나 Random walk 등의 임의성이 보장된 분산 분포의 경우 대체로 Gaussian 분포를 따르며, 공명의 성질이 있는 피크의 경우 Lorentzian 분포를 따를 것을 기대해 볼 수 있다. 흡광도, 자기 공명 등 공명에 의해 측정치를 정량 분석하는 경우가 존재하는데, 이러한 경우 Lorentzian 회귀는 매우 중요한 역할을 한다.

$$\text{Gaussian 회귀} : f_{\mu,\sigma}(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

$$\text{Lorentzian 회귀} : L_{\mu,\Gamma}(X) = \frac{1}{\pi} \frac{\Gamma/2}{(X - \mu)^2 + (\Gamma/2)^2}$$

한편, 서로 다른 분산을 가진 Gaussian 분포 여러 개가 더해진 경우도 있을 수 있다. 이러한 예시는 트랩에 담긴 입자들이 에너지가 서로 다른 여러 상태를 점유하며 섞여 있는 경우에서 찾을 수 있다. 에너지가 큰 상태에 놓인 입자들은 Momentum Space에서 더 큰 분산을 가진 Gaussian Distribution을 가질 것이며, 에너지가 낮은 상태에 놓인 입자들은 더 작은 분산을 가질 것이다. 이들의 수밀도 분포를 측정하면 bimodal/multimodal 분포를 얻을 수 있다. 다음 그림은 ToF(Time-of-Flight) 기법 (Free Expansion을 이용해 입자가 가진 momentum distribution을 spatial distribution으로 변환하는 실험 설계)을 통해 바닥 상태와 들뜬 상태를 구분해 내는 회귀 분석의

예시를 보여 준다. 이 경우, 회귀 분석은 육안으로는 명확한 기준을 제시하기 힘든 자료에 대해 기존의 상과 구분해낼 수 있는 새로운 상이 공존하고 있음을 뒷받침하는 통계적인 근거를 제공해 줄 수 있다.

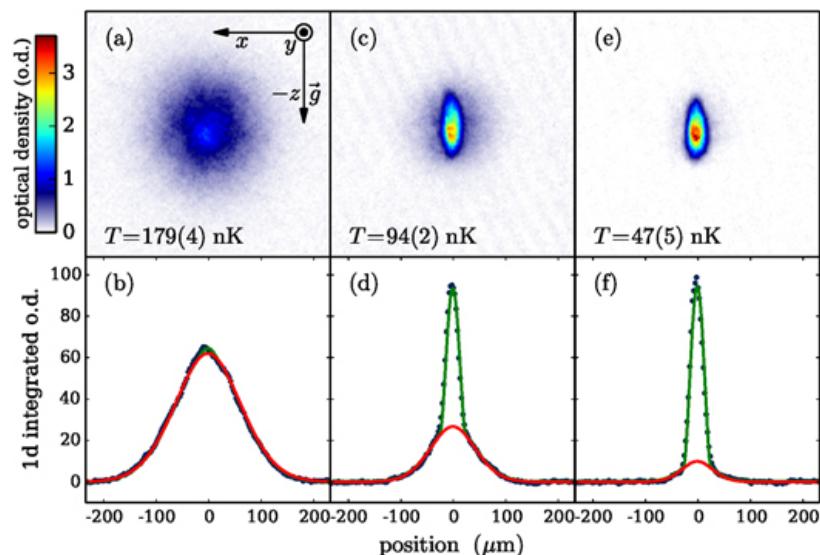


Figure 7.1: 온도에 따른 Bimodal Fitting - BEC

이외에도 여러 가지 응용 사례를 찾을 수 있다. 광학 기기의 분해 능을 결정하는 Rayleigh 기준의 경우 중심 위치가 다른 두 Gaussian 분포의 합으로 얻어지는 bimodal 분포를 따르는 것으로 간주할 수 있으며, 이러한 모델 하에 전산 회귀를 수행하면 Rayleigh 기준 안쪽의 육안으로 구분 불가능한 두 봉우리의 중심 위치 간 간격의 최적치를 추산해 낼 수 있다.

정리하자면, Lorentzian은 Peak 자체의 성질을 잘 반영하며, Gaussian은 Curve Tail의 성질을 잘 담고 있다. 또한 피크가 넓은 분포를 가지고 있는 경우 극값 주변에서의 이차 함수 회귀를 통해 안정성이나 곡률에 대한 정보를 추산해 낼 수 있다.

# 8

## 불확실도의 전파

### 8.1 다변수함수 관계를 가진 물리량에서 오차의 전파

8.1 다변수함수 관계를 가진 물리량에서 오차의 전파 . . . . .	63
8.2 회귀 계수의 오차 . . . . .	66
8.3 유효 숫자 . . . . .	69

다변수함수의 형태로 다른 물리량의 영향을 받는 어떤 물리량의 값을 정밀/정확하게 결정하려면, 그 값을 구하기 위한 다른 변수들 역시 낮은 불확실도로 측정되어야 한다. 이는 다변수 함수의 파라미터가 변하면 함숫값도 변화하므로, 파라미터들이 가지는 불확실도는 함숫값의 불확실도로 전파되기 때문이다.

전파된 불확실도의 크기를 구하기에 앞서, 유의 수준을 통일해 주어야 한다. 유의 수준이 바뀌면 불확실도의 크기도 바뀌기 때문에, 같은 유의 수준을 적용하여 파라미터들의 불확실도를 추산하고, 이를 토대로 얻어진 결과값의 불확실도 역시 같은 유의 수준을 갖는다. 이처럼 유의 수준을 해치지 않으며 결과값이 파라미터의 불확실도에 의해 얼마나 흔들리는지를 얻는 과정이 불확실도의 전파 추산이다.

불확실도의 전파는 조작변인과 종속변인 간의 기울기에 의해 결정된다.  $\epsilon - \delta$  방법을 이용해 수렴성을 검정하는 것과 같이, 함수의 파라미터가 좁은 폭의 불확실도를 가지면 이에 대응되는 함숫값도 좁은 폭으로 흔들리게 된다. 이 때문에, 조작 변인에 대한 일변수 함수로 표현되는 종속변인의 경우에는, 조작 변인의 불확실도에 기울기의 절댓값을 곱한 만큼의 불확실도를 가지게 된다.

기본적으로 불확실도의 덧셈 시에는 각각의 오차 요인이 모두 독립변수로 거동하는 것을 상정한다. 이에 따라 여러 오차 요인이 동시에 존재하는 경우 불확실도의 덧셈 방법은 피타고拉斯 합을 적용하게 되며, 이 경우 종속 변인을 조작변인에 대한 다변수함수로 기술한 뒤 불확실도를 알고 있는 각각의 조작변인으로 편미분하여 기울기를 얻을 수 있다.

$$\sigma_f = \sqrt{\sum_i \left( \frac{\partial f}{\partial X_i} \sigma_{X_i} \right)^2}$$

최확치 대비 불확실도의 비율이 작은 경우에는 제곱합을 하지 않아도 비교적 쉽게 전파되는 불확실도의 크기를 추산할 수 있다. 덧셈과 뺄셈의 경우, 부호에 관계없이 불확실도는 항상 커지는 방향으로 더해지며, 곱셈과 나눗셈의 경우 상대 오차, 즉 최확치 대비 불확실도의 비율끼리 선형적으로 더해진다. 한편 Power law를 따르는 경우 ( $X^n$ ) 상대 오차는  $n$ 배 증가하게 된다. 이외에도 함수식이 간단하게 알려져 있는 경우 기울기를 이용해 불확도를 추산할 수 있다.

그러나, 근의 공식과 같이 복잡한 형태의 식을 사용할 경우 오차의 전파 계산은 상당히 어려워지며, 변수마다 결과값을 변화시키는 방향(증가 / 감소)이 다르므로 매우 복잡한 과정으로 오차의 전파가 일어나게 된다. 또 다른 예시로는, 절댓값 함수의 원점 근처에서 조작변인의 오차는 종속 변인에 비대칭적인 오차 범위를 주게 되는데, 이는 monotonic하지 않은 관계에서 극값 근처의 조작변인을 취급할 때 주의해야 할 오류의 예시이다. 이러한 문제들을 종합하여 보면, 오차가 크고 관계식이 복잡한 경우, 오차의 전파 시 유의 수준을 보존하는 간단한 계산식을 찾기 어려워질 수 있다.

이러한 경우, 결과값을 가장 크게 만드는 조건으로 변수를 대입하여 얻은 최대 결과값과, 결과값을 가장 작게 만드는 조건의 변수에 대해 얻은 최소 결과값을 빼 얻은 차이를 이용해 우연오차를 추산할 수 있다. 이를  $2\sqrt{3}$ 으로 나누어 주면  $1\sigma$  유의 수준의 우연 오차를 가

하는데,  $1\sigma$  유의 수준의 우연 오차는 분포의 표준 편차와 같다. 이의 증명은, 복잡한 관계 속에서 파라미터 불확도들의 기여가 혼합되어 균일한 분포로 fade-out됨을 상정하여 분산을 계산하는 것이다.

균일 분포를 상정하고, 주어진 파라미터들의 신뢰 구간 내에서 얻을 수 있는 함숫값의 최솟값과 최댓값의 차이를  $L$ 이라 하자. 이때 신뢰 구간의 중심으로부터 떨어진 거리를  $X$ 라 하면  $X$ 는  $-L/2$ 에서  $L/2$ 까지의 값을 가진다. 균일 분포이므로  $X^2$ 의 평균을 구해 분산을 얻을 수 있는데, 이는  $L^2/12$ 이다. 이 때의 표준편자는  $(X_{max} - X_{min})/2\sqrt{3}$ 으로 얻어지며, 우연오차에 대한 추정에 해당한다. 이는 균일한 막대의 중심을 축으로 하는 경우의 회전관성 계산과 본질적으로 같다. 이때 유의할 점은 표준편자는  $1\sigma$  유의수준의 우연오차와 비슷한 크기를 가지며,  $\sigma$  표기 신뢰 수준이 두 배 높아지면 신뢰구간이 두 배 넓어지므로 (우연오차가 두 배 커지므로) 파라미터들의 유의 수준이  $1\sigma$ 가 아니라  $n\sigma$ 인 경우에도 최대-최소법을 이용해 추산한 함숫값의 불확실도 역시 약  $n$ 배가 되어  $n\sigma$ 에 해당한다는 것을 알 수 있다. 이는 이 방법이 유의 수준을 보존하며 불확실도의 전파량을 추산할 수 있는 방법임을 의미한다.

이러한 방법으로 불확실도를 얻어내는 방법을 최대-최소법이라 한다. 이때 최확치의 결정은 최대-최소 사이의 중심이 아니더라도, 각 파라미터들의 최확치를 대입해 얻어진 값을 채택하는 것이 바람직하다. 함숫값의 분포를 임의로 균일 분포로 상정한 것은 유의 수준을 보존하며 분산에 대한 대략적인 추산을 가능케 하지만, 최확치에 대한 대칭 분포임이 보장되지는 않는다. 따라서 최대-최소법을 사용한 결과는 불확실도 추산을 위해서만 사용하는 것이 바람직하다.

## 8.2 회귀 계수의 오차

회귀를 통해 원하는 물리량을 얻기 위해서 기울기나 절편의 값 을 사용한다. 이때 회귀선은 최적 추세선 하나만 있는 것이 아니라, 상한선부터 하한선까지 다양한 기울기와 절편을 가지는 여러 가지 추세선이 모두 가능하다. 이들은 제곱오차를 최소로 만들지는 않지만, 최적 추세선에 비해 제곱 오차가 그렇게 커지지는 않는 다른 ‘대안 추세선’이며, 최적 추세선과는 약간 다른 기울기와 절편을 가진다. 이들을 반영하는 방법은 기울기와 절편에 각각 최적치와 불확실도가 존재함을 알고 추산하는 것이다. 즉, 회귀 계수의 불확실도를 분석하는 방법이 필요하다.

회귀의 오차는 데이터 포인트의 오차로부터 전파된 것이다. 이는 (1) 데이터 포인트 각각이 얼마나 불확실하게 측정되었는지 (데이터 포인트의 우연 오차)와 (2) 데이터 포인트들이 회귀선으로부터 얼마나 멀리 떨어져 있는지 (회귀 편차)에 따라 각각 두 가지 오차 (1 : 측정 오차, 2 : 회귀 오차)로 구분할 수 있다. 등분산 선형 회귀의 경우에는 측정 오차가 회귀 오차에 비해 무시될 만큼 작음을 상정하고, 회귀 편차를 최소화하는 방향으로 추세선을 찾는다. (회귀선이 데이터에 가장 가깝게 다가가도록 : 최소 제곱오차) 반면, 데이터들이 상당히 선형으로 놓인 경우(이 경우 회귀선이 거의 데이터 포인트와 아주 가까운 지점들을 지나므로 회귀 편차가 작아진다.), 혹은 데이터의 측정 불확실도가 회귀 편차에 비해 많이 큰 경우, 데이터의 측정 불확도가 큰 데이터의 가중치를 줄여야 한다. 이 경우에는 등분산 가정을 폐기하고, 분산이 다른 선형 회귀를 사용하여야 한다.

측정 오차가 회귀 편차보다 큰 경우에는, 이분산 선형회귀를 하고 데이터 포인트들의 측정 불확실도가 회귀 오차로 전파되는 것을 간주한다. 반면 회귀 오차가 측정 오차보다 큰 경우에는, 데이터들은 대체로 정밀하게 측정되었으나 선형성이 매우 높지는 않아 데이터

들의 회귀 편차 (회귀선으로부터 떨어진 거리)가 큰 것으로, 이 경우에는 회귀 편차가 회귀 오차로 전파된다. 이러한 개념은 9장에서 오차 막대의 plot 시 주의 깊게 살펴야 한다.

먼저, 간단하고 가장 널리 사용되는 것은 후자에 해당한다.(디지털 측정이 이루어진 경우 측정 오차는 매우 작아지므로 회귀 편차가 회귀 오차의 주 원인이 된다.) 등분산 선형회귀의 경우, 회귀계수  $A, B$ 가 모두  $Y_i$ 와  $X_i$ 로 표현된다. 이 때, 오차의 전파를 고려하면  $X_i$ 는 무분산 가정하고,  $Y_i$ 는 등분산 가정하므로 회귀 계수를  $Y_i$ 로 편미분한 기울기를 제곱하여 합한 뒤 제곱근을 취하고 (피타고拉斯 합), 그에  $\sigma_Y$ 를 곱하면 각 회귀계수의 불확실도를 구할 수 있다.

$$\sigma_A = \sqrt{\sum \left( \frac{\sum X_i^2 - X_i \sum X_i}{N \sum X_i^2 - (\sum X_i)^2} \right)^2 \sigma_Y^2} = \sqrt{\frac{\sum X_i^2}{\Delta}} \sigma_Y$$

$$\sigma_B = \sqrt{\sum \left( \frac{NX_i - \sum X_i}{N \sum X_i^2 - (\sum X_i)^2} \right)^2 \sigma_Y^2} = \sqrt{\frac{N}{\Delta}} \sigma_Y$$

한편, 직접 제곱 합들을 계산한다면 이 공식을 그대로 사용해도 좋지만, 등분산 선형 회귀는 계산기나 Excel 등의 범용 스프레드시트로도 손쉽게 수행할 수 있으므로  $A$ 와  $B$ ,  $R^2$ 값만을 이용하여 회귀계수의 불확실도  $\sigma_A, \sigma_B$ 를 추산할 수 있다면 매우 편리하게 사용할 수 있다. 이를 위해서는 앞서 얻은 상관계수와  $\sigma_Y$ 의 관계를 이용할 수 있다.  
(7장 참고)

$$SSR^2 = \sum_i (\bar{Y} - A - BX_i)^2 = N(\bar{Y} - A)^2 + B^2 \sum_i X_i^2 + 2B(A - \bar{Y}) \sum_i X_i$$

$$\text{한편, } A - \bar{Y} = -B\bar{X} = -B \frac{\sum_i X_i}{N}$$

$$\therefore SSR^2 = \frac{N \sum_i X_i^2 - (\sum_i X_i)^2}{N} B^2 = \frac{\Delta}{N} B^2$$

$$\sigma_Y = \sqrt{\frac{\frac{1}{R^2} - 1}{N-2}} \sqrt{SSR} = \sqrt{\frac{\frac{1}{R^2} - 1}{N-2}} \sqrt{\frac{\Delta}{N}} B$$

이를 기반으로 다음 관계를 얻을 수 있다. (매우 유용하다!)

$$\boxed{\begin{aligned}\sigma_B &= \sqrt{\frac{N}{\Delta}} \sigma_Y = \sqrt{\frac{\frac{1}{R^2} - 1}{N - 2}} B \\ \therefore \quad \sigma_A &= \sqrt{\frac{\sum_i X_i^2}{N}} \sigma_B\end{aligned}}$$

이렇게 얻어낸 불확실도들은 모두 통계적 기법으로 얻어내었으므로 (우연오차와 표준편차가 같아지는 유의수준은  $1\sigma$ 이다.)  $1\sigma$ 의 신뢰수준을 가진다.

이외에도 그래프의 plot과 오차 막대 표기를 통해 회귀선 및 회귀선의 오차를 육안 보간하는 방법도 간편하게 사용할 수 있다. 이 방법에 대해서는 9장에서 소개하도록 한다.

원점을 지나는 등분산 선형회귀의 경우에도 마찬가지로 회귀계수  $B = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}$ 를  $Y_i$ 로 편미분하여 다음과 같이 불확실도를 얻어낼 수 있다.

$$\frac{\partial B}{\partial Y_i} = \frac{X_i}{\sum_i X_i^2}$$

$$\sigma_B = \sqrt{\sum_i \left( \frac{X_i}{\sum_i X_i^2} \right)^2} \sigma_Y = \frac{\sigma_Y}{\sqrt{\sum_i X_i^2}}$$

비선형 회귀 모형에서는 회귀계수를  $Y_i$ 들로 나타내는 표현식을 알고 있는 경우 제곱오차의 크기와 회귀계수의  $Y_i$  방향 편미분값들을 바탕으로 회귀의 불확실도를 추산할 수 있다. 이 과정에서  $i$ 번째 회귀계수  $A_i$ 의 불확도는 다음과 같이 산정된다. (물론 이 경우에도 유의수준은  $1\sigma$ 이다.) 이때 조작변인의 무분산성, 종속변인의 등분산성이 보장됨을 가정한다.

$$\sigma_{A_i} = \sqrt{\sum_i \left( \frac{\partial A_i}{\partial Y_i} \right)^2} \sigma_Y$$

이 때,  $\sigma_Y$ 의 산출 과정에서 선형회귀의 경우 분모에  $N - 2$ 가 사용

되었으나, 여기에서는 다른 자유도가 사용되어야 한다. 선형회귀의 경우  $A, B$  2개의 회귀계수를 얻어내야 하므로 최소 3개 이상의 데이터를 수집하여 불확실도를 논할 수 있다. 이 때문에, 비선형 회귀의 자유도는 측정 횟수  $N$ 에서 얻어내어야 할 계수의 개수를 빼준 값을 사용해야 한다.

한편, 이분산 가정이 요구되는 경우 회귀 계수를  $X_i, Y_i$ 로 편미분한 기울기에 각각의  $\sigma_{X_i}, \sigma_{Y_i}$ 를 곱한 뒤 각각 제곱하여 모두 더한 뒤 제곱근을 취하면 (다변수함수의 편미분을 이용한 오차의 전파 관계식) 회귀 계수의 불확실도를 얻어낼 수 있다.

### 8.3 유효 숫자

불확실도의 존재는 참값의 유효 숫자를 제한한다. 유효 숫자는 표기값 중 불확실성을 고려하더라도 신뢰할 수 있는 자릿수를 의미하며, 보다 구체적으로는 그 아래 자릿수에서 발생한 반올림(혹은 불확실도의 더해짐)에 의해 자릿값이 조금 정도 바뀔 수 있는 경우에도 유효한 자릿수로 간주하여, 그것까지 포함한 것을 유효 숫자의 개수로 논하게 된다.

이에 따라 유효 숫자의 기준은 불확실도의 크기를 기준으로 정해지게 된다. 기본적인 규칙은 불확실도보다 작은 값을 나타내는 자릿수는 유효 숫자로 간주될 수 없다는 것인데, 이는 불확실도의 정의를 생각해 보면 자명하게 알 수 있다. 따라서 불확실도의 가장 큰 자릿수가 최학치의 가장 작은 자릿수의 유효 숫자가 된다. 가령 중력 가속도의 측정에서  $0.2 [m/s^2]$ 의 불확실도가 얻어졌다면, 최학치의 유효숫자는  $9.8 [m/s^2]$ 까지 두 개만 사용할 수 있다. 최학치가  $9.81 [m/s^2]$ 이라 주장하더라도, 마지막 자리의 1은 불확실도 범위

안에 들어와 신뢰할 수 없는 자릿수이므로 유효 숫자로 인정되지 않는다.

물리량의 연산을 수행할 때 유효 숫자의 개수는 연산에 의해 전파되는 불확실도의 크기를 기준으로 정해진다. 예를 들면  $A = 3.14 \pm 0.03[-]$ 이며,  $B = 2.7 \pm 0.2[-]$ 인 경우 이 둘의 합의 최학치는  $5.84[-]$ 이며 불확도는 더해져  $0.23[-]$ 이 되는데, 후술하겠으나 일반적으로 불확실도의 유효숫자는 한 자리만을 감안하면 약  $0.2[-]$ 의 불확실도를 얻을 수 있다. 이에 따라 최학치의 유효숫자는 불확실도 첫 번째 자리에 의해 재단되어  $5.8[-]$ 이 된다. 이로부터 물리량의 덧셈에서 유효숫자는 더 큰 자릿값을 가지는 마지막 자리에 의해 재단된다는 것을 알 수 있다. 반면 이는 아주 정확한 백만 규모의 수 (유효 숫자 6~7개)에 유효 숫자 한 자리인 한 자릿수를 더하더라도, 원래 값의 유효 숫자 개수에는 차이가 없음을 뜻한다. 곱셈과 나눗셈의 경우 이는 완전히 달라진다.

불확실도의 전파 규칙에서 곱셈과 나눗셈의 경우 상대 오차가 누적되어 더해짐을 논하였다. 이를 토대로 곱셈과 나눗셈은 자릿수의 위치에 무관히 유효 숫자의 총 개수가 작은 쪽에 맞추어 재단된다. 가령 13487에 2를 곱할 경우, 2라고 표기된 값은 1.5부터 2.5까지의 값을 가질 수 있는 매우 불확실한 정보이므로 (상대 오차의 크기가 큼) 곱셈의 결과에서도 유효 숫자를 한 자리밖에 보장할 수 없게 된다. 곱셈과 나눗셈 이외에, 어떤 일변수 함수의 함숫값의 유효 숫자의 개수는 함수의 파라미터가 가지는 유효 숫자의 개수를 따라가는데, 이는 기울기에 기반한 불확실도의 전파로부터 상대 오차의 크기가 대략 보존된다는 것으로부터 알아낼 수 있다. 이는 어느 정도의 선형성을 상정 (오차 범위 내에서 도함수값의 큰 변화가 없음)하고 얻은 결론이 된다. 유효 숫자 기준은 최소 눈금에 대해서도 유사하게 적용되기도 한다.

유효 숫자의 개수를 논할 때에는, 본질적으로 규칙을 딱딱하게 적용하기보다는 불확실도를 제대로 된 방법으로 유의 수준을 보존하며 추산해낸 뒤 그에 맞게 최확치를 재단한다는 개념으로 판단하는 것이 가장 바람직하다. 상술한 유효 숫자 기준은 어디까지나 근사적인 규칙일 뿐, 유의 수준을 보존하지 않고 크기 규모만 따져 자릿수를 추산하기에 불확실도 추산을 통한 최확치 재단에 비해 정교하지 못하다. 이러한 점을 고려하여, 최종 결과를 얻을 때까지의 중간 과정에서는 자릿수가 부족하여 불확실도가 과도하게 누적되는 것을 막기 위해 유효숫자를 한 자리 늘려 추산하는 것이 보다 바람직하다. (물론 기입할 경우 유효숫자 기준을 따지는 것이 좋다. 이러한 문제를 막기 위한 근본적인 해법은 불확실도를 2자리 유효숫자로 기입하고, 최확치의 유효숫자를 불확실도의  $1/10$  자릿수에서 재단하는 것이다.)

불확실도의 유효 숫자는 일반적으로 한 자리이며, 이에 맞추어 최확치의 표기 자릿수도 재단하게 된다. 불확실도는 유의 수준에 따라 엄격하게 통계 처리를 할 수도 있지만, 근본적으로 불확실한 양이므로 크기 규모만 보면 되는데, 반올림할 경우 두 번째 자리를 생략하더라도 10% 정도 규모의 오차로 크기를 추산할 수 있게 되기 때문에 한 자리의 유효숫자만으로 기입하는 것이 보통이다.

물론 두 번째 자리의 생략이 불확실도의 값을 큰 폭으로 바꿀 수 있는 경우에는 불확실도를 유효숫자 두 자리까지 기입하는 경우도 있다. 예를 들어, 첫 번째 유효숫자가 1로 시작하는 경우 반올림에 따른 결과가 최대 2배까지의 차이를 허용하므로 불확실도의 크기 규모를 크게 흔들 수 있다. 이 경우 명확한 크기 추산을 위해 불확실도를 두 자리 기입해 주는 것이 바람직하다. 이외에도 앞서 언급된 유효 숫자 부족으로 인한 불확실도의 과잉 추산을 막기 위해 정밀한 학술 연구에서는 두 자리의 불확실도를 기입하는 경우도 종종 있다.

불확실도를 두 자리 기입하는 경우에는 최확치의 자릿수도 불 확실도의 끝자리에 맞추어  $3.14 \pm 0.15$ 와 같이 기입한다. 한편 이러한 경우 유효숫자는 표기값이 3개의 자리를 가지고 있으나 여전히 2 개에 불과하다. 이러한 표기를  $3.14(15)$ 의 형태로 나타내기도 한다.

데이터의 경향을 파악하기 위해서 적절한 회귀와 함께 scatter plot을 통해 2차원 공간에 데이터 포인트들을 나타낼 수 있다. 이 경우, X축에 조작 변인, Y축에 종속 변인을 나타내어 데이터가 가지는 개형을 살펴 볼 수 있으며, 데이터 포인트들과 회귀된 추세선 사이의 편차를 한 눈에 알아보기 쉽게 Visualize할 수 있다. 실험 결과의 Plot은 단순히 실험 결과를 가시화할 뿐 아니라, 조작 변인과 종속 변인 간의 관계와 기울기들을 명확하게 확인할 수 있어 오차의 전파를 손쉽게 계산할 수 있고, 오차 막대 및 회귀선의 불확실도를 가시화하여 실험계가 가지는 오차의 특성을 파악하여 통계론적 오차 분석을 적용하는 데 유용한 도구가 될 수 있다.

Plot은 모눈종이를 이용하여 수기로 작성될 수도 있으나, 최근에는 전산 기술의 발달로 쉬운 스크립트 언어를 이용해 Plot을 작성할 수 있다. Python을 이용한 matplotlib 라이브러리나 이를 포함한 pylab, 혹은 GNUPlot 등의 툴을 이용하면 원하는 요소를 원하는 형태로 배치하여 고도로 사용자화된 그래프를 작성할 수 있으며, 비교적 쉽게 수행할 수 있고 스크립트가 남아 유사한 그래프를 다시 그릴 때 약간의 수정만으로 다시 그려 낼 수 있다는 장점이 있고, 전산 회귀 분석과 연계하여 사용할 수 있다는 장점이 있다. C 언어를 사용한 old-fashioned way도 물론 가능하지만, 라이브러리 제공이 번잡하기 때문에 필요한 과정이 많아 그리 권장되지 않는다. MATLAB이나 Origin과 같은 상용 데이터 처리 소프트웨어를 이용하여 자료 해석을 수행할 수도 있으며, 이들 또한 강력한 그래프 Plot 기능을 제공한다. 급한 경우, Excel 등의 범용 스프레드시트 프로그램을 이용해 그래프를 그리는 경우도 있으나, 이 경우에도 약간의

9.1 Plot의 형태 . . . . .	74
9.2 Plot의 오차 분석 . . . . .	75
9.2.1 오차 막대 (Error Bar)	75
9.2.2 회귀선의 불확실도 . . .	78
9.2.3 유효 데이터의 수 . . . . .	79
9.3 Eye-Balling Estimation . . .	79
9.4 Plot을 이용한 interpolation	82
9.5 Plot 작성 팁 : 들어가야 할 내용 . . . . .	85
9.5.1 타이틀 . . . . .	85
9.5.2 축 . . . . .	86
9.5.3 범례 (legend) . . . . .	87
9.5.4 회귀선 . . . . .	88
9.5.5 데이터 포인트 . . . . .	89
9.5.6 회귀 모델 . . . . .	89
9.5.7 물리량의 단위 . . . . .	89

신경을 써 주면 필요한 요소를 많이 담고 있는 미려한 그래프를 Plot 해 낼 수 있다. 이 장에서는 이처럼 그래프를 그리는 데 필수적으로 반영되고 고려되어야 할 요소들에 대해 소개한다.

## 9.1 Plot의 형태

데이터 plot은 크게 두 가지의 형태로 나눌 수 있는데, 첫 번째는 조작변인과 종속변인의 correlation을 나타내는 선형 관계 plot이며 두 번째는 극값 peak를 가지는 곡선 형태의 관계에 해당한다. 전자는 선형회귀를 통해 기울기와 절편 정보를 얻어내는 데 사용되며, 회귀 변수를 잘 설정하여 지수적인 증가/감소 등의 반감기 등을 분석하는 데에도 사용될 수 있고, 이론적인 예측치와 실험치를 비교하여 선형성을 분석하면 어느 구간에서 이론과 실험이 잘 일치하는지 편차를 분석하기 유용하다. 후자의 경우 peak의 위치와 높이, 폭 등으로부터 중요한 공명곡선 등의 정보를 얻어낼 수 있다. 피크를 가지는 곡선의 경우, 이차함수를 이용한 quadratic fit을 국소적으로 수행하여 극값의 위치와 높이, 곡률을 찾아내거나, 공명곡선의 정보를 얻기 위해 Lorentzian fit을 수행할 수 있으며, 정규분포를 따르는 경우(열적 들뜸 등으로 얻어지는 분포를 확인할 때 등등) Gaussian fit 등을 수행하여 피크의 높이, 폭(FWHM; Full Width Half Maximum), 표준편차, 밀넓이 등을 얻어낼 수 있고, bimodal, Poisson distribution 등 여러 가지 함수식을 사용해 회귀하고 plot에 나타낼 수 있다. 이외에도 회절 무늬 등을 회귀하기 위해 sinc, bessel, airy function 등등 다양한 고급 회귀 모델들 역시 적용 가능하다.

상술한 바와 같이 Plot에서는 선형 자료의 경우 8개, 극값 근처의 경우 5개, 변곡점 근처의 경우 3개 이상의 서로 구분되는 데이터 포인트를 수집하여 나타낼 것이 권장된다. 이때 데이터 포인트끼리의 오차 막대가 겹친다면 이는 통계적으로 서로 다른 데이터 포인트가

아니라는 귀무가설을 기각할 충분한 근거가 부족하므로 추가적인 데이터 포인트를 보태어 plot으로 나타내는 것이 바람직하다.

## 9.2 Plot의 오차 분석

### 9.2.1 오차 막대 (Error Bar)

Plot에서 오차를 만드는 요인은 데이터 포인트들의 측정 오차이다. 회귀의 종류에 따라 다르지만, 기본적으로 모든 실험의 데이터 셋은 조작 변인 (X축)의 불확실도와 종속 변인 (Y축)의 불확실도를 포함한다. 이 때 이를 간편하게 나타내기 위해 오차 막대 표기를 사용한다.

오차 막대(Error Bar)는 I자 형태로 그려지며, X축, Y축 방향으로 모두 기입되거나 혹은 조작 변인의 불확실도가 무시되어 Y축 방향으로만 기입되는 경우도 있다. 오차 막대의 의미는 신뢰 구간과 같은데, 이는 회귀선이 측정된 데이터 포인트를 엉나갈 확률 분포를 제시하는 것이다. 회귀선이 오차막대 바깥을 지날 확률은 신뢰구간의 유의 확률보다 낮은 것으로, 매우 일어날 가능성이 낮다. 다르게 말하면, 오차 막대를 벗어나게 그릴 회귀선은 최확 회귀선일 확률이 매우 낮다. 이에 따라, 가급적 모든 데이터 포인트의 오차 막대 안쪽을 지나가도록 회귀선을 그려 주는 것이 제곱오차를 최소화하는 방법이 된다.

오차 막대를 그릴 때 유의 수준은  $1\sigma$ 나  $2\sigma$ 가 사용되는데,  $2\sigma$ 로 사용할 경우 회귀선이 오차막대를 벗어날 확률이 매우 작아 거의 대부분의 데이터가 회귀선을 담는 오차 막대를 가지게 된다. 반면,

$1\sigma$  오차 막대를 그린다면 이론적으로 얻은 값과 비교하여 그려 넣기 용이하지만, 회귀선이 오차 막대 바깥으로 나갈 확률이 36% 가량이므로 육안 보간을 통해 회귀 오차를 추산하기 어려워진다.

이에 따라 오차 막대가 큰 데이터는 가중치가 줄어든다는 사실을, 이분산 회귀에서 회귀선이 데이터 포인트로부터 멀리 떨어져 있어도 여전히 오차 막대 안에 들어올 확률이 크게 줄어들지 않는다는 점으로부터 알 수 있다. 이는 제곱 오차의 성질로부터 알아낼 수 있는데, 지수오차  $F = \exp(-E^2)$ 을 정의하면  $E$ 가 최소가 될 때  $F$ 가 최대의 값을 가진다. 이 때 제곱오차는 각각의 데이터 포인트들에 대한 제곱합이므로  $E^2 = \sum_i \delta_i^2$ 인데 ( $\delta_i$ : deviation), 이를  $F$ 에 넣으면 각각의 데이터 포인트들이 가지는 분산에 대해 추세선이 편차만큼 떨어진 지점을 지날 확률의 곱이 얻어진다. 즉,  $F = \prod_i \exp(-\delta_i^2)$ 이다. 이로부터 각각의 오차 막대가 가지는 의미는 최확 추세선이 각각의 데이터들로부터 지금의 편차만큼 떨어져 있을 확률을 제공하는 것임을 알 수 있다. 오차 막대가 확률의 의미를 가진다는 것을 알면, 데이터 포인트의 분포와 오차 막대의 크기만으로 회귀선을 육안 보간할 수 있으며, 신뢰수준을  $2\sigma$ 로 놓고 불확실도를 추산할 경우 비교적 잘 들어맞는 값을 얻는다. 이러한 기법의 강력함은 앞서 7장에서 살펴본 것과 같이 통계적으로 복잡하고 어려운 이분산 회귀를 손쉽게 눈대중으로 추산할 수 있으며, 이 때 부정확한 데이터 포인트에 대해서는 낮은 가중치를 주는 것까지 반영된다는 점에서 확인할 수 있다.

오차 막대의 크기를 기입하는 방법은 측정 오차와 회귀 편차의 대소 관계에 따라 달라진다. 등분산 가정이 유효한 경우 회귀에서 얻어지는  $\sigma_Y$ 를  $Y$ 방향 오차 막대의 반 폭 (half-width) 이 되게 기입(이 경우에는  $X$ 방향 오차 막대를 무시한다)하면 되며, 그렇지 않은 경우 측정에서 얻어진 측정 불확실도를 이용해 오차 막대를 작도하며 이 경우 필요시  $X$ 방향 오차 막대도 같은 방법으로 기입해 줄 수 있다.

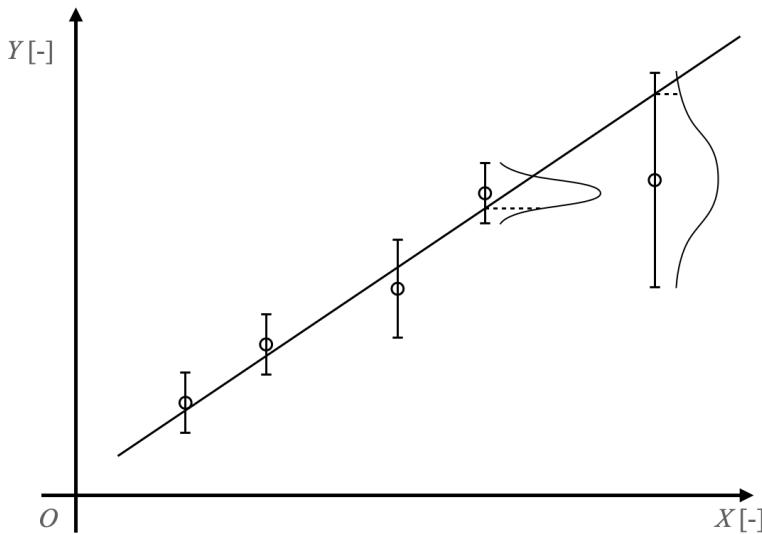
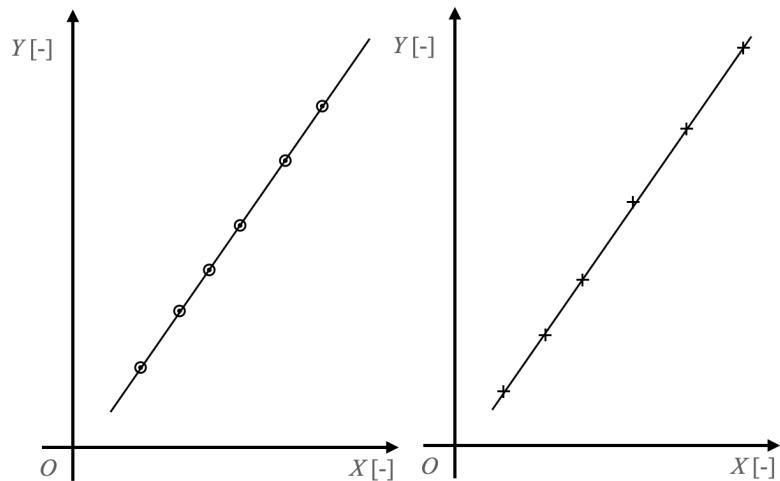


Figure 9.1: 오차 막대와 회귀선 표기

신뢰 구간이 비대칭 분포로부터 얻어질 경우 오차 막대 역시 비대칭으로 기입되어야 한다. 이는 주로 비선형 함수의 극값 근처에서 (곡률이 큰 구간) 오차의 전파를 사용한 결과로 얻어지거나, 회귀 변수가 특수하게 정의된 경우 발생한다. 가령,  $X$ 를 바꾸어 가며 실험하고  $1/X^3$ 을 회귀 조작변인으로 삼는다고 하자. 먼저,  $X$ 가 작은 곳에서 등분산 측정을 했다고 하면  $1/X^3$ 의 최확치는 아주 큰 값을 가진다. 이 때,  $X$ 의 불확실도가 최확치의 절반 정도로 상대적으로 큰 (측정이 정말해도 최확치가 작은 구간을 지나는 경우) 불확실도를 가지므로, 이에 역수를 취하고 세베곱을 걸면 매우 비대칭적인 신뢰구간을 얻게 된다. 반대로, 이번에는  $X$ 가 큰 영역에서 실험을 수행한다고 하자. 이러한 경우 아주 큰 값  $X$ 의 불확실도가 아무리 크더라도,  $1/X^3$ 의 작아지는 방향 불확실도의 크기는 거의 0일 수밖에 없다. 이처럼 정의상 신뢰 구간이 0보다 작아질 수 없는 경우에도, 이러한 점을 유의하여 오차 막대를 기입해 주어야 한다.

한편, 불확실도가 너무 작은 (회귀 편차와 측정 오차 모두) 경우에는 다음 그림과 같이 동심원과 점 혹은 작은 십자 표시로 오차 막대 표기를 대체할 수 있으며, 그 의미는 불확실도의 크기가 Plot에서 구분하여 식별 가능한 최소 눈금보다 작다는 것을 의미한다.



**Figure 9.2:** 불확실도가 작고 선형성이 높은 (회귀 편차가 작은) 자료

### 9.2.2 회귀선의 불확실도

8장에서 논한 것과 같이, 회귀선의 불확실도는 데이터 포인트들의 오차 막대로부터 전파된 것이다. 선형 회귀의 경우, 이를 그래프 상에 나타낼 때에는 3개의 직선을 이용해 나타낼 수 있는데, 최적 추세선 (Best estimated trendline)과 상한선, 하한선을 그릴 수 있다. 최적 추세선은 최소자승법으로 얻어진 회귀계수들로 기술되는 회귀선이며, 상한선과 하한선은 각각 최대 기울기, 최소 기울기를 가지도록  $A, B$ 의 불확도 범위를 고려하여 그려지는 선이다.

이상적인 경우 세 직선은 한 점에서 교차하는데, 이는 데이터들의 조작 변인 RMS 평균에 해당한다.  $Y = (A + dA) + (B - dB)X$ 와  $Y = (A - dA) + (B + dB)X$ 의 교점을 찾으면,  $X = \frac{dA}{dB} = \sqrt{\frac{\sum_i X_i^2}{N}}$ 을 얻으므로, 이는 데이터 포인트의 조작 변인의 RMS 평균에 해당한다. 이때 교점의 좌표 중  $Y = A + B \frac{dA}{dB}$ 이므로, 이는 최적 회귀선 위에 놓인다. 즉, 세 직선은 한 점에서 교차한다. 세 자세한 논의는 9.3절을 참고하라.

비선형 회귀의 경우 극값 주변에서 Quadratic Fit을 수행하거나, 점근 성질(Asymptotic)을 가지고 있는 자료를 분석할 때 회귀선의

불확도를 고려할 필요가 있다. 각각 적절한 회귀 모형의 불확도를 바탕으로 최적 회귀선과 최대 편향 회귀선을 찾아 Plot할 수 있다.

### 9.2.3 유효 데이터의 수

한편 실험 검정을 통해 서로 다른 데이터 포인트라고 말하기 충분한 통계적 유의성이 없는 경우, 그 데이터들은 서로 다른 데이터 개수로 세기에 적합하지 않다. 물론 통계 처리에서 사용될  $N$ 값은 이들을 서로 다른 값으로 세 주어야 하지만, 데이터 샘플링 간격을 얻기 위해서는 이들은 분간되지 않는 데이터로 간주하여, 9.1절에서 제시한 데이터 측정 횟수를 어느 정도 충족시키도록 간격을 설정하는 것이 좋다.

## 9.3 Eye-Balling Estimation

복잡한 이분산 분포를 가진 자료의 선형회귀나, 간단하게 극값 근처의 국소 quadratic 회귀를 통해 극값의 위치와 그 불확도를 알 아내는 데에는 plot을 이용한 오차 추산이 유용하게 사용될 수 있다. 이는 육안으로 불확실도의 간단한 규모를 얻어내는 방법인데, 오차 막대를 바탕으로 가능한 회귀선의 범위를 찾을 수 있다. 가능한 한 오차 막대의 중앙을 모두 지나도록 그어 준 추세선이 최적 회귀선이 되며, 이 때 오차막대가 큰 데이터보다 오차막대가 작은 데이터를 더 우선 순위에 두고 작도한다. 이는 이분산 가중 평균을 기하학적으로 눈대중하여 추산할 수 있는 간편한 방법이다. 물론 정교한 불확실도 추산이나 기울기를 이용한 물리량의 최적치 결정에 사용하기에는 신뢰도가 떨어지거나, 잡음 등의 오차 요인들로 인해 분산이 큰 실험 자료들을 이용해 실험을 진행할 적절한 구간을 찾는 등의 예비 실험

을 할 때 빠르고 편리하게 회귀분석이 가능하다는 장점이 있다.

회귀선의 육안 추정에서도 등분산 회귀와 이분산 회귀는 다른 방법으로 이루어진다. 측정 오차가 회귀 편차보다 큰 경우, 데이터들에 대해 등분산 가정이 성립하지 않는다. 이 때, Plot 상에 데이터 포인트들의 위치와 오차 막대의 크기를 측정된 불확실도를 기반으로 표기한다. 각 데이터 포인트의 오차 막대의 크기는 6장에서 논의한 방법과 같이 반복측정 혹은 계측기기의 보간을 이용해 얻어낼 수 있는 우연 오차와 같은데, 이 때 유의 수준을  $2\sigma$ 로 그리도록 한다.  $2\sigma$  유의 수준을 사용할 경우  $1\sigma$  수준 하에서 얻어진 불확실도의 약 2배 가량의 불확실도를 얻으며, 불확실도는 양측 신뢰 구간을 가지므로 오차 막대의 길이는  $2\sigma$  불확실도의 2배, 그리고  $1\sigma$  수준으로 추산된 불확실도의 4배가 된다. 컴퓨터를 이용한 Plot에서는  $1\sigma$  오차 막대를 사용하는 것이 유의 수준의 보존 측면에서 편안하지만, 육안으로 추세선과 불확도를 추정할 때에는  $2\sigma$ 의 유의수준을 사용하는 것이 유의 확률(5%)이 낮아져 보다 편리하다.

이때 회귀선을 오차 막대들의 안쪽을 최대한 지나게 하며, 가급적 데이터 포인트들로부터의 상대 편차(오차막대의 길이 대비 실제 편차) 가 어느 정도 일정하도록 최적 회귀선을 눈대중하여 찾을 수 있다. 이 경우 상대 편차를 일정하게 조율하는 과정에서 분산 역수가 중치가 자연스럽게 반영된다. 이후 오차 막대를 벗어나기 직전까지 기울기를 키우거나 줄여 각각 상한 회귀선과 하한 회귀선을 작도한다. 이 때, 상한 회귀선과 하한 회귀선은 대략 최적 회귀선의 중앙 부근에서 교차하도록 작도할 수 있다. (이상적인 경우, 세 추세선은 데이터 포인트의 평균에서 교차한다.)

회귀 편차가 측정 오차보다 큰 경우에도 유사한 방법을 사용할 수 있다. 이 경우에는 등분산 가정이 가능한데, 이 때는 데이터 포인트를 먼저 Plot하고, 편차가 최소화되도록 최적 추세선을 눈대중으로

작도한 후, 각각의 데이터 포인트들에 대해 추세선을 거의 포함하도록 똑같은 길이의 Y방향 오차 막대를 기입한다. 이는 종속변인의 등분산 가정과 조작변인의 무분산 가정을 반영한 것이다. 이후 상한 회귀선과 하한 회귀선을 그리는 방법은 동일하다.

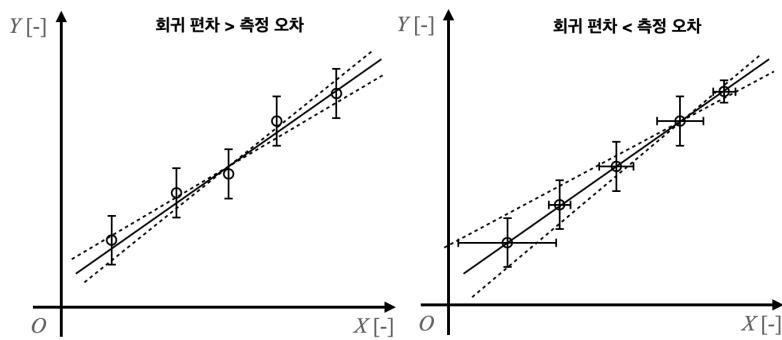


Figure 9.3: 회귀 편차와 측정 오차의 비교

이때 상/하한선의 교차점으로부터 같은 거리만큼 가로축 방향으로 이동한 눈금 대비 상/하한선 및 최화 추세선까지의 세로 눈금을 읽어 기울기의 최확치 및 불확도를 추산한다. 절편은 축의 원점 근처에 물결 표기를 이용한 축약이 없는 경우 직접 읽을 수 있다. 그러나 절편 추정의 불확도가 크므로, 육안 보간이 이루어지는 경우에는 절편의 불확도를 그래프로부터 읽어 내는 것보다, 기울기의 불확도를 이용하여 등분산 근사 하의 이론 관계를 빌려 대강의 크기를 추산하는 것이 유리하다. (물론 등분산 가정이 성립하지 않는 경우 이 관계는 유효하지 않으나 대강의 경향은 일치한다.)

앞서  $y$ -절편의 불확도  $\sigma_A = \sqrt{\frac{\sum_i X_i^2}{N}} \sigma_B$  임을 등분산 선형회귀의 경우 확인한 바 있는데, 앞의 계수는 조작변인의 rms 평균에 해당한다. 육안 보간의 경우, 이 값은 대강 상한선과 하한선의 교차점의 X좌표와 비슷한 크기를 가진다. (30% 이내 오차이므로 불확실도 추산에 크게 불리하지 않다.)

등분산 선형회귀의 경우 오차 막대의 크기는 회귀선을 대략 포함하도록 설정되는데, 이 때 오차 막대의 크기는 대략 회귀선 기울기 불확실도에 조작 변인의 구간 길이(기울기가 너무 크지 않은 경우, 이 값은 대략 상/하한선 교차점 X좌표의 절반과 비슷하다.)의 절

반을 곱한 것의 두 배(양축 구간이므로) 정도가 된다. 한편 절편의 불확실도는 회귀선 기울기 불확실도에 상/하한선 교차점 X좌표를 곱한 값이므로, 기울기가 너무 크지 않은 경우 대략 절편의 불확실도의 두 배는 오차 막대의 크기와 거의 유사하게 된다. 이는, 회귀선의 상/하한선이 Y축과 교차하는 절편의 폭이 대략 오차 막대의 길이와 유사한 규모를 가진다는 의미이며 이를 통해 절편의 불확도 규모를 추산할 수도 있다.

육안 추정 시 불확실도의 규모는 유의 수준  $2\sigma$  하에서 추정되므로, 최소 자승법으로 예측한 값( $1\sigma$ )에 비해 두 배 가량 커짐에 유의 하여야 한다. 유의 수준과 오차의 전파 관계를 잘 인지하고 있으면, 눈대중으로도 쉽고 빠르게 회귀 분석 및 오차 처리를 할 수 있다는 것은 실험자에게 강력한 직관을 줄 수 있는 도구가 될 것이다.

극값 근처에서의 Quadratic 회귀를 눈대중으로 수행할 경우에 는, 측정 오차로서 얻어진  $X, Y$  방향 불확실도를 둘 다 오차 막대로 나타내고, 최소 5개 이상의 데이터들을 이용해 좌측 최대편향 상한 회귀선과 우측 최대편향 하한 회귀선을 그리고, 편차를 가장 작게 만드는 최화 회귀선을 그린 뒤 피크의 위치를 Y축에 평행한 점선으로 X축 눈금 위까지 내려 극값 위치의 최화치와 불확도를 추산할 수 있다. 이는 분포가 꽤 넓은 경우 유효하며, 좁고 날카로운 피크에 대해서는 데이터 수를 늘려 수치 방법을 이용한 Lorentzian 혹은 Gaussian Fit을 수행하는 것이 합리적이다.

## 9.4 Plot을 이용한 interpolation

조작 변인에 따른 실험계의 응답을 얻어 이론적으로 예측된 상관 관계와 비교하고자 한다. 이 때, 실험계의 응답은 샘플에 의한

것 이외에도 계측기기의 비선형성으로부터 올 수도 있다. 이 때, 빈 샘플을 넣고 실험에 사용할 정의역에서 조작 변인을 바꾸어 가며 측정치를 얻는다고 하자. 이 때 얻어진 값이 일정하지 않다면 이 값은 온전히 계측기기로부터 온 것이다. 따라서 샘플을 실험하기 전에 이 값을 미리 측정해 효율이나 전달함수를 보정해 주어야 할 필요가 있다.

이 때, 정석적으로는 조작 변인을 고정시켜놓고 샘플이 있는 경우와 그렇지 않은 경우를 비교해야 한다. 그러나 경우에 따라 조작 변인을 고정시킨 채 실험하지 못하는 경우가 있을 수 있는데, 이 경우 고정된 조작변인에서 샘플이 없을 때의 측정치 뿐 아니라, 임의의 조작 변인에 대한 연속함수로서의 측정치를 알아야 한다. 이를 위해 서는 실험치의 내삽 (Interpolation)이 필요하다. 즉, 분모와 분자에 들어가야 할 측정치들이 같은 조작변인으로부터 측정되지 못할 경우, 분모에 들어갈 효율에 대한 정보를 주변 조작변인의 측정치들로부터 인터폴레이션하여 얻어낼 수 있다.

인터폴레이션과 회귀의 차이점은 크게 두 가지로, 인터폴레이션은 회귀 모형의 식을 몰라도 가능하다는 점과, 불확실도를 허용하지 않는다는 점이다. 이러한 차이는 인터폴레이션으로 데이터를 얻을 때, 주변 조작변인에서의 측정치가 잡음이 심하거나 불확실하면 좋은 결과를 기대하기 힘들다는 것을 의미한다. 즉 인터폴레이션을 시행하기 전에 적절한 Sampling과 Smoothing을 거친 뒤, 오차의 전파 기법을 이용해 인터폴레이션 된 연속함수에 대한 불확실도 논의를 가져야 한다. 회귀에서는 등분산 가정의 경우 회귀 편차, 즉 데이터들이 회귀선으로부터 얼마나 벗어났는지 - 회귀 모델로부터 얼마나 우연오차를 가지는지가 회귀선의 불확실도로 전파되었으며, 그렇지 못한 경우에는 각각의 측정 오차들이 회귀선의 불확실도로 전파되었다. 이 경우 인터폴레이션은 측정오차들의 불확도가 그대로 인터폴레이션 불확실도로 이어진다.

인터플레이션은 주어진 데이터 포인트들을 불확실도를 허용하지 않고 모두 지나는 연속함수를 찾아내는 것을 의미한다. 자료에 노이즈가 많을 경우, 이들을 인터플레이션하면 구불구불하고 복잡한 쓸모없는 결과를 얻게 되는데, 이를 방지하기 위해 여러 데이터들을 묶어 평균을 취하는 등의 방법으로 Smoothing을 수행할 수 있다. 이 경우 데이터들의 분산은 Smoothen된 하나의 데이터 포인트가 가질 수 있는 우연 오차로 반영되며, 6장에서 언급된 반복 측정에 의한 불확실도 추정의 방법으로 기술될 수 있다. 결과적으로, 점들을 모두 지나도록 이어도 곡률이 너무 커지지 않는, 완만한 경향을 잘 반영하는 데이터포인트들을 얻은 뒤 이들을 인터플레이트하면 원하는 연속 함수를 얻어낼 수 있다.

수학적으로 엄밀하고 정확한 보간 방법은 11장에서 논의하며, 대체로 전산 처리를 통한 수치적 방법으로 얻어진다. 본 장에서는 plot을 이용한 기하학적인 육안 보간에 대해 논한다.

Plot된 데이터로부터의 육안 보간은 Sampling과정을 생략하고, 각각의 데이터 포인트들을 모두 오차 막대와 함께 plot한 뒤, 이들을 적당히 부드럽게 지나는 최적 내삽선을 그린 뒤 회귀선을 그려주는 것과 같이 상/하한선을 제시해 불확도를 얻는다. 이는 회귀와 매우 비슷하나, 회귀 과정에서는 어떤 개형을 따라야 할지를 이론적으로 알 수 있는데 비해, 여기서는 실험적으로 얻어진 개형을 눈대중으로 보간한다는 점에 차이가 있다. 이 과정에서 여러개의 데이터 포인트를 정확하게 지나는 대신, 몇 개의 Smoothen된 데이터 포인트들을 잇는 곡선이 되는 형태로 sampling이 자연스럽게 이루어진다. 보다 정확한 방법은 컴퓨터를 이용해 여러 가지 형태의 보간함수를 제시할 수 있는데, 이 경우 주어진 유한 개의 정보로부터 얻어낼 수 있는 연속함수의 수는 무한히 많게 된다.

이때 불확실도를 반영한 상/하한선의 경우, 회귀의 경우에는

최확 추세선과 교차하나, 인터폴레이션의 경우 단순히 오차막대의 위쪽 끝들과 아래쪽 끝들을 지나게끔 각각 상한선, 하한선을 얻으면 된다. 이는 회귀 모델이 가지는 통계적 성질 없이, 각각의 조작 변인에 대한 종속 변인이 이 정도 신뢰구간 안에 포함될 것으로 기대될 수 있음을 제시하는 것으로, 육안 보간 시 약  $2\sigma$ 의 신뢰수준을 가진다.

육안 보간은 태생적으로 불확실할 수밖에 없으며, 노이즈에 취약하므로 통계적으로 강한 근거를 제공하지 못한다. 정확한 데이터 포인트의 위치를 기반으로 한 전산 인터폴레이션은 여러 수학적 가능성을 제시하지만, 불확실도를 허용하지 않으므로 샘플링을 주의 깊게 거치지 않으면 신뢰할 수 없는 결과를 얻게 된다. 인터폴레이션의 통계적 유의성 및 여러 수학적 보간 함수 제안은 11장에서 자세히 논한다.

## 9.5 Plot 작성 팁 : 들어가야 할 내용

Plot을 작성할 때는 가급적 다음의 요소들을 빠짐없이 기입하는 것이 권장된다 :

### 9.5.1 타이틀

Plot이 어떤 조작 변인에 의해 어떤 종속 변인이 변화하는 경향을 보고자 하는 것인지 기입한다. 선형 관계의 경우 <Y – X Correlation>의 형태, 비선형 데이터의 경우 <Y – X Curve>의 형태로 타이틀을 붙이며, 그래프/Figure의 번호나 간단한 물리적 의미의 설명을 덧붙일 수 있다. 타이틀을 생략한 채 plot의 의미에 대한 간략한 설명으로 대신하는 경우도 있으나, Y – X의 순서에서 반드시 종속 변인이 앞에

오게 기입하는 것이 관례이다. ( $v - t$  그래프,  $I - V$  특성곡선 등의 예시를 생각해 보자.)

### 9.5.2 축

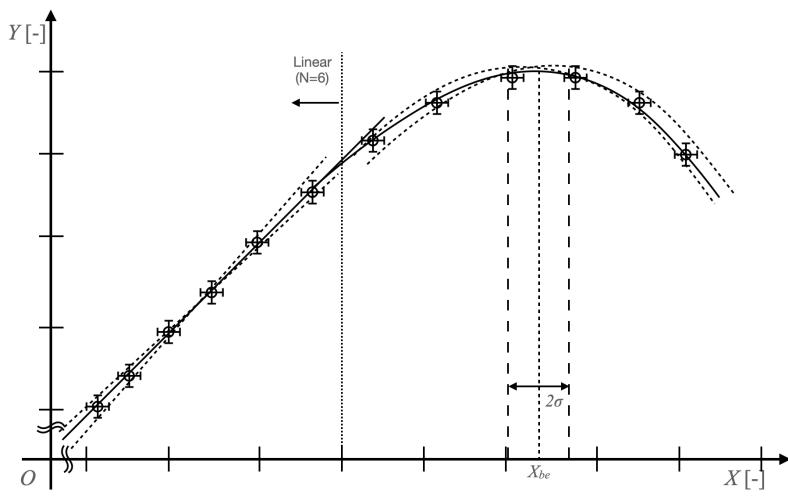
축은 데이터 포인트의 조작 변인과 종속 변인이 놓이는 공간 기준이다. 축은 눈금과 축 변수, 원점 등으로 이루어져 있으며, 축 화살표 끝에 어떠한 변수가 놓이는지와 그 물리량의 단위가 기입되어야 한다. 축은 로그 스케일이나 선형 스케일 중 적절한 것을 선택할 수 있으며, 원점에서 멀리 떨어진 값들이 주로 표시되는 경우 축에 물결 표시를 통해 경향을 잘 보여줄 수 있는 구간의 데이터들이 넓게 보일 수 있도록 축의 눈금을 설정하는 것이 좋다. 축의 축척이나 눈금의 설정 등은 그래프 영역이 공간의 80% 이상을 차지하도록 그려 주는 것이 바람직하다. 그래야 데이터 셋의 경향을 효과적으로 가시화하기 유리하다.

축 눈금의 크기는 대략 5개에서 10개 사이의 눈금 개수를 가지고 록 선형적으로 설정하는 것이 일반이며, 로그 스케일을 사용할 경우 10이나  $2, e$ 의 밑을 설정하는 것이 일반이다. 또한 원점을 반드시 기입하는 것이 권장된다.

통상적으로 세로축에 종속 변인을, 가로축에 조작 변인을 기입한다. X축은 데이터 셋의 조작 변인을 의미하는데, 조작 변인의 정의역이 선형 구간과 비선형 구간을 모두 포함할 경우, Plot 영역에 회귀에 포함하는 선형 구간과 그에 속하는 데이터 수를 표시해 줄 필요가 있다.

한편 Plot 상에서 특정 점이나 구간의 위치를 읽어 주기 위해

축 상에 Reading을 기입할 수 있다. 주로 피크의 극값을 읽는데 사용되며, 극값 근처에서의 2차 함수 (Quadratic) 회귀와 같이 종종 사용된다. 예를 들면, 아래 그림과 같이 위의 내용들을 기입하여 나타낼 수 있다.



**Figure 9.4:** 그래프의 축과 관련된 정보의 Plot

### 9.5.3 범례 (legend)

서로 다른 여러 개의 상관관계를 가지는 데이터 셋들을 동시에 plot할 경우, 어떠한 데이터가 어떠한 실험군으로부터 얻어진 것인지 구분하기 위해 범례를 기입한다. 주로 데이터 포인트 점의 모양이나 색깔, 회귀선의 색이나 모양 (실선/점선 등) 을 달리해 서로 다른 실험군들의 자료를 구분하고, Plot 영역의 한쪽 구석에 작은 상자 를 만들어 어떠한 모양이나 색이 어떠한 실험군에서 얻어진 것인지 기입한다. 불확실도가 큰 경우 구분이 어려우므로 동시에 여러 실험 군의 자료를 기입하는 것이 권장되지는 않지만, 같은 구간에서 다른 개형을 보이는 경우나 피크 위치의 차이를 알아야 하는 경우, 이론 적 예측치와 비교하거나 자료의 접근 성질 (Asymptotic Behaviour) 등을 보여 주어야 하는 경우 여러 자료를 한 번에 나타내면 개형을 비교하여 가시화하는 데 유리할 수 있다.

#### 9.5.4 회귀선

데이터에 선형 구간이 존재하는 경우, 적절한 구간 내에서 선형 회귀를 수행할 수 있다. 이 경우 최적화 회귀선(최소 자승법으로 얻어진 데이터 포인트와의 편차를 가장 줄일 수 있는 회귀계수로 기술되는 추세선)을 실선으로 그려 주고, 회귀 계수(기울기, 절편)의 불확실도를 고려하여 기울기가 큰 상한선과, 기울기가 작은 하한선을 점선으로 그려 준다. 이상적인 경우 상한선과 하한선, 최적화 추세선은 데이터 포인트의 평균에서 교차하나, 실제 분석 결과는 약간 벗어날 수 있다.

극값에 대한 정보가 필요할 경우 2차 회귀를 수행하여 극값의 위치 및 불확실도를 기입해 줄 수 있다. 이 경우, 2차회귀 계수의 불확실도에 기반하여 양측으로 최대 편향된 회귀선을 각각 점선으로 기입해 줄 수 있다. 피크를 Lorentzian/Multimodal Gaussian 등으로 회귀하는 경우, 회귀선의 불확실도를 기입하기 위해 최대 편향 회귀선을 그려 넣으면 피크의 정보를 알아보기 불편하므로 최적화 회귀선만을 기입하며, FWHM, 밀넓이 등 몇몇 중요한 값을 기입한다.

이론 모델이 없는 영역(이론 오차가 커 선형성이 떨어지기 시작하는 구간, 혹은 서로 다른 모델 사이에 이론이 잘 맞지 않는 영역 등)에서는 Patching Function을 찾아 그려 줄 수 있다. 수기로 작성할 경우 육안 보간을 통해 회귀 가능한 구간에서의 추세선을 연장하여 대략 이어 줄 수 있으며, 디지털 자료의 원본을 가진 경우 적절한 회귀 모형을 찾아 (이론적 예측 없이 구간 내 개형을 잘 맞추기 위한 회귀 모형의 경우, 회귀계수의 수를 최소화하는 것이 바람직하다.) 선형 중첩하거나, 편차가 크지 않고 부드럽게 연결된 개형을 가진 자료인 경우 인터플레이션을 사용해 Patching Function을 얻을 수도 있다.

### 9.5.5 데이터 포인트

각 조작 변인에 대응되는 종속 변인의 위치를 2차원 공간 상의 점으로 나타낸다. 유효숫자는 일반적으로 2개 정도이며, 데이터 포인트는 항상 오차 막대를 수반해야 한다. 오차 막대를 기입하는 자세한 방법은 9.2절을 참고할 수 있다.

### 9.5.6 회귀 모델

회귀에 사용된 회귀 모델을 plot 공간이나, 범례 칸 안에 간략히 기입하는 것이 권장된다. 선형회귀의 경우를 예시로 하면,  $Y = A + BX$ 의 회귀 모델과,  $A, B$ 의 값과 불확실도, 단위를 기입하고  $R^2$  값을 기입하는 것이 권장된다. 이때,  $R^2$  값에서 이어지는 9는 유효숫자가 아닌 자릿수이므로, 유효숫자의 개수는 처음으로 9가 아닌 자릿수의 숫자로부터 두 자리 이상 기입해 주는 것이 권장된다.

### 9.5.7 물리량의 단위

가급적 모든 물리량의 단위를 기입하도록 한다. 선형 회귀 그래프에서 단위 기입이 요구되는 물리량은 조작 변인, 종속 변인, 기울기, 절편 등이 있다.  $R^2$ 과 같은 무차원량도 단위가 없음을 기입해 주는 것이 권장되며,  $\ln, \sin, \exp$  등 변수의 무차원화가 요구되는 함수의 경우에는 함수 내부에도 변수의 단위를 기입해 주는 것이 권장된다. (무차원화되지 않은 경우) 예를 들면  $\ln(P/\text{[atm]})[-]$  등의 표기를 사용할 수 있다.



앞서 이론적 예측을 기반으로 회귀 모델을 선정하고, 변수의 선형화를 통해 절편과 기울기로 분석하는 방법에 대해 논하였으며, 여러 가지 기본 회귀모델에 대해 살펴보았다. 이 장에서는 복잡한 회귀 모델에 대해 어떻게 대응할지에 대해 논의한다.

10.1 수치 회귀 . . . . .	91
10.2 Gradient Descent . . . . .	92
10.3 전산 회귀에서의 불확실도 분석 . . . . .	95
10.4 Convolution Fit . . . . .	97

## 10.1 수치 회귀

Gaussian, Lorentzian Fit 등은 대칭분포 피크의 분석에 광범위하게 적용된다. 그러나, 물리학에서는 비대칭적이며 극값을 갖는 분포를 가지는 개형의 현상이 자주 등장한다. 대표적인 예시로는 Planck Curve와 Maxwell-Boltzmann Distribution을 생각해 볼 수 있다. 이처럼 weight function이 붙어 비대칭적인 분포를 가지는 경우, 이론적인 개형 예측이 가능하다면 회귀를 통해 비례계수나 중요한 의미를 갖는 상수들을 실험치로부터 얻어낼 수 있으나 (가령 Planck Curve로부터 Boltzmann Constant의 추산을 시도하는 경우), 이를 위해서는 복잡한 함수에 대한 회귀를 시도해야 한다. 선형 회귀의 경우 계산기나 스프레드시트를 이용해 쉽게 기울기와 절편, 불확실도 등을 얻어낼 수 있고, Gaussian / Lorentzian Fit 등은 Origin 등의 여러 상용 자료해석 툴에서 기본적으로 지원하므로 용이하게 회귀할 수 있다. Python이나 MATLAB 등의 고급 프로그래밍 언어를 사용하는 경우에도 Curve fit 기능을 제공하기도 한다. 그러나 복잡한 관계식을 가지거나, 등분산 가정을 적용하지 않는 등의 Customized된 분석을 수행하고자 할 경우 전산 수치 해법을 통해

회귀를 수행할 수 있다.

## 10.2 Gradient Descent

수치적인 방법을 이용해 회귀를 수행하는 경우, 등분산 가정이거나 이분산 가정이거나 등에 의해 제곱오차의 정의가 달라지는 것은 7장과 동일하나 이 제곱오차를 최소화 (최적화) 하는 방법이 편미분 대신 수치해석에 의존한다는 점이 다르다. 함수의 극값을 구하기 위해서는 구속 조건을 걸고 Lagrange 승수법을 쓰는 경우도 있으나, Iterative algorithm을 구현할 경우 수렴성이 보장되지 않는 경우도 있다. 최근에는 인공지능 및 뉴럴 네트워크 기술의 발달로 Gradient Descent를 이용한 함수 최적화 기법이 널리 사용되고 있는데, 이는 복잡한 형태의 회귀 모형도 비교적 손쉽게 수렴시킬 수 있다는 장점을 가지는 Easy and Robust method라는 장점을 가지므로 간략히 소개하도록 한다. 이 기법은 Python이나 MATLAB, C언어 등 기본적인 반복연산을 수행할 수 있는 기초 수준의 프로그래밍 경험이 있다면 손쉽게 적용해 볼 수 있다.

Gradient Descent는 복잡한 형태의 다변수 회귀 모델을 제곱 오차를 줄이는 식으로 최적화하기에 적합한 모델이다. 이는 다변수 연립방정식의 수치해를 구하는 Newton-Raphson법과 거의 비슷하게, 기울기를 이용해 현재 값에서 최적점을 향해 수렴하는 방향으로 반복 계산하는 알고리듬인데, Newton-Raphson법과의 차이점은 도함수로 함수값을 나누어 한 번에 해로 추정되는 위치에 보내는 것이 아닌, 매우 느린 Learning Rate을 두어 도함수의 방향으로, 도함수에 비례하지만 매우 작은 변화만을 주어 수렴성을 해치지 않게 주의하는 것이다.

등분산 회귀의 경우 제곱오차를 정의할 때 회귀 편차 중  $Y$ 방향 성분만을 고려하며, 측정 오차는 반영하지 않는다. 이에 따라 얻어진 회귀 편차를 각각의 데이터 포인트들에 대해 균일한 가중치로 제곱합하여 제곱오차를 얻는다.

이분산 회귀의 경우, 조작 변인의 무분산성을 상정할 수 있는 경우와 그렇지 않은 경우로 나누어 논한다. 전자의 경우,  $Y$ 방향으로 회귀 편차를 계산하되, 측정 오차의 크기로부터 얻은 분산의 역수를 가중치로 평균을 내어 제곱오차를 얻는다. 이때 측정 오차의 크기는 회귀 계수의 변화에 의존하지 않으므로 가중치를 규격화하지 않아도 제곱 오차를 최적화하는 데에는 지장이 없다.

후자의 경우,  $Y$ 방향 뿐 아니라  $X$ 방향에 대해서도 편차를 계산하여야 한다. 그러나 직선에서의 경우와 달리, 복잡한 회귀선의 경우에는 회귀선으로부터 데이터 포인트가 떨어진 편차를 2차원 거리로 계산하기 쉽지 않다. 이 때는, 데이터 포인트  $(X_i, Y_i)$ 에 대하여  $Y_i$ 와,  $X_i$ 를 회귀식에 대입하여 얻어진  $Y_i$  사이의 편차를  $\Delta_{Y,i}$ ,  $X_i$ 와,  $Y_i$ 를 회귀식에 대입하여 얻어진  $X_i$  사이의 편차를  $\Delta_{X,i}$ 라 하자. 이 때 데이터 포인트가 회귀선으로부터 떨어진 거리  $\Delta_i$ 는 다음 관계식을 통해 얻어낼 수 있다.

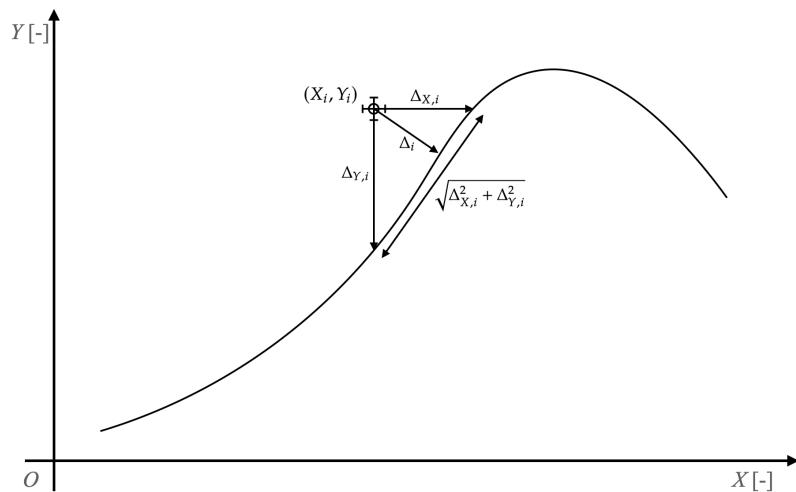
$$\Delta_i = \frac{\Delta_{X,i}\Delta_{Y,i}}{\sqrt{\Delta_{X,i}^2 + \Delta_{Y,i}^2}}$$

이 때,  $\Delta_i^2$ 을 더해 주되, 이들에 대해 분산 역수만큼의 가중치를 반영해 주어야 한다. 결과적으로 얻어지는 제곱 오차는 다음과 같이 표현된다.

$$E^2 = \sum_i \frac{\Delta_{X,i}^2 \Delta_{Y,i}^2}{(\Delta_{X,i}^2 + \Delta_{Y,i}^2)(\sigma_{X_i}^2 + \sigma_{Y_i}^2)}$$

이 때에도  $\sigma_{X_i}$ 와  $\sigma_{Y_i}$ 의 경우 회귀 계수의 값에 의존하지 않으므로,

가중치를 규격화하지 않아도 제곱 오차의 최적화에 아무런 문제가 없다.



**Figure 10.1:** 2차원 거리를 이용한 회귀 편차 계산

이러한 방법으로 정의된 제곱 오차를 최적화하기 위한 기법을 Gradient Descent를 이용해 구현한다. 이 경우 각각의 회귀계수를  $Y_i$ ,  $X_i$ 로 편미분한 기울기를 알아야 하는데, analytic한 표현식이 없는 경우에도 수치 미분을 통해 기울기를 얻어낼 수 있다.

등분산 선형회귀의 경우 다음과 같은 예시로 Gradient Descent 법을 적용할 수 있다. (python 사용) 배열  $X, Y$ 에는 회귀를 위한 자료가 들어 있으며, csv 자료를 읽어올 때 python의 경우 pandas 라이브러리를 이용하면 편리하게 자료를 입력받을 수 있다. (참고: <https://towardsdatascience.com/linear-regression-using-gradient-descent-9>)

```

1 | X = [0., 1., 2., 3., 4., 5., 6.] # Sample Data Set ; Y = 1 + 2X
2 | Y = [1., 3., 5., 7., 9., 11., 13.]
3 |
4 | n = len(X)
5 |
6 | def f(A, B, x) : return A+B*x # Fitting Model
7 |
8 | def err (A, B) : # Squared Error : to be minimized
9 |   E = 0.
10|   for i in range (n) :
11|     E += pow( Y[i] - f(A, B, X[i]) , 2.)
12|   return E
13|
14| B = 0. # slope
15| A = 0. # Y-intersect
16|
17| L = 1.e-4 # learning rate
18| epochs = 10000 # gradient descent loop
19|
20| e = 1.e-10 # Very small value
21|
22| # Gradient Descent
23| for i in range(epochs):
24|   dA = (err(A+e,B) - err(A,B)) / e
25|   dB = (err(A,B+e) - err(A,B)) / e
26|   # Update A, B ; if diverge, reduce L
27|   A -= L * dA
28|   B -= L * dB
29| print (A, B)
30|
31| # Result : 0.9915966663559919 2.0019837579155686

```

Figure 10.2: Gradient Descent : Example Code (Python)

## 10.3 전산 회귀에서의 불확실도 분석

전산 회귀에서도 불확실도의 논의가 가능하다. 앞서 기술한 바와 같이 회귀 불확실도는 (1) 측정 오차가 회귀 편차보다 큰 경우와, (2) 측정 오차보다 회귀 편차가 큰 경우에 대해 다른 방법으로 추산되어야 한다. 전자의 경우 이분산 회귀를 수행하여야 하며, 측정 오차로 얻어진 분산이 회귀 계수의 오차에 전파된다. 후자의 경우 조작변인의 무분산성과 종속변인의 등분산성을 상정하여 회귀할 수 있으며, 회귀 계수들의 불확실도는 데이터 포인트의  $Y$ 방향 회귀 편차로부터 전파된 것이다.

(2)의 경우 비교적 간단한 기술이 가능하다. 이는 등분산 선형 회귀에서와 같이  $X$ 방향 편차가 무시되며,  $Y$ 방향 편차의 기댓값  $\sigma_Y$ 는  $Y$ 방향으로의 균평균제곱 (rms) 편차를 얻어 알 수 있다. 이를 구할 때  $Y$ 방향 편차를 제곱하여 모두 더한 후 자유도로 나누어 제곱근을 취하면,  $1\sigma$  유의수준의  $\sigma_Y$ 를 구할 수 있다. 이는 RMSE에  $\sqrt{\frac{N}{v}}$ 를 곱한

값과 같은데, 이는 표본분산에 대한 계산이기 때문에  $N$  대신 자유도  $\nu$ 로 나누어 주어야 하기 때문이다. 자유도는  $N$ 에서 얻어내고자 할 회귀계수의 수를 빼준 값과 같으며, 절편이 있는 선형회귀의 경우  $A, B$  두 개의 회귀계수를 얻어야 하므로 자유도  $\nu = N - 2$ 가 된다.

이후  $\sigma_Y$ 로부터 회귀계수의 불확실도를 얻어내는 방법은, 회귀계수를  $Y_i$ 로 편미분한 기울기를 rms 평균하여 평균 기울기를 얻는다. 이 때 rms 평균에는 위와 다르게 자유도 대신  $N$ 을 분모에 사용한다. (8장에 같은 논의가 있다.) 각각의 편미분값은 수치 미분을 통해 구할 수 있으며, 데이터 포인트 중 하나의  $Y$ 에 아주 작은 값  $e$ 를 더해 얻어진 회귀계수가 최종 회귀계수로부터 가지는 편차를  $e$ 로 나누어 주면 수치 편미분값을 얻을 수 있다. 이를 각각의 데이터 포인트들에 대해 거듭하여, 수치적 방법을 통해 회귀계수의 불확실도가  $\sigma_Y$ 의 몇 배인지를 얻어낼 수 있다.

한편 (1)의 경우 등분산 가정이 성립하지 않으므로 각각의 회귀계수의 불확실도는 회귀 편차  $\sigma_Y$ 가 아닌 각각의 데이터 포인트들이 가지는 측정오차  $\sigma_{X_i}, \sigma_{Y_i}$ 로부터 얻어진다. 위와 같은 방법으로 회귀계수가 각각의 데이터 포인트의 위치에 얼마나 가파르게 의존하여 변화하는지 회귀 계수를  $X_i$ (등분산 회귀의 경우에는 해당 없음),  $Y_i$ 로 편미분한 기울기를 수치 미분을 통해 얻을 수 있다. 이때 등분산 가정이 성립하지 않으므로 이 기울기들을 rms 평균을 내는 대신, 회귀계수의  $Y$ 방향 기울기에 측정치의  $Y$ 방향 측정 불확실도  $\sigma_{Y_i}$ 를 곱한 값을 각각 제곱하여 모두 더한 뒤 ( $X$ 방향에 대해서도 마찬가지), 제곱근을 취하면 각 회귀계수의  $1\sigma$  신뢰 수준 우연오차를 얻을 수 있다. 이는 8장에서 논의한 것과 같은 내용이며, Gradient Descent 알고리즘과 연계하여 손쉽게 구현할 수 있다. 다만 반복 연산으로 인해 계산량이 많으므로 전산 수치처리 기법에 의존해야 한다.

## 10.4 Convolution Fit

물리량의 실험적 측정에서 측정기기의 응답 특성 (Response)이 균일하지 않은 경우가 있다. 이 경우, 실제 실험계가 가지는 데이터의 개형과 측정기기의 응답특성이 Convolution된 채로 측정되게 되는데, 이 경우 측정기기의 응답을 역산하는 것은 실험 잡음에 취약 하므로, 대신 회귀 모델에 측정 기기의 응답을 Convolution을 통해 반영해 회귀하는 방법이 있다. 이 경우 측정기기의 응답에 대한 정확한 값을 모른 채 개형만 알더라도, 회귀를 통해 실제 실험계의 거동 뿐 아니라 측정기기의 응답에 대한 정보까지 얻어낼 수 있다는 장점을 누릴 수 있다.

간단한 예를 들면, 단색광(Monochromatic light)에 의한 간섭 무늬는 이상적인 Sinusoidal 꼴을 가질 것이다. 그러나 빛의 파장에 약간의 분산이 있을 경우, linewidth에 따라 간섭 무늬에는 약간의 퍼짐이 발생할 것이며 이는 광원의 스펙트럼 분포를 간섭 무늬에 Convolution을 취한 형태로 얻어진다. 이러한 자료를 회귀 분석할 때 Convolution fit을 적용할 수 있다.

Gradient Descent 알고리즘은 매우 Robust한 수렴성을 보장하기 때문에, convolution을 수치 적분한 함수를 이용해서도 수행할 수 있다. 이는 앞 절에서 언급된 내용의 범주 내에서, 조금 더 복잡한 함수를 정의하여 사용하는 식으로 해결해 볼 수 있다.

상용 라이브러리를 사용하는 경우, OriginLab의 다음 자료를 참고하여 Convolution fit을 시도해 볼 수 있다. (본 절의 첫 번째 문단은 아래 링크를 참고하여 작성하였다.)



## 11.1 개요

이산적인 분포를 가진 데이터 포인트들로부터 연속된 함수를 얻기 위해서는 데이터의 보간이 필요하다. 이론적으로 예측된 회귀 모형이 있는 경우 회귀 기법을 사용하는 것이 더욱 강한 통계적 근거를 가지는 방법이 되며, 이론적 예측이 없더라도 비슷한 개형을 제시할 수 있는 회귀 모형을 찾을 수 있다면 계산량이나 오차 분석 측면에서 그러한 접근이 훨씬 유리하다. 그러나, 가령 온도에 따른 비열 정보를 측정하였는데 6차 이상의 다항식을 사용해도 개형을 잘 반영하는 회귀 모형을 찾아내지 못하고, 큰 회귀 편차를 가지는 경우 인터플레이션을 검토할 수 있다. 이는 9.3절에서 소개한 회귀 계수의 육안 추정과 다르며, 9.4절의 보간법과 결을 같이 한다.

데이터의 경향이 평범하지 않은 곡선을 따르는 경우, 적절한 회귀 모형을 적용하지 못할 수도 있다. 가령, 쉬운 예시로 Gaussian과 Lorentzian Fit을 널리 적용할 수 있지만, 한쪽이 가파르고 한쪽이 완만한 비대칭 분포는 쉽게 피팅하기 어렵다. 플랑크 곡선이나 맥스웰-볼츠만 분포를 생각해 보면, 얼추 비슷한 개형을 가진 곡선임에도 위의 대칭 분포를 가지는 회귀 모델을 적용하면 매우 큰 편차를 가진다.

이외에도 이론적인 예측 모델이 없거나, 측정치 플롯의 개형이 복잡하다면 회귀를 수행할 수 없다. (개형이 단순하다면 이론 예측이

11.1 개요 . . . . .	99
11.2 보간 함수의 종류 . . . . .	100
11.2.1 상수 보간 (Nearest-neighbor interpolation) . .	100
11.2.2 선형 보간 (Linear interpolation) . . . . .	101
11.2.3 다항 보간 (Polynomial interpolation) . . . . .	101
11.2.4 3차 스플라인 (Cubic Spline Interpolation) . . .	102
11.2.5 Padé 근사 . . . . .	104
11.3 데이터 샘플링과 불확실도	104
11.4 맷는 말 . . . . .	108

없어도 선형화 / 이차 함수 근사 / 가우시안 / 로렌치안 핏으로 높은 상관 계수를 얻을 가능성도 있다.) 이러한 경우 측정치들로부터 측정되지 않은 값을 얻어내기 위해 Interpolation<sup>o</sup> 사용된다.

보간법의 특징은 오차를 무시한다는 점에 있다. Discrete한 데이터를 continuous function으로 이어주는 과정이므로, 각각의 데이터 포인트들을 오차 없이 지나는 (혹은 점으로부터 궤적을 특정지을 수 있는 구속조건들을 정확하게 부여받은) 연속함수를 찾아내는 과정이며, 이들은 몇몇 수학적 규칙을 통해 제시된다. 얼마나 부드러운 함수를 찾을지, 혹은 얼마나 간편한 함수를 찾을지는 인터폴레이션 방법의 설정에 따라 좌우된다. 불규칙한 이산 자료로부터 연속함수를 만들어낼 때에는 무한한 자유도가 있는데, 그 중 많이 쓰이는 방법들을 소개하자면 다음과 같다.

## 11.2 보간 함수의 종류

### 11.2.1 상수 보간 (Nearest-neighbor interpolation)

가장 간단한 보간은 상수 보간이다. 이는 측정된 조작 변인 중 가까운 쪽의 종속 변인을 가지고록 보간함수를 제시하는 것이며, 측정된 조작 변인 근처에서의 좁은 구간에서는 함숫값이 일정하고, 그 경계에서는 불연속적으로 함숫값이 튀게 된다. 보간함수  $f(X)$ 를 제시하면 다음과 같다.

$$f(X) = Y_i \text{ s.t. } |X - X_i| \leq |X - X_j| \text{ for } \forall j$$

위와 같은 정의는 측정된 조작변인 두 개의 정 가운데에서 어떤 값을 가지느냐에 대한 Ambiguity가 있는데, 물리학에서 쓰이는 연속함수를

이 보간법으로 다룰 경우에는 대체로 그런 특이점의 거동을 신경쓰지 않아도 될 정도인 경우가 많다. 실제 프로그래밍에 적용할 경우에는 작은 쪽이나 큰 쪽 중 적당한 방향 기준을 잡아 중앙에서의 함숫값을 한쪽으로 몰아 주는 방식으로 재정의될 것이다.

### 11.2.2 선형 보간 (Linear interpolation)

상수 보간은 매우 단순하지만, 보간의 의미가 크게 없다. 이를 약간 보정한 것이 선형 보간인데, 이는 각각의 데이터 포인트들을 모두 지나되 보간함수를 인접한 두 데이터 포인트들을 잇는 선분들로 구성하는 것이다. 이 경우 보간함수는 연속하나, 보간함수의 기울기는 각각의 데이터포인트들의 위치에서 연속성을 보장받지 못한다.

얻고자 하는 보간값이  $X$  위치에서의  $Y$ 값이라 할 때, 좌측으로 인접한 데이터 포인트를  $(X_l, Y_l)$ , 우측으로 인접한 데이터 포인트를  $(X_r, Y_r)$ 이라 하면 다음과 같이 선형 보간을 수행할 수 있다.

$$Y = Y_l + (Y_r - Y_l) \frac{X - X_l}{X_r - X_l}$$

이때 불확실도는 오차의 전파를 이용해 다음과 같이 얻어진다.

$$\sigma_Y = \sqrt{\sigma_{Y_l}^2 \left( \frac{X_r - X}{X_r - X_l} \right)^2 + \sigma_{Y_r}^2 \left( \frac{X - X_l}{X_r - X_l} \right)^2}$$

### 11.2.3 다항 보간 (Polynomial interpolation)

선형 보간은 매끄럽지 않다. 기울기까지 연속한 보간함수를 제시하기 위해 다항 보간이 사용될 수 있는데, 이는  $N$ 개의 데이터

포인트를 모두 지나는  $M$ 차의 다항식을 찾아내는 것이다. 자유도를 고려하면,  $M = N - 1$ 임을 알 수 있다. 이 때,  $N - 1$ 차의 다항식은  $N$ 개의 계수를 가지며 이들은  $N$ 개의 점을 지나도록 요구받는 조건 하에 있으므로, 유일한 해를 가지게 된다. 이 경우 해들이 뚜렷한 경향을 가지지 않는다면 실제 실험계의 응답과는 완전히 다른 양상의 구불구불한 결과를 얻을 것을 유의하여야 한다.

다항 보간은 선형 보간과 달리 미분가능성이 보장되므로 기울기까지 연속된다. 다항 보간을 위해 다음과 같은 Lagrange Basis Function을 정의한다. 이는  $N$ 차 다항함수이면서  $X = X_j$  일 때는 1을 얻고,  $X = X_{k \neq j}$  일 때는 반드시 0을 얻는다.

$$L_{N,j}(X) = \prod_{k \neq j} \frac{X - X_k}{X_j - X_k}$$

이로부터 다항 보간함수를 구하면 방법으로는,  $(N + 1)$ 개의 데이터 포인트를 기술하는  $N$ 차 다항 보간 함수를  $N$ 개의 Lagrange Basis의 선형 결합으로 나타낼 수 있다.

$$\tilde{Y}(X) = \sum_{j=0}^N Y_j L_{N,j}(X)$$

이때 불확실도는 다음과 같이 기술된다.

$$\tilde{\sigma}_Y(X) = \sqrt{\sum_{j=0}^N \sigma_{Y_j}^2 L_{N,j}^2(X)}$$

#### 11.2.4 3차 스플라인 (Cubic Spline Interpolation)

계산량이 너무 많은 다항 보간 대신, 데이터 전체를 지나는 것 이 아닌 일부 구간의 데이터만을 지나는 스플라인 함수 여러 개를 구간별로 지나게 하고, 각각의 스플라인끼리 원하는 수준의 연속성

(함수값, 기울기, 곡률 등)을 가지도록 조건을 부여하여 구간함수로의 보간을 제시할 수 있다.

많이 쓰이는 것은 3차함수를 이용한 스플라인 보간(Cubic Spline)인데, 이 때에는 인접한 2개의 점을 잇는 3차함수를 구하되, 구간의 경계에서 함수값과 기울기, 그리고 곡률까지 연속하도록 경계 조건을 부여한다. 이 때 다음 과정을 통해 구간별로 스플라인 보간 함수를 얻어낼 수 있다.

먼저  $H_i = X_{i+1} - X_i$ ,  $C_i = (Y_{i+1} - Y_i)/H_i$ ,  $U_i = 2(H_{i-1} + H_i)$ ,  $V_i = 6(C_i - C_{i-1})$ 를 정의할 수 있다. 이때 삼중대각행렬로 표현할 수 있는 (가우스 소거법이나 Numpy 등의 라이브러리를 이용해 풀 수 있음) 다음 연립방정식의 해  $\{Z_2, \dots, Z_{N-1}\}$ 을 구한다.

$$\begin{cases} Z_1 = 0 \\ H_{i-1}Z_{i-1} + U_iZ_i + H_iZ_{i+1} = V_i ; \quad (2 \leq i \leq N-1) \\ Z_N = 0 \end{cases}$$

이때, 각각의 구간 위에서 다음과 같은 스플라인 보간 함수를 얻는다.

$$S_i(X) = \frac{Z_{i+1}}{6H_i}(X-T_i)^3 + \frac{Z_i}{6H_i}(T_{i+1}-X)^3 + \left( \frac{Y_{i+1}}{H_i} - \frac{H_i}{6}Z_{i+1} \right)(X-T_i) + \left( \frac{Y_i}{H_i} - \frac{H_i}{6}Z_i \right)(T_{i+1}-X)$$

$$\tilde{Y}(X) = \begin{cases} S_1(X) ; & X_1 \leq X \leq X_2 \\ \vdots \\ S_{N-1}(X) ; & X_{N-1} \leq X \leq X_N \end{cases}$$

불확실도의 계산은 복잡하므로 생략하나, 이전의 경우와 같은 방법으로 얻어낼 수 있다.

### 11.2.5 Padé 근사

사용할 수 있는 또 다른 인터폴레이션 기법은 Padé 전개를 이용하는 것이다. Padé 전개는 유리 다항식 (두 다항식의 비율)으로 합수값을 근사하는 것으로, 여러 초월함수들이나 특이점 (pole을 가지는 등)을 가지는 함수에 대해서도 근사가 가능하다. 얻어진 데이터들의 Padé 근사는 Scipy나 MATLAB 등의 상용 데이터 처리 라이브러리에서 제공하며, 이를 이용하면 쉽게 데이터의 보간을 할 수 있다. Padé 근사/전개에 대한 자세한 기술은 이 책의 범위를 넘어선다.

## 11.3 데이터 샘플링과 불확실도

인터폴레이션을 할 때 데이터 포인트들의 우연오차가 크면 그들을 모두 지나기 위해 매우 구불구불하고 부정확한 곡선을 얻게 된다. 이들을 막기 위해서는 적절한 샘플링을 통해 Smoothen된 데이터 포인트들을 얻어야 한다. 이 때, 몇몇 Smoothing 기법에 대해 논의할 수 있는데, (1) 몇몇 데이터들을 그룹으로 묶어 평균내어 하나의 데이터 포인트와 불확실도를 얻는 방법과, (2) 각각의 데이터 포인트들에 대해 주변 몇 개의 데이터들에 대한 가중 평균으로 부드러워진 데이터들을 얻는 기법, 그리고 (3) Gaussian Blur를 적용해 볼 수 있다.

이러한 기법을 적용하기 위해서는 결과로 얻어지는 함수가 급격한 변화나 큰 곡률 없이 완만한 경향을 가지고 있음을 예비 실험이나 대강의 정성적인 이론적 예측을 통해 보장할 수 있어야 한다. 위의 세 샘플링 방법에 대해 간단히 다루면 다음과 같다.

### (1) 데이터들의 그룹으로부터 얻은 평균과 불확실도

가장 간단한 샘플링의 방법으로는, 데이터 셋을 작은 크기를 가진 몇 개의 그룹으로 나눠 이들의 평균을 구하는 것이다. 이 경우 각각의 데이터 포인트들이 가진 측정 오차를 무시하고, 각 데이터 포인트들을 등분산 가정 하에 t-분포를 이용한 반복 측정에 의한 불확실도 및 최확치 (= 평균) 추산을 할 수 있다. 그 과정은 6장에 기술된 것과 같다. 여러 개의 데이터를 묶을수록 표본이 커져 표본분산이 감소하고, 우연오차의 영향을 더 잘 걷어낼 수 있다. 그러나 이 경우 데이터 포인트의 수가 줄어 일부 구간에서 원래의 개형을 잘 못 잡아낼 수 있다는 한계를 가진다.

### (2) 가중치를 둔 주변 데이터들의 반영

위에서는 데이터 포인트 여러 개를 하나의 그룹으로 묶어 하나의 데이터로 취급하였기 때문에, 사용할 수 있는 데이터의 갯수가 줄어든다. 이 경우 단순히 노이즈를 제거하는 것을 넘어 데이터의 편차를 늘릴 우려도 있으므로, 데이터의 수를 줄이지 않으면서 노이즈를 줄여 부드럽게 만든 데이터를 만드는 방법도 고려해 볼 수 있다. 이를 위해서는, 주변 값들의 영향을 어느 정도 받아들이되, 멀리 있는 값은 덜 받아들이고, 가까이 있는 값의 영향은 강하게, 그리고 원래 위치에서 얻어진 데이터의 영향은 가장 강하게 받아들이도록 할 수 있다.

이 때 몇 가지 규칙이 있는데, 먼저 가중치를  $w_i$ 로 정의하여  $w_0$ 은 자기 자신의 반영 가중치,  $w_i$ 들은 자신으로부터  $i$ 번째 떨어진 이웃이 반영되는 가중치라 하자. 이 때, 모든  $i$ 에 대해  $w_i > w_{i+1} > 0$ 이 성립하도록, 즉 monotonic한 가중치를 갖도록 가정할 필요가 있다. 또한, 가중치가 대칭 분포를 가지고 해 줄 필요가 있다. 이 때 주변 데이터의 영향을 반영하는 것이 데이터의 크기 규모를 해치지 않기

위해서는 가중치의 규격화가 이루어져야 한다.

가중치가 대칭적으로 정의되었으므로,  $X_j$  위치의 데이터가  $X_i$  위치의 데이터에 반영되는 비율을 합할 때  $i$ 를 고정시키고  $j$ 를 합하는 것이나,  $j$ 를 고정시키고  $i$ 를 합하는 것이나 동일하다. 이 때  $X_j$  위치의 데이터가 여러  $X_i$  위치에서 영향을 받는다고 할 때, 이러한 ‘퍼짐’의 총 합은  $X_j$  위치 데이터의 원래 값과 같은 크기를 가져야 한다. 이를 위한 조건은  $w_0 + 2 \sum_{i>0} w_i = 1$ 이 될 것을 요구한다.

이에 따라 얻어지는 주변 데이터들의 반영은 다음과 같이 나타낼 수 있다.

$$\tilde{Y}_j = Y_j w_0 + \sum_{i>0} (Y_{j+i} + Y_{j-i}) w_i$$

이들의 불확실도는 오차의 전파 기법을 통해 다음과 같음을 얻는다.

$$\tilde{\sigma}_{Y_j} = \sqrt{\sigma_{Y_j}^2 w_0^2 + \sum_{i>0} (\sigma_{Y_{j+i}}^2 + \sigma_{Y_{j-i}}^2) w_i^2}$$

### (3) Gaussian Blur

이는 근본적으로 (2)와 유사하나, (2)의 경우 가중치를 두고 주변 값을 참조하여 일부 반영한다는 개념이지만 이 경우에는 주변 값에 가우시안 형태의 영향을 전파하는 형태로 Smoothing을 수행한다는 차이가 있다. 가우시안 블러를 수행할 경우 규격화된 (그러나 분산은 조절할 수 있는) 정규 분포를 사용함으로서 주변 값과의 혼합으로 스케일이 변하는 것을 막을 수 있다. 즉, 주변으로 영향을 미친 만큼 중앙값은 낮아지도록, 가우시안 분포의 밀넓이 총합은 일정하게끔 기술된다면 각각의 데이터 포인트들이 주변에 가우시안의 형태로 가산되더라도 데이터 셋의 전체 합의 크기는 크게 변하지 않는다. 물론 이는 데이터의 수가 충분히 많아 연속분포와 비슷하게 기술할 수 있는 경우의 이야기이며 이산 분포에서는 오차가 생길 수 있다.

연속 분포의 극한에서, 이는 함수에 Gaussian Convolution을 취한 것과 같은 결론을 얻는다.

한편, 가우시안은 대칭 분포이므로 부호를 바꿔 주어도 (순서를 바꿔 주어도) 같은 가중치를 가진다. 이는 (3) 역시 (2)의 일종으로 기술될 수 있음을 의미한다. 이 경우 (3)은 (2)에서의 가중치가 가우시안 분포를 가지는 것과 같은 것으로 간주할 수 있다. 각각의 가중치들을 조절하여 최적치를 찾아 주어야 하는 (2)와 달리, 분산만을 조절하여 단 한 개의 파라미터만을 이용하여 재현 가능한 Smoothing<sup>o</sup> 가능하게 하므로 편리하다.

공식은 규격화된 가우시안 분포를  $g(X, \sigma)$ 라 하고, 얻어진 이산 데이터를  $(X_i, Y_i)$ 라 할 때,  $X_i$  위치에서의 처리된 데이터  $\tilde{Y}_i(\sigma) = \sum Y_j g(X_j - X_i)$ 로 얻을 수 있다. 이 때  $\sigma$ 를 바꾸어 주면 분산에 따라 얼마나 데이터를 blur할 것인지의 정도를 조절할 수 있다. (이 때 조작변인의 무분산성을 상정한다.)

이 경우 불확실도는 오차의 전파를 이용해 구할 수 있다. 조작변인의 무분산성을 상정하므로, 이 경우 선형합으로 구해진 데이터와 달리, 불확실도는 피타고拉斯 합으로 합산되게 된다.

$$\tilde{\sigma}(\sigma) = \sqrt{\sum_j (g(X_j - X_i)\sigma_{Y_j})^2}$$

한편 인터폴레이션의 통계적 유의성은 쉽게 보장되지 않는다. 데이터가 충분히 완만하다는 것이 정성적인 이론 분석이나 손행연구, 예비 실험 등으로 보장되는 경우에는 인터폴레이션으로 얻은 보간함수가 어느 정도의 불확실도 이내로 잘 들어맞음이 보장되지만, 스무딩 과정에서 우연오차(노이즈)의 제거 시 물리적으로 유의미한 피크가 사라졌을 수 있으며, 인터폴레이션 과정에서는 이러한 피크의 존재성에 대해 통계적인 근거를 제공하지 못한다. 가령 예를 들면 흡광도

측정에서 몇몇 공명 피크가 얻어졌을 때, 공명 근처는 Lorentzian Fit을 하고 나머지 구간을 인터폴레이션하려 하나, 크기가 작은 피크들이 묻히는 것에 대해서는 이것이 물리적인 의미를 가진 피크인지 노이즈인지 구분할 수 있는 근거를 제공하지 못한다는 것이다. 이는 주변 노이즈의 크기와 비교하여 앞에서 언급된 실험 검정 기법을 적용하여 유의해야 하는 outlier가 있는지를 항상 살펴야 하며, 이러한 점이 인터폴레이션을 적용하기 전의 주의 사항이다.

실험계의 거동이 충분히 완만하다고 가정할 수 있는 경우 (숨겨지거나 묻힌 피크가 존재하지 않음을 보장할 수 있는 경우 : 예를 들어 이론적인 연속성이 예측되는 경우 등) 보간함수가 실제 실험계의 거동을 얼마나 잘 표현하는지는 오차의 전파를 이용하여 얻어낼 수 있다. 전산 기법으로 계산되는 인터폴레이션에서 데이터 포인트 하나를  $Y$ ,  $X$ 방향으로 각각 아주 조금 움직였을 때의 기울기를 이용하여 수치미분을 하고, 이를 토대로 각각의 측정 오차를 반영하여 피타고라스 합을 하면 보간 함수의 불확실도를 얻는다. 이 때 피타고라스 합은 유의 수준을 보간하므로 얻어지는 보간함수의 불확실도는 각각의 측정 오차들의 유의수준을 따른다.

## 11.4 맷는 말

넓은 의미의 인터폴레이션은 본래 실험으로 얻어진 데이터 셋의 정의역 범위 내에서, 측정하지 않은 조작변인에 대한 종속변인의 값을 추정하는 것으로, 앞서 논의한 회귀 분석도 인터폴레이션의 일종으로 포함하는 경우도 있다. 그러나 앞서 살펴본 바와 같이 회귀 분석과 이외의 인터폴레이션은 그 목적과 사용 방법에 있어 다소 상이한 특성을 가지고 있으며 오차의 전파나 통계적 유의성의 측면에서 약간 결을 달리하고, 수학적 기반 또한 다르기 때문에 구분지어 논의하였다. 이러한 데이터 보간 기법들을 잘 이해하는 것은 실험의

정밀도를 어느 정도로 확보하여야 원하는 결과를 얻을 수 있는지와, 어떠한 변인을 제어하여 어떠한 결과를 얻을 것인지의 실험 설계를 최적화하여 수행하는 데 필요한 강력한 도구가 될 것이다.

자연계는 본래 불확실하다. 이러한 불확실도의 장막을 걷어내고, 편향된 추정 없이 최확치를 참값에 최대한 가까이 들여다 볼 수 있는 실험을 설계한다면 상상하지 못했던 일들이 가능해질 수도 있다. 물리학에 남겨진 여러 미해결 문제들은 매우 작은 차이를 분간해내기 위한 날카로운 실험적 근거를 요구하는 경우가 많은데, 오차 해석 및 실험 통계 기법의 적절한 적용을 숙지하는 것은 실험 물리를 다루는 물리학자들에게 비슷한 환경, 비슷한 예산으로 훨씬 잡음이 적고 정밀한 실험을 설계할 수 있는 직관을 제공할 것이다.