Historia Mathematica Vol. 12, No. 1, 1999, 1-12

가우스의 오차론에 근거한 정규분포 배경의 역사적 고찰*

인하대학교 수학통계학부 **具滋興**

Abstract

The first part of this thesis discusses the types and the properties of errors, one of which makes up systematic errors of measurements, removable by detecting their causes, the other errors of accidental causes which can not be removed.

The final part of this thesis deals with the historical background of the Gaussian distribution by Hershel, Hagen, Laplace and Gauss from the late 18th century to the early 19th century

It can be concluded that the accidental idea and the treatment of accidental error distribution by Gauss is the best one based on the assumption that the most probable value of true value is the arithmetic mean of data, obtained by repeated measurements of a given quantity.

0. 서론

주어진 양(量)의 관측값에 대한 참값(眞値, true value)과의 오차(誤差, absolute error)의 절대값을 $\varepsilon(>0)$이라 하자. 그러면 일반적으로 주어진 길이를 재거나 무게를 달아 저울의 눈금을 읽어 기록하는 과정에서 범하게 되는 오차의 값으로는 ① 양의 오차, ② 음의 오차, ③ 영의 오차 중에서 어느 한 경우를 상정할 수 있다.

한편, 실험의 설계나 얻어진 결과들을 분석하기 위해서는 여러 가지 양들의 측정과정을 통하여 얻어지는 측정값들의 오차와 목적값(目的値)들의 오차를 최소화(最小化)하는 것이 바람직하다. 그리고 실제의 경우 모든 측정값들은 참값이 아닌 오차를 수반한 근사값(近似値)으로써 얻어진다.

* 본 연구는 1998년도 인하대학교 교내 연구비에 의하여 수행된 것임.

그러므로 주어진 실험에서 소정의 목표를 달성하기 위해서는 필요로 하는 측정값들의 오차를 분석하고 측정과정에 개입(介入)되는 계통오차(systematic error)들의 원인(cause)들을 찾아내어 제거(除去)하고, 각각의 직접 측정값들의 오차들이 간접 측정값인 목적값의 오차에 끼치는 관계를 규명하여 목적값 오차를 최소화하는 것이 바람직하다고 하겠다.

그러나 측정이나 관찰과정에서 수반되는 오차들 중 우발오차(accidental error)와 같이 그 원인을 밝혀내어 그 원인을 제거할 수 없는 오차도 수반된다. 이 우발오차에 대하여서는 그 분포 법칙을 찾아내어 주어진 크기의 오차가 발생하게 될 확률을 제시해줄 수 있고 또 분포 상태를 보여줄 수 있는 확률분포 모형이 오래 전에 여러 학자들에 의하여 제시되었다.

이러한 연구들 중 대표적인 오차분포가 바로 가우스 분포(Gaussian distribution)이다.

주어진 양의 측정값이 그 원인들을 제거할 수 없는 우발오차만을 수반하는 경우를 고려할 때, 측정값은 매번 다른 데이터 값으로 얻어지는 통계량(확률변수)으로 볼 수 있다. 그리고 이때 측정값을 주는 통계량의 분포는 주로 가우스(Gauss), 하겐(Hagen), 라플라스(Laplace) 등 저명한 수학자들에 의하여 가우스 분포의 수리적모형으로 밝혀졌다.

그 당시만 해도 "모든 통계량의 확률분포는 오직 가우스 분포에 따른다."고 믿고 있었다. 그러던 중 푸아송 분포(Poisson distribution)의 발견으로 다른 확률분포를 따르는 통계량도 존재한다는 사실이 밝혀졌다.

사실이 그렇다고는 하지만 가우스 분포의 연구와 그 분포에서 유래된 기대값(expectation), 표준오차(standard error) 등 확률론에서의 기본 용어들과 정의들은 오늘날까지 확률론과 통계학의 이론적 토대가 되었음은 주지의 사실이며, 가우스 분포에 관해 여러 가지 고전적 연구들을 재조명해볼 가치가 충분하다고 하겠다.

이상에서도 언급하였듯, 가우스 분포는 확률론의 기원이 되었으며, 그 당시 확률론의 중심 연구 과제가 오차론에 바탕을 둔 가우스의 오차법칙의 규명에 있었다.

따라서 본 연구에서는 고전적 확률론의 중심과제였으며, 현대 통계학의 이론적 바탕이 되는 정규분포의 효시(嚆矢)라고 볼 수 있는 가우스 분포에 관한 고전적 연구를 재조명하고, 아울러 그 모수(parameter)들의 성격들도 해석(解釋)해 보기로 하겠다.

1. 오차론과 오차분포

1.1. 오차론의 기원

오차론에 관한 연구는 최초의 정밀과학(精密科學)이라고 할 수 있는 천문학(天文學), 측지학(測地學) 등의 발달과 연관되어 18세기말부터 19세기초에 왕성하게 이루어져왔다.

이들 학문들의 현장연구를 통하여 주어진 하나의 양(量)을 측정해야 할 경우 그 측정환경의 완전 무결한 정비와 통제란 불가능하였으므로 매 측정마다 다소 서로간의 상이한 값들이 관측되었다. 즉 하나의 유성(遊星)의 궤도요소(軌道要素, orbit element)를 결정하기 위한 주어진 요소(要素)에 대한 측정에서 서로 다른 데이터 값이 얻어졌을 경우 이들 값을 어떻게

조정할 것인가? 하는 것이 절실한 연구과제로 제기되었다.

한편 18세기말에서 19세기초에 행하여진 유럽의 대규모 토지측량(土地測量)의 경우에도 같은 유형의 문제가 제기되어 그 조정(調整)의 필요성이 같은 과제로 되었다. 좀더 구체적 사료(史料)로서 가우스도 1821년에서 1825년 사이에 행하여진 『독일국토의 삼각측량(三角測量)』에 참가하였고, 이때 얻어진 자료정리과정에 오차론에 바탕을 두고 최소제곱법(least square method)을 사용하였다. 이때 오차의 조정방법으로서는 대개 반복측정을 통한 산술평균(arithmetic mean)이 타당한 조정 방법이었으며 따라서 산술평균과 관련지어 오차의 성질을 발견하려는데 노력이 경주되었다.

위와 같은 오차론에 관한 연구들은 아주 오래 전부터 코테(Cote), 심프슨(Simpson), 베르누이(Bernoulli) 등에 의하여 이미 이루어져왔었으나 가장 확실하고 위에서 언급한 성질에 부합되는 오차분포법칙을 제안한 저작물들로서는 라플라스의 『확률의 해석적 이론(Théorie analytique des probabilités(1812)』과 가우스의 『측지문제에 관하여』 등이었다.

그 후 더 구체적으로 오차 분포에 관한 여러 가지 증명들이 Enche(1831), Bessel(1838), Hagen(1837), Tait(1867), Crofton(1870) 등에 의하여 주어졌다.

그러나 오차분포를 토대로 한 자료들의 조정을 위한 최소제곱법의 제창자는 르장드르(Legendre, 1806)와 가우스(1821)이었다.

1.2. 오차의 분포법칙에 관하여

일반적으로 오차는 주어진 양의 측정값에서 그 참값을 뺀 차이값(差異値)이다. 그리고 관례상 오차를 그 원인에 따라 계통오차와 우발오차로 구분한다. 그런데 전자인 계통오차는 그것을 일으키는 원인이 분명하고 그 원인을 가려내어 제거 가능한(removable) 경우이다. 이에 반하여 후자인 우발오차는 그 원인이 판명(判明)될 수 없어 제거할 수 없는 원인에 의하여 발생되는 오차이다.

뿐만 아니라, 우발오차는 그 원인들이 규명될 수 없는 여러 가지 미세오차(微細誤差)들의 누적으로 초래되는 오차로서 측정과정에 불가피하게 수반되는 백색잡음(白色雜音, white noise)으로서 일종의 확률변수의 실현치들로 해석할 수 있는 오차이다.

우발 오차의 분포법칙에 관한 고전적 연구들은 다른 확률분포 법칙들보다 훨씬 앞서 선행되었으며, 이 오차분포에 관한 연구들은 확률론에서 여러 가지 극한정리(limiting theorem)들의 최초의 발단(發端)이라는 관점으로도 그 역사적 의미가 깊은 것이라 할 수 있다.

오차분포는 그 기대값이 영인 정규분포(normal distribution)로 그 확률밀도함수(probability density function)는 다음과 같은 가우스 분포의 확률 밀도 함수로 상정되었다.

$$f(x)=\frac{h}{\sqrt{\pi}}\cdot\exp[-\frac{1}{2}h^2x^2] \quad \cdots(1)$$

위와 같은 분포법칙의 밀도함수에 관한 유도과정으로서는 허셜(Hershel), 하겐, 가우스 등에 의하여 수행되었다.

① 허셜의 해석적 유도과정

다소 약식 증명이기는 하지만 그 접근방식이 평이하고 해석적(解釋的)이다.

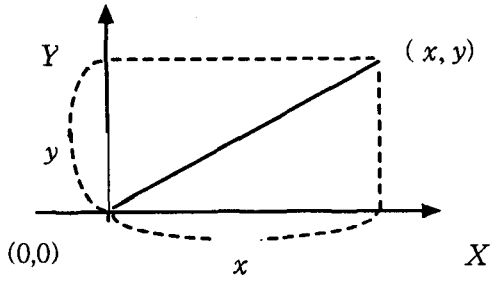
좌표평면 상에서 그 참값이 원점 (0, 0)이라고 하자. 그리고 그 측정값(observed value)을 좌표점 (x, y)이라 할 때, 만일 오차의 분포가 그 방향(方向)과는 무관(無關)하고, 단지 원점으로부터의 거리만이 관계되며 또 직교(直交)하는 두 방향의 오차는 서로 독립(independence)이라고 가정한다.

구체적인 예로서, 좌표평면의 원점 위에서 원점을 겨냥하고 작은 공을 떨어뜨려 공이 떨어진 위치를 기록하도록 한 경우가 바로 위 실험모형의 적합한 본보기라고 할 수 있다.

이때 오차분포의 밀도함수를 φ라고 하면 주어진 가정에 의하여 다음 관계식이 성립한다.

$$\varphi(x^2+y^2)=\varphi(x^2)\cdot\varphi(y^2) \quad \cdots(2)$$

이때 x와 y는 각각의 원점 (0, 0)에서 직각 방향으로의 빗나간 변의 길이이다. 즉 떨어진 공이 X-축 방향으로 x만큼의 오차(빗나감)를 일으키고 이에 직교축인 Y-축으로 y만큼의 오차를 일으켰으며 이들 오차들로서 크기 $\sqrt{x^2+y^2}$ 만큼의 오차를 발생하였다고 해석할 수도 있다. 그리고 오차 x와 y는 서로 독립이다.



한편 함수 방정식

$$f(x+y)=f(x)\cdot f(y) \quad \cdots(3)$$

의 해 $f(z)$는 다음과 같다.

$$f(z)=\exp[hz]$$

한편 $f(z)$가 확률밀도함수임을 감안하면 임의의 실수 h에 대하여 식 (2)의 해는 다음과 같이 표현된다.

$$\varphi(x^2)=k\cdot\exp[-h^2\cdot x^2] \quad \cdots(4)$$

그리고 $\varphi(x^2)$이 확률밀도함수가 되도록 비례상수 k를 결정하여 주면 오차분포법칙에 관한 확률밀도함수 식 (1)을 얻게 된다.

② 하겐의 유도

오차 ξ가 ξ와 $\xi+d\xi$ 사이의 값으로 일어나게 될 확률, 즉 위 구간에서의 발생도수의 전체 발생도수에 대한 비(比)를 $f(\xi)d\xi$로 나타낼 때, $f(\xi)$를 오차 ξ의 발생 확률밀도(發

生確率密度)라 한다.

그러면 확률밀도의 정의에 의하여 다음이 성립한다.

$$\int_{-\infty}^{\infty} f(\xi)\,d\xi = 1$$

이때, 오차 ξ에 관한 확률밀도함수 $f(\xi)$는 앞서의 식 (1)로 주어진 가우스 분포의 밀도함수로 얻어진다.

이제부터 논하려는 오차는 모두 우발오차로서 다음 조건들에 부합되는 것들이라야 한다.

<조건 1> 각 관측오차는 모두 총 n개(상당히 큰 n)의 소위 근원오차요소(根源誤差要素)들의 총합으로 이루어지고,

<조건 2> 근원오차요소들은 어느 것이나 일정한 절대오차 ε를 가지며, 양의 오차 ε과 음의 오차 $-\varepsilon$의 발생확률은 동등하다고 가정한다.

이때 1회의 관측에서 나타나는 n개의 근원오차들 중 p개가 음($-\varepsilon<0$)인 것으로, 나머지 $(n-p)$개가 양($\varepsilon>0$)인 경우, 우발오차 ξ의 값은 다음과 같다.

$$\xi = (n-2p)\varepsilon \quad \cdots(5)$$

그리고 그 발생도수는 다음과 같이 n개 중 p개를 뽑아내는 방법의 수와 같다.

$$\binom{n}{p} = \frac{n!}{(n-p)!\,p!} \quad \cdots(6)$$

마찬가지로, n개의 근원오차 중 $(p-1)$개가 음으로 나타나게 되는 경우에 우발오차와 그 발생도수는 각각 다음과 같다.

$$(n-2p+2)\varepsilon = \xi + 2\xi \quad \cdots(7)$$

$$\binom{n}{p-1} = \frac{n!}{(n-p+1)!\,(p-1)!} \quad \cdots(8)$$

따라서 위 두 경우 오차 발생의 확률밀도의 비는 다음과 같다.

$$\frac{f(\xi+2\varepsilon)}{f(\xi)} = \frac{p}{n-p+1} \quad \cdots(9)$$

그러므로 다음을 얻는다.

$$\frac{f(\xi+2\varepsilon)-f(\xi)}{f(\xi)} = \frac{f(\xi+2\varepsilon)}{f(\xi)} - 1 = \frac{p}{n-p+1} - 1 = \frac{2p-n-1}{n-p+1}$$

마찬가지 방법으로 다음을 얻는다.

$$\frac{f(\xi+2\varepsilon)+f(\xi)}{f(\xi)} = \frac{f(\xi+2\varepsilon)}{f(\xi)} + 1 = \frac{p}{n-p+1} + 1 = \frac{n+1}{n-p+1}$$

그러므로 n이 아주 클 경우 근사적으로 다음과 같이 쓸 수 있다.

$$\frac{f(\xi+2\varepsilon)-f(\xi)}{f(\xi+2\varepsilon)+f(\xi)} = -\frac{n-2p+1}{n+1} \fallingdotseq -\frac{n-2p}{n} = -\frac{\xi}{n\varepsilon} \quad \cdots(10)$$

한편 $f(\xi)$를 ξ의 연속함수라고 할 때, 그것의 테일러 급수전개에 의하면 다음과 같이 둘 수 있다.

$$f(\xi+2\varepsilon) \fallingdotseq f(\xi)+2\varepsilon\frac{df(\xi)}{d\xi} \quad \cdots(11)$$

또 식 (10)은 근사식으로 다음과 같이 쓸 수 있다.

$$\frac{f(\xi+2\varepsilon)-f(\xi)}{2f(\xi)} = \frac{2\varepsilon\frac{df}{d\xi}}{2f(\xi)} = -\frac{\xi}{n\varepsilon} \quad \cdots(12)$$

그러므로 식 (12)에서 우변의 등식에서 다음 등식을 얻는다.

$$\frac{1}{f(\xi)} \cdot \frac{df}{d\xi} = -\frac{\xi}{n\varepsilon^2} \quad \cdots(13)$$

$$\ln f(\xi) = -\frac{\xi^2}{2n\varepsilon^2} + \ln k \quad \cdots(14)$$

여기서 k는 상수이다.

이제 $h^2 = 1/2n\varepsilon^2$이라 두면, 식 (14)에서 다음을 얻는다.

$$f(\xi) = ke^{-h^2\xi^2}$$

그리고 함수 $f(\xi)$가 앞서 언급한 확률밀도 함수의 성질을 만족하도록 k를 결정하면 오차 ξ의 확률밀도 함수 식 (1)의 결과를 얻는다.

③ 드 무아브르 -라플라스의 정리

위의 하겐의 유도과정에서 주어진 <조건 1>과 <조건 2>와 똑같은 조건 밑에서 우발오차 ξ의 n개의 근원오차들 중 k개가 음의 오차 $-\varepsilon$으로 발생하게 됨으로써 일어나게 될 우발오차 ξ의 확률밀도는 이항분포 $B(n, ½)$에 따르게 된다. 즉

우발오차 $\xi = (n-2k)\varepsilon$ $\cdots(15)$

가 발생하게 될 확률은 이항분포에 의하여 다음과 같이 주어진다.

$$f(\xi) = \Pr(\xi=(n-2k)\varepsilon) = \binom{n}{k}\left(\frac{1}{2}\right)^n \quad \cdots(16)$$

여기서 $k=1,2,3,\cdots,n$이다.

또 근원오차의 총 개수 n이 아주 클 때 $n!$에 관한 스털링(Stirling)의 공식에 의한 정규근사법(normal approximation to binomial distribution)에 의하면 식 (16)은 정규분포(또는 가우스 분포) $N(0, \sigma^2)$에 분포수렴하게 된다. 즉 우발오차 ξ의 값이 실수구간 (s, t) 사이의 값으로 발생하게될 확률은 다음과 같이 구할 수 있다.

$$P(s<\xi<t) \fallingdotseq \frac{1}{\sqrt{2\pi}\sigma}\int_s^t \exp\left[-\frac{x^2}{2\sigma^2}\right]dx \quad \cdots(17)$$

여기서 $\sigma^2 = n\xi^2$이다.

위 우발오차 ξ에 대하여 그 평균과 분산은 각각 다음과 같이 계산된다.

$$E(\xi)=\sum_{i=0}^{n}E(\xi_i)=\sum_{i=1}^{n}[-\varepsilon\frac{1}{2}+\varepsilon\frac{1}{2}]=0$$

$$Var(\xi)=Var(\sum\xi_i)=E[\sum\xi_i]^2-(E(\xi))^2$$

$$=\sum E(\xi_i^2)+\sum_{i\neq j}E(\xi_i\cdot\xi_j)=\sum E(\xi_i^2)=n\varepsilon^2$$

위와 같은 정규근사법에 의한 오차분포의 유도법은 드 무아브르-라플라스의 정리로 전해지며, 대표본론(large sampling theorey)의 핵심원리를 이루는 중심극한정리(central limit theorem)를 암시하는 형태의 발단으로 중요한 의미를 내포하고 있다.

④ 최소제곱법과 가우스 분포

우발오차의 확률분포를 위하여, 앞에서 허셜, 하겐, 라플라스 등의 방법을 고찰하였으나, 좀더 일반적이고, 보다 체계적인 확률론적 접근 방법은 참값 a의 측정값들의 산술평균이 그 최확값(most probable value)이라 가정하고 유도된 가우스에 의한 유도방법이다.

즉 이제 n회의 측정값 $x_1, x_2, \cdots, x_n$들이 얻어졌을 때, 이들 측정치들에 의하여 알려지지 않은(unknown) 참값 a를 추정하기 위한 오차론(오차의 분포)에 대하여 고찰하기로 한다.

이때, 그 특징은 원인확률(原因確率)에 관한 소위 베이즈의 정리(Bayes Rule)를 이용한다는 점이다.

또 측정값들의 분포는 단지 측정값과 참값의 차에만 의존된다고 가정하고 그 밀도함수를 $\Phi(x_i-a)$라고 둔다.

이 경우, 참값 a의 사전확률(a priori probability)을 $\upsilon(a)$라고 가정하고, 또 a의 n개의 측정값 $x_1, x_2, \cdots, x_n$이 얻어졌다면, 그 참값이 a와 $a+da$ 사이에 들어있게 될 확률은 베이즈의 정리에 의하여 다음과 같이 나타낼 수 있다.

$$\upsilon(a\mid x_1, x_2, \cdots, x_n)da=\frac{\upsilon(a)\,\Phi(x_1-a)\cdots\Phi(x_n-a)da}{\int_{-\infty}^{\infty}\upsilon(a)\,\Phi(x_1-a)\cdots\Phi(x_n-a)da} \quad \cdots(18)$$

두 번째 가우스의 가정은 우선 위에 언급한 사전확률 $\upsilon(a)$가 일정(一定)하다고 할 때, a의 분포밀도는 a의 값이 그것의 n개의 측정값들의 산술평균값과 일치(一致)될 때 최대값(maximum value)이 된다는 것이다.

그러기 위하여서는 식 (18)로 주어진 오차분포 밀도 $\Phi(x_i-a)$들의 곱

$$L(a)=\Pi\Phi(x_i-a), \quad (i=1,2,\cdots,n) \quad \cdots(19)$$

이

$$a=\frac{1}{n}(x_1+x_2+\cdots+x_n) \quad \cdots(20)$$

일 때 최대값이 되도록 오차분포의 밀도함수 Φ를 결정하여야 한다.

이제 식 (19)에 대수(對數)를 취하여 $l(a)$라 하면 다음과 같다.

$$l(a)=\sum_{i=1}^{n}\ln\Phi(x_i-a) \quad \cdots(21)$$

또 식 (21)의 양변을 a로 편미분하고, 그 결과식을 영으로 두면 다음과 같다.

$$\frac{\partial l}{\partial a}\equiv\sum_{i=1}^{n}\frac{\Phi'(x_i-a)}{\Phi(x_i-a)}=0 \quad \cdots(22)$$

다시 $F(x)=\Phi'(x)/\Phi(x)$라 둘 때, 식 (22)를 다음과 같이 다시 쓸 수 있다.

$$\frac{\partial l}{\partial a}=\sum_{i=1}^{n}F(x_i-a)=0 \quad \cdots(23)$$

그리고 또 $z_i=x_i-a$라 두었을 때, 만일 $\sum z_i=0$일 때 $\sum F(z_i)=0$이 성립하면, 가우스의 조건이 모두 충족된다. 그러면 함수 $F(z)$와 z 사이에는 서로 정비례(正比例)하게 되므로 이때 비례상수(比例常數)를 C라 하면, 다음과 같은 $F(z)$와 z의 관계식을 얻는다.

$$F(z)=\Phi'(z)/\Phi(z)=Cz \quad \cdots(24)$$

그러므로 식 (24)의 양변을 z에 관해 적분하고 $\Phi(z)$가 확률밀도함수의 성질 $\int_{-\infty}^{\infty}\Phi(z)dz=1$을 만족하도록 적분상수를 결정해주면 가우스 분포의 확률밀도함수가 얻어진다. 즉 식 (24)의 양변을 적분하면, 다음과 같다.

$$\ln\Phi(z)=C\frac{z^2}{2}+\ln k,$$

$$\ln\frac{\Phi(z)}{k}=-h^2z^2 \quad (\because C=-2h^2)$$

$$\Phi(z)=ke^{-h^2z^2} \quad \cdots(25)$$

다음 $\Phi(z)$를 z에 관해 전실수구간 R 상에서 적분하여 1로 두면 식 (25)의 상수값 $k=h/\sqrt{\pi}$를 얻는다. 그러므로 위의 k값을 식 (25)에 대입하면 가우스 분포의 밀도함수가 다시 얻어진다.

그러면 이제부터 가우스 오차분포의 모수(母數) h에 관하여 고찰해 보기로 하겠다. 식 (18)에서, 만일 $v(a)$가 상수일 경우 다음을 얻는다.

$$v(a\mid x_1, x_2, \cdots, x_n)=\prod_{i=1}^{n}\Phi(z_i)$$

$$=\left(\frac{h}{\sqrt{\pi}}\right)^n\exp\left[-h^2\sum(x_i-a)^2\right]$$

$$=\left(\frac{h}{\sqrt{\pi}}\right)^n e^{-h^2s^2n+nh^2(a-\alpha)^2} \quad \cdots(26)$$

여기서 $a=\sum x_i/n$이다. 즉 식 (26)에서 h의 값이 주어지면 미지(未知)인 참값 a의 확률분포가 확정된다.

다른 한편, 참값 a의 값이 주어지면 그것의 측정값들로 $x_1, x_2, \cdots, x_n$이 얻어지게 될 확률밀도는 마찬가지로 다음과 같이 주어진다.

$$\left(\frac{h}{\sqrt{\pi}}\right)^n \exp\left[-h^2\{(x_1-a)^2+\cdots+(x_n-a)^2\}\right]=\left(\frac{h}{\sqrt{\pi}}\right)^n \exp^{-nh^2s^2+nh^2(a-a)^2}$$

그런데 만일 $a=a$인 경우, 식 (26)으로 주어진 확률밀도는 최대값을 갖게되며, 또 위 확률밀도가 최대값을 갖도록 모수 h의 값을 정해주는 것이 가우스의 두 번째 가정이었다. 즉 $a=a$로 두었을 때 식 (26)의 양변에 자연로그(e를 밑으로 하는 로그)를 취하고 다음과 같이 h로 미분한 다음 영으로 둔다.

$$\frac{\partial}{\partial h}\left[\ln h-\ln\sqrt{\pi}-h^2s^2\right]=n\left[\frac{1}{h}-2hs^2\right]=0 \quad \cdots(27)$$

위 방정식 (27)의 해는 다음과 같다.

$$h^2=\frac{1}{2s^2} \quad \cdots(28)$$

따라서 h의 값이 크면 클수록 표본 표준편차 s^2의 값은 작아지며 측정값의 산포도(dispersion)은 작아진다. 그러므로 h를 참값 a에 대한 측정값들의 정도(精度, precision)이라고 부른다.

또 가우스의 분포의 밀도함수에 정도 $h=1/\sqrt{2}s$를 대입하면

$$\Phi(z)=\frac{1}{\sqrt{2\pi}s}e^{-\frac{1}{2s^2}z^2} \quad (-\infty<z<\infty)$$

가 얻어지는데, 측정횟수 n이 매우 클 경우 이것은 바로 평균이 0이고, 그 분산이 s^2인 정규분포 $N(0, s^2)$의 확률밀도함수에 수렴하게 됨을 추측할 수 있다.

2. 관측값의 신뢰도에 대하여

주어진 양(量) a를 측정해서 얻어진 여러 개의 관측값 $x_1, x_2, \cdots$의 분포상태(分布狀態)에서 판단하여 그 관측값들이 어느 정도 신뢰(信賴)할 수 있는가를 고찰하는 수단으로 다음과 같은 오차들을 생각하게 된다.

① 제곱평균오차(error of mean square)

다음과 같이 n개의 측정오차들의 제곱평균의 제곱근를 표준편차(standard deviation)이라고 한다.

$$\sigma=\sqrt{(\xi_1^2+\xi_2^2+\cdots+\xi_n^2)/n} \quad \cdots(29)$$

여기서 $\xi_i=x_i-a$ 이고 a는 참값이다.

우발오차 ξ를 연속확률변수로 보고, 또 가우스 오차 법칙에 따른다고 볼 때, ξ의 분산은 다음과 같이 계산된다.

$$\sigma^2=\int_{-\infty}^{\infty}\xi^2 f(\xi)\,d\xi \qquad (\because E(\xi)=0)$$

$$=\frac{h}{\sqrt{\pi}}\int_{-\infty}^{\infty}\xi^2 e^{-h^2\xi^2}d\xi=\frac{h}{\sqrt{\pi}}\frac{1}{2h^2}\int_{-\infty}^{\infty}e^{-h^2\xi^2}dw$$

$$=\frac{h}{\sqrt{\pi}}\frac{1}{2h^2}\frac{1}{h}\int_{-\infty}^{\infty}e^{-w^2}dw=\frac{1}{2h^2}$$

즉 오차의 분산 σ^2과 측정의 정도 h 사이에는 다음 관계식이 성립한다.

$$\sigma^2=\frac{1}{2h^2} \quad \cdots(30)$$

또 그 제곱근

$$\sigma=\frac{1}{\sqrt{2}h} \quad \cdots(31)$$

를 평균오차(mean error)라고 부른다. 이것은 표준편차 또는 표준오차라고도 부른다. 뿐만 아니라 식 (31)을 h에 관해 풀면, 다음과 같다.

$$h=\frac{1}{\sqrt{2}\sigma} \quad \cdots(32)$$

이것은 이미 언급하였듯이 측정의 정도로서 h가 커지면 표준편차 σ의 값이 작아지므로 측정치 $x_i(i=1,2,\cdots)$들의 평균값인 참값 a 근방에 밀집(密集)되는 것으로 간주할 수 있으므로, h값이 크다는 것은 측정의 정밀도(精密度)가 높아지다는 것을 의미하게 된다. 다른 한편, 위의 가우스 분포의 밀도함수에 모수 $h=1/\sqrt{2}\sigma$를 대입하면

$$\Phi(z)\,dz=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}z^2}dz$$

을 얻게 되고, 또 $z=x-a$라 두면 x의 확률요소(probability element)는 다음과 같다.

$$f(x)dx=\Phi(x-a)\,dx=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-a)^2}dx \quad \cdots(33)$$

즉 다음과 같이 측정치를 나타내주는 확률변수 x의 확률밀도함수를 얻는다.

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-a)^2} \qquad (-\infty<x<\infty)$$

이때 확률변수(참값 a의 측정치를 나타내주는 확률변수) X는 정규분포 $N(a,\sigma^2)$에 따른다는 것을 알 수 있다.

② 절대 평균오차

다른 한편, 각 측정오차의 절대값의 평균값 θ도 역시 평균오차라고도 한다. 즉, θ는 다음과 같이 구해진다.

$$\theta = \frac{|\xi_1| + |\xi_2| + \cdots + |\xi_n|}{n}$$

$$= \int_{-\infty}^{\infty} |\xi| f(\xi) d\xi = 2\int_0^{\infty} \xi f(\xi) d\xi = 2\frac{h}{\sqrt{\pi}} \int_0^{\infty} \xi e^{-h^2\xi^2} d\xi$$

$$= 2\frac{h}{\sqrt{\pi}} \frac{1}{2h^2} = \frac{1}{\sqrt{\pi}h} \quad \cdots(34)$$

그런데 이 측정 오차들도 하나의 연속확률변수 ξ의 n개의 실현치들이라 간주하면, 식 (34)에서 h의 값이 커지면 평균오차 θ의 값이 작아지므로 역시 h를 측정의 정도(精度)로 해석할 수 있다.

③ 확률오차(probable error)

오차의 절대값이 임의의 양의 실수 ε보다 작아질 확률이나 또는 ε보다 커질 확률이 서로 같다고 할 때, ε를 확률오차(確率誤差), 개연오차(蓋然誤差) 또는 공산오차(公算誤差)라고 부른다. 즉, 다음을 만족하는 ε을 확률오차라고 부른다.

$$2\int_0^{\varepsilon} \xi f(\xi) d\xi = 2\int_{\varepsilon}^{\infty} \xi f(\xi) d\xi = \frac{1}{2}$$

따라서 가우스 분포의 경우는 다음과 같다.

$$2\frac{h}{\sqrt{\pi}} \int_0^{\varepsilon} e^{-h^2\xi^2} d\xi = \frac{1}{2} \quad \cdots(35)$$

식 (35)의 좌변의 적분을 I, $h\xi = t$라 두면, $I = \frac{2}{\sqrt{\pi}} \int_0^{h\xi} e^{-t} dt$이다. 또 $v = h\xi$라 두면,

$$\Phi(v) = \frac{2}{\sqrt{\pi}} \int_0^{v} e^{-t^2} dt \quad \cdots(36)$$

와 같이 고쳐 쓰고, 이것을 확률적분(probability integral)이라고 한다.

그리고 확률오차 ε의 값을 구하면,

$$\Phi(h\varepsilon) = \frac{1}{2}$$

을 만족하는 $h\varepsilon$의 값이 0.47694이므로 $\varepsilon = 0.47694/h$이다. 또 $h = 1/\sqrt{2}\sigma$에서 $\varepsilon = 0.47694 \cdot \sqrt{2}\sigma \fallingdotseq 0.6745\sigma$ 인 관계식을 얻는다.

3. 결론 및 요약

① 우발오차의 분포법칙을 유도하는 하나의 고전적 경향으로는 형식상 그것이 따르는 것이 가장 이상적이라고 생각되는 조건을 오차법칙에 달아서 이 조건을 만족하는 함수방정식을 꾸미고, 또 그것을 풀어서 오차분포 법칙을 끌어내는 방법이 바로 허셜의 유도방법이었다.[1)]

② 다음 드 무아브르, 라플라스 등에 의한 근원오차(elementary errors) 분포의 극한정리(approximation theorem)에 의한 유도방법은 그 후 여러 학자들에 의하여 더욱 일반화되었고, 당시 확률론의 중심과제 이었던 중심극한정리의 발견의 단서(端緖)가 되기도 하였다.

③ 더 나아가, "주어진 양(量) a의 참값(神만이 아는 값)의 최확값이 a의 n회의 측정값 $x_1, \cdots, x_n$의 산술평균과 일치한다."는 조건하에 오차분포법칙을 이끌어 낸 것이 확률론적으로 더욱 타당한 유도방법으로 그 유명한 가우스의 오차법칙이다.

④ 본 연구에서는 자세히 언급하지는 않았지만 가우스와 유사한 조건으로 "a의 최확값이 측정값들의 중위수(Median Value) 와 일치한다."는 가정 하에도 가우스 오차 분포의 밀도함수가 유도되었다.[2)]

참고 문헌

1. 瀨正已, **誤差論**, 培風館, 日本, 1953, 14-20.
2. **統計學辭典**, 東洋經濟新報社, 일본, 1970, 475.
3. 구자홍, **확률론**, 자연 과학 대우 학술총서 권 58, 민음사, 1988, 12, 13, 75, 206
4. 주금종, 구자홍, **오차론과 Gauss 분포에 관한 연구**, 인하대학교 교육대학원 석사논문, 1999, 13-18.
5. 정한영 편저, **통계학사개론**, 한림대학교 출판부, 1995, 139-141.

1) Edinbourgh Review (1850)
2) Pólya; Annales de l'Institute Henri Poincaré(1, 1931)