Chapter 10

Data Analysis

10.1 Mean Value and Standard Deviation

Measured value in experiment is usually different from the real value of the physical quantity. The accuracy of measured value depends on measuring device, environment, human error, etc.

Error: difference between a measured value and the real value.
Relative error: ratio of the error to the real value.

Source of systematic error:

- Measuring device; scale, accuracy, ...
- Environment of measurement; temperature, pressure, humidity, ...
- Habit of person.
- Theoretical relation between variables of measurement.

The systematic error can be reduced by careful measurement using more accurate measuring device in well controlled environment. Systematic error can also be compensated after measurement.

Source of random error:

- Small variation of environment.

- Thickness of scale tick mark.
- Small change of human factor.

The random error has no control.
Characteristics of random error in repeated many measurements:

1. The probability occurring negative error is same as the probability of positive error with the same magnitude.
2. The probability of occurring small error is larger than the probability of larger error.
3. The probability occurring a very large error (order of the smallest scale of measuring device) is very small.

Thus if we repeat measurement infinitely many times, then the most probable measured value would be closest to the real value.

According to the characteristics 1 and 2 of random error, for the real value X, the sum of errors of the N measured values x_i is zero if the number of measurement N is infinitely large.

$$\sum_{i=1}^{\infty}(x_i - X) = 0 \implies X = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}x_i$$

For a large but finite N,

$$\sum_{i=1}^{\infty}(x_i - X) \approx \sum_{i=1}^{N}(x_i - x_0) = 0 \implies x_0 = \frac{1}{N}\sum_{i=1}^{N}x_i \tag{10.1}$$

Thus the arithmetic mean value x_0 approaches to the real value X as the number of measurements N becomes larger and larger. On the other hand, according to the characteristics 2 and 3 of random error, as the measured values (data) x_i distribute more sharply around the real value X, the standard deviation σ of the data

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - X)^2 \tag{10.2}$$

is smaller. The standard deviation of data σ is independent of the number of measurements N if N is not too small. This σ does not represent the

reliability of the mean value x_0 to be close to the actual value X. It just represents how sharply the data is distributed.

For the reliability of the mean value x_0, we define the standard deviation of mean value σ_0 as the magnitude of difference between the mean value of data and the real value,

$$\begin{aligned}
\sigma_0^2 &= |x_0 - X|^2 = \left[\frac{1}{N}\sum_{i=1}^{N} x_i - X\right]^2 \\
&= \frac{1}{N^2}\left[\sum_i (x_i - X)^2 + \sum_{i\neq j}(x_i - X)(x_j - X)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}(x_i - X)^2 = \frac{1}{N}\sigma^2. \qquad (10.3)
\end{aligned}$$

The second term in the second line vanished due to the characteristic 1 of random error. Smaller value of the standard deviation of the mean value σ_0 means higher reliability of the mean value x_0 to be close to the actual value X. Thus the reliability of the mean value x_0 becomes better as the number of measurements N becomes larger as $1/\sqrt{N}$.

Since we do not know the real value X we cannot use Eq.(10.2). However we know the mean value x_0. Using Eqs.(10.1) – (10.3),

$$\begin{aligned}
N\sigma^2 &= \sum_{i=1}^{N}(x_i - X)^2 = \sum_i (x_i - x_0 + x_0 - X)^2 \\
&= \sum_i (x_i - x_0)^2 + \sum_i (x_0 - X)^2 = \sum_i (x_i - x_0)^2 + \sigma^2. \qquad (10.4)
\end{aligned}$$

Thus the standard deviation σ of the data and the standard deviation σ_0 of the mean value x_0 are

$$\sigma^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i - x_0)^2, \qquad (10.5)$$

$$\sigma_0 = |x_0 - X| = \frac{\sigma}{\sqrt{N}} \qquad (10.6)$$

The standard deviation of data σ is independent of the number of measurements N for not too small N. It depends only on the distribution form of the measured data. But the standard deviation of the mean value σ_0 is proportional to $1/\sqrt{N}$. Thus the reliability of the mean value is better for larger number of measurements. Finally, the measured value for X is $x_0 \pm \sigma_0$.

10.2 Chi-Square Fitting and Error Analysis

For an experimental value which are measured directly, the arithmetic mean value given by Eq.(10.1) is good for the mean value of data which replaces the real value X. For a physical quantity which is not determined by direct measurement but determined through a relation with some other directly measured quantities, the simple arithmetic mean value cannot represents the real value X. For a physical quantity which can be measured directly, the arithmetic mean value of Eq.(10.1) can be looked as minimizing the standard deviation σ^2 of Eq.(10.2) with respect to the mean value x_0, i.e.,

$$\frac{d}{dx_0}\left[\sum_{i=1}^{N}(x_i - x_0)^2\right] = 2\sum_{i=1}^{N}(x_i - x_0) = 0. \qquad (10.7)$$

This is the least chi square method for mean value of directly measured data.

Various physical quantities are connected through a relation according to the physical law of the phenomena. For indirectly measurable variables a, b, c, ... and directly measurable variables x, y, z, ..., a functional relation

$$f(x, y, z, \cdots; a, b, c, \cdots) = 0 \qquad (10.8)$$

represents the law of the underlying physical phenomena. As an example, Ohm's law is $f(V, I; R) = V - RI = 0$ for determining resistance R by measuring voltage V and current I. For N sets of directly measured values $\{x_i, y_i, z_i, \ldots\}$, the value of $f_i = f(x_i, y_i, z_i, \cdots; a, b, c, \cdots)$ may not be zero exactly due to the error of measurement. The systematic error of f_i can be controlled but the random error of f_i has the same characteristics as the random error of direct measurement. Thus the chi square is defined as the weighted sum of squared error f_i^2

$$\chi^2 = \sum_{i=1}^{N} w_i f_i^2 = \sum_{i=1}^{N} w_i \left|f(x_i, y_i, z_i, \cdots; a, b, c, \cdots)\right|^2 \qquad (10.9)$$

where the weight w_i represents the reliability of the data set $\{x_i, y_i, z_i, \ldots\}$. The chi square per data is χ^2/N. The least chi square fitting method (최소제곱법) determines the values of a, b, c, ... by minimizing the chi square. Through the weight factor w_i in the χ^2, the evaluated values $\{a, b, c, \cdots\}$ are fitted to give smaller error of $|f_i|$ for the more reliably measured value set with larger weight w_i than the less reliable data set with smaller weight w_i.

If a systematic error σ_i varies for data set by set then the weight factor can be set to $w_i = \sigma_i^{-2}$. Eq.(10.7) for direct measurements corresponds to the case of least chi square fitting for $a = X$ with $w_i = 1$, i.e., with a non-varying systematic error ($\sigma_i = \sigma$).

As an example, for the case of determining resistance R through Ohm's law, $V = RI$,

$$\chi^2 = \sum_{i=1}^{N} |f(V_i, I_i; R)|^2 = \sum_{i=1}^{N} (V_i - RI_i)^2, \tag{10.10}$$

$$\frac{\partial \chi^2}{\partial R} = 2\sum_{i=1}^{N} (V_i - RI_i) I_i = 0 \implies R = \frac{\left[\sum_i V_i I_i\right]/N}{\left[\sum_i I_i^2\right]/N} \tag{10.11}$$

Here the resistance R is given as the ratio of the arithmetic mean of power $\overline{VI}$ to the arithmetic mean of current square $\overline{I^2}$. On the other hand, if we use the form of $f(V, I; R) = R - V/I$ for the Ohm's law, then we have

$$\chi^2 = \sum_{i=1}^{N} \left(R - \frac{V_i}{I_i}\right)^2 = \sum_{i=1}^{N} \frac{1}{I_i^2} \left(V_i - RI_i\right)^2, \tag{10.12}$$

$$\frac{\partial \chi^2}{\partial R} = 2\sum_{i=1}^{N} \left(R - \frac{V_i}{I_i}\right) = 0 \implies R = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{V_i}{I_i}\right] \tag{10.13}$$

Now the resistance R is the arithmetic mean $\overline{V/I}$ of resistance $R_i = V_i/I_i$ for each measurement. The Eq.(10.12) is different from the Eq.(10.10) by the factor of I_i^{-2}. In Eq.(10.10), every data set $\{V_i, I_i\}$ are treated with the same weight $w_i = 1$ while each data set $\{V_i, I_i\}$ are weighted with $w_i = I_i^{-2}$, i.e., the smaller current has larger weight in Eq.(10.12). The Eq.(10.12) corresponds to giving a same weight to the data $R_i = V_i/I_i$ rather than to the data set $\{V_i, I_i\}$. If one set of voltmeter and ammeter is used for whole data sets then Eq.(10.10) would be more proper.

As another example, consider measurement of the resistivity ρ of metal at various temperature T (assume here that both ρ and T are measured directly). From these directly measured data sets, we extract two physical quantities of temperature coefficient α of resistivity and the resistivity ρ_0 at temperature T_0. The resistivity ρ and the temperature T are related by

$$f(\rho, T; \rho_0, \alpha) = \rho_0[1 + \alpha(T - T_0)] - \rho = 0. \tag{10.14}$$

The chi square fitting becomes, with $\beta = \rho_0 \alpha$,

$$
\begin{aligned}
\chi^2 &= \sum_{i=1}^{N} \left[\rho_0 + \beta(T_i - T_0) - \rho_i\right]^2, \\
\frac{\partial \chi^2}{\partial \rho_0} &= 2\sum_{i=1}^{N} \left[(\rho_0 - \rho_i) + \beta(T_i - T_0)\right] = 0, \\
\frac{\partial \chi^2}{\partial \beta} &= 2\sum_{i=1}^{N} \left[(\rho_0 - \rho_i) + \beta(T_i - T_0)\right](T_i - T_0) = 0.
\end{aligned}
\tag{10.15}
$$

From these coupled equations for ρ_0 and $\beta = \alpha\rho_0$,

$$
\begin{aligned}
\rho_0 &= \frac{\left[\sum_i \rho_i/N\right]\left[\sum_i (T_i - T_0)^2/N\right] - \left[\sum_i (T_i - T_0)/N\right]\left[\sum_i \rho_i (T_i - T_0)/N\right]}{\left[\sum_i (T_i - T_0)^2/N\right] - \left[\sum_i (T_i - T_0)/N\right]^2} \\
\beta &= \frac{\left[\sum_i \rho_i (T_i - T_0)/N\right] - \left[\sum_i \rho_i/N\right]\left[\sum_i (T_i - T_0)/N\right]}{\left[\sum_i (T_i - T_0)^2/N\right] - \left[\sum_i (T_i - T_0)/N\right]^2} \\
\alpha &= \frac{\left[\sum_i \rho_i (T_i - T_0)/N\right] - \left[\sum_i \rho_i/N\right]\left[\sum_i (T_i - T_0)/N\right]}{\left[\sum_i \rho_i/N\right]\left[\sum_i (T_i - T_0)^2/N\right] - \left[\sum_i (T_i - T_0)/N\right]\left[\sum_i \rho_i (T_i - T_0)/N\right]}
\end{aligned}
\tag{10.16}
$$

Here β is used instead of α itself in minimizing χ^2 since ρ_0 and β form a set of coupled linear equations.

Finally, consider the case of determining the time constant $\tau = RC$ of an R-C circuit. The voltage across the capacitor during the charging process with emf of E is $V(t) = E(1 - e^{-t/\tau})$ which can be represented either by

$$
f(V, t; \tau) = V - E(1 - e^{-t/\tau}) = 0 \tag{10.17}
$$

or by

$$
f(V, t; \tau) = t + \tau \ln\left(1 - \frac{V}{E}\right) = 0. \tag{10.18}
$$

with the directly measured data sets of $\{V_i, t_i\}$. Since the time t appears in the exponent, for the case of Eq.(10.17), the chi square has a larger weight for the data set measured at smaller time than the data set at larger time. In contrast to this, the time t appears linear for the case of Eq.(10.18) and all the measured data sets have the same weight. However more importantly, the form of Eq.(10.18) is much easier in extracting the value of the time constant τ using least χ^2 fitting than the form of Eq.(10.17) since τ appears

linearly in Eq.(10.18) while τ appears exponentially in Eq.(10.17). If we can make $f(x, y, z, \cdots; a, b, c, \cdots)$ linear in the variables a, b, c, $\cdots$ which should be determined by chi square fitting, it is better using linear form of f than using nonlinear form unless the reliability of data set requires otherwise.

For the reliability of mean values of the indirectly measured quantities $\{a, b, c, \cdots\}$, we can define the standard deviation for each mean value a_0, b_0, c_0, $\cdots$ in the same way as the standard deviation Eq.(10.6) of a directly measured quantity. That is, for the standard deviation of mean value of quantity a,

$$\sigma_a^2 = |a_0 - a|^2 = \frac{1}{N}\left[\frac{1}{(N-1)}\sum_{i=1}^{N}(a_i - a_0)^2\right] \tag{10.19}$$

Through the functional relation f, the deviations of data a_i, b_i, c_i, $\cdots$ from their corresponding mean values a_0, b_0, c_0, $\cdots$ are related with the deviation of directly measured data x_i, x_i, z_i, $\cdots$ from their mean values x_0, y_0, z_0, $\cdots$;

$$\begin{aligned} df &= \left(\frac{\partial f}{\partial a}\right) da + \left(\frac{\partial f}{\partial b}\right) db + \left(\frac{\partial f}{\partial c}\right) dc + \cdots \\ &\quad + \left(\frac{\partial f}{\partial x}\right) dx + \left(\frac{\partial f}{\partial y}\right) dy + \left(\frac{\partial f}{\partial z}\right) dz + \cdots \end{aligned} \tag{10.20}$$

The deviation $da_i = a_i - a_0$ of a_i from its mean value a_0 is then, with $db_i = 0$, $dc_i = 0$, $\cdots$

$$da_i = -\left(\frac{\partial f}{\partial a}\right)^{-1}\left[\left(\frac{\partial f}{\partial x}\right) dx_i + \left(\frac{\partial f}{\partial y}\right) dy_i + \left(\frac{\partial f}{\partial z}\right) dz_i + \cdots\right]$$

where $dx_i = x_i - x_0$, $dy_i = y_i - y_0$m and $dz_i = z_i - z_0$. Thus the standard deviation of the mean value a_0 is

$$\sigma_a^2 = \left(\frac{\partial f}{\partial a}\right)^{-2}\left[\left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2 + \cdots\right] \tag{10.21}$$

The standard deviation of mean values b_0, c_0, $\cdots$ are similarly given. Thus once we find the standard deviation σ_x, σ_y, σ_z, $\cdots$ of mean values x_0, y_0, z_0, $\cdots$ for the directly measured quantities, then we can find the standard deviation σ_a, σ_b, σ_c, $\cdots$ of mean values a_0, b_0, c_0, $\cdots$ for the indirectly measured quantities for their reliability.

Some data has a form of linear superposition of special functions of some variables as parameter;

$$y(x) = C_1 g_1(x) + C_2 g_2(x) + C_3 g_3(x) + \cdots \tag{10.22}$$

where $g_1(x)$, $g_2(x)$, $g_3(x)$, $\cdots$ are known functions of parameter x. For this case, the coefficients C_1, C_2, C_3, $\cdots$ can be determined from the measured data sets of $\{x_i, y_i\}$ using the least chi square fitting with

$$f(x, y; C_1, C_2, C_3, \cdots) = y(x) - [C_1 g_1(x) + C_2 g_2(x) + C_3 g_3(x) + \cdots] .$$

The reliability of the mean value of the coefficients can be obtained in terms of the standard deviation of directly measured x and y through Eq.(10.21). Expanding angular distribution in terms of Legendre polynomial is an example of this method. Fourier series expansion or wavelet analysis can also be considered as expanding in terms of special functions using chi square fitting.

10.3 Fast Fourier Transform (FFT)

If the functions $g_l(x)$ in Eq.(10.22) are orthonormalized functions, i.e.,

$$\int_0^L g_l^*(x) g_{l'}(x) dx = \delta_{ll'} \tag{10.23}$$

then the expansion coefficients C_l can be obtained by

$$C_l = \int_0^L g_l^*(x) y(x) dx = \frac{L}{N} \sum_{i=1}^{N} g_l^*(x_i) y_i \tag{10.24}$$

from the measured data sets $\{x_i, y_i\}$.

One such example is the Fourier series expansion of a periodic function $y(x)$ with period L:

$$y(x) = \frac{1}{2} A_0 + \sum_{n=1}^{\infty} [A_n \cos(nk_0 x) + B_n \sin(nk_0 x)] , \tag{10.25}$$

$$A_n = \frac{2}{L} \int_{x_0 - L/2}^{x_0 + L/2} y(x) \cos(nk_0 x) dx, \tag{10.26}$$

$$B_n = \frac{2}{L} \int_{x_0 - L/2}^{x_0 + L/2} y(x) \sin(nk_0 x) dx \tag{10.27}$$

where $k_0 = 2\pi/L$. Since $e^{ix} = \cos x + i \sin x$, Fourier series expansion can also be represented as

$$y(x) = \sum_{n=-\infty}^{\infty} C_n e^{ink_0 x}, \tag{10.28}$$

$$C_n = \frac{1}{L}\int_{x_0-L/2}^{x_0+L/2} y(x) e^{-ink_0 x} dx, \tag{10.29}$$

$$A_n = C_n + C_{-n} \quad \text{and} \quad B_n = i(C_n - C_{-n}). \tag{10.30}$$

For $L \to \infty$, Eq.(10.28) becomes Fourier integral,

$$y(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \eta(k) e^{ikx} dk, \tag{10.31}$$

$$\eta(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y(x) e^{-ikx} dx. \tag{10.32}$$

This is called Fourier transformation.

Fast Fourier transform (FFT) is a computational algorithm to make fast the Fourier transform in discretized $x_j = (j-1)\Delta$ space (data sets of $\{x_j, y_j\}$). As an example, using Eq.(10.29) for Fourier series of Eq.(10.28),

$$C_n = \frac{1}{N}\sum_{j=1}^{N} y_j e^{-ink_0 x_j} = \frac{\Delta}{L}\sum_{j=1}^{N} y_j \left(e^{-ik_0\Delta}\right)^{(j-1)n} = \frac{\Delta}{L} w^{-n} F_n. \tag{10.33}$$

In the second form $w = e^{-ik_0\Delta}$ is calculated once and only the simple multiplication of w^{jn} is needed for each term of $F_n = \sum_j y_j w^{jn}$. The number of multiplications of w for w^{jn} in each term is N^2. For even N,

$$F_n = \sum_{j=1}^{N/2} w^{(2j-1)n} y_{2j-1} + w^n \sum_{j=1}^{N/2} w^{2jn} y_{2j} = F_n^o + w^n F_n^e. \tag{10.34}$$

Since only $N/2$ terms appear in each of F_n^o and F_n^e with multiplication of w^2, the number of multiplication of w^2 for each of these is $(N/2)^2$ and thus $N^2/2$ in total. For N is a power of 2, this reduction can be done for $\log_2 N$ steps and reduces the number of multiplication w can be reduced to be order of N.

10.4 Wavelet Analysis

Fourier series expansion is good only for a periodic function which repeats a same form of function for infinitely many times. Thus we can expand in terms of trigonometric functions which oscillate for infinite number of periods. For the case of function with a form of many localized functions superposed, such as heart beat signal or superposition of many Gaussian functions, we expand using wavelet which corresponds to a wave function in a short range of the parameter variable x. In a wavelet analysis for heart beat, the position and amplitude of the wavelet representing the basic form of heart beat can be extracted from the data. For the case of superposed Gaussians, the amplitude A_i, the position x_i, and the width σ_i of the Gaussian function

$$G_i(x) = A_i e^{-(x-x_i)^2/(2\sigma_i^2)} \tag{10.35}$$

can be extracted in a wavelet analysis. The least chi square method can be used in wavelet analysis.