EXP. 1 **Data and Error**

1. Section Purpose:

A guide to data and error analysis, and an introduction to the relevant application of PC and handheld computers.

2. Theory:

(1) Experimental error

Theoretically, there are real values of the measured physical quantity (e.g., temperature, time, length, etc) However, because of the existence of error, the real value will never be measured accurately, which can only be estimated through data analysis. For example, the arithmetic mean of the physical quantity, which is measured N times with the same measurement method, can be considered as the real physical value（μ'）; that is,

$$\mu' = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (1)$$

Of course, the error between estimated value (μ') and true value (μ) comes from each measurement. Sources of experimental error can be divided into "system errors, " and " random errors". "System error" refers to the error with a certain magnitude and a clear cause. It can be further subdivided into "instrument error", "method error", "personal error" and "environmental error, caused by the instrument itself, human factors, environmental method and the environment of the experiment. If the magnitude of the error is presented in numerical size scattered randomly and no clear form source, it is called "random error." "System error" cannot be easily controlled or analyzed, but there is a set of statistical analysis standard of "random error". Only through such a standard error analysis can we correct expression, analyze data, and further explain the physical meaning of the data.

If N-pieces of measurements are mostly located in the vicinity of $\bar{x}$, we would have more confidence in the repeatability of measurement repeatability have more confidence, feeling the precision of data correct enough. But high “precision” do not represent high “accuracy”. Only after the real value is compared with the measured physical quantity can the "accuracy" level of data or the magnitude of the error be determined. The commonly used data expression way is, Measurement results = $\bar{x}$ +ε% , (2)

And, ε = Relative percentage error = $\frac{|\bar{x}-\mu|}{\mu}\cdot 100$ (2')

What circumstances will the estimate value (μ') become identical to the real value of the measured physical quantity (μ)? The answer to this problem is relatively simple when the measurement data is under "random error" and no "systematic error", that is, when the number of measurements approaches infinity, $\mu = \mu'$, or

$$\mu = \lim_{N\to\infty} \mu' = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} x_i}{N} \quad (3)$$

In fact, we can not measure the data for infinite times, so the error must exist between μ' and μ . How do we calculate, represents the estimate value and the error of measurement data with just N times of random error? The answer to this problem is not only related to a special distribution function of the random error data, but also with the data of the "standard deviation".

(2) Standard Deviation

Standard deviation (σ) reflects the level of accuracy of measurement data. The smaller the standard deviation, the more measurement data which concentrate in the vicinity of the average. Conversely, the larger the standard deviation value, the more dispersed the measurement data. Statistically, the definition of the standard deviation is,

$$\sigma = \sqrt{\lim_{N\to\infty} \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \quad (4)$$

But for finite N times measurement data, what we are concerned about is "sample standard deviation", the formula is,

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}} \qquad (5)^{\text{note1}}$$

The average value $\bar{x}$ is also the estimate value of the measured physical quantity. Of course, when N is large, we can assume that $s \approx \sigma$. Note that though we use the N times measurement data to calculate (samples) standard deviation, but each single measurement data (x_i) corresponds to the (sample) standard deviation σ (or s).

It means that, if the standard deviation of measurement is σ , and data errors are random, then there is a 68% chance of each single measurement, which is within the certain plus or minus scope of the standard deviation.[note2]

If the measurement value corresponds to each of the standard deviation σ , (that is, $x_i \leftrightarrow \sigma$), what the corresponding standard deviation of the estimated value (μ') calculated according to N measured values should be?

According to statistical analysis, we can prove that, [note3]

$$\sigma_\mu = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{N}} \qquad (6)$$

Data and error distribution: Gaussian distribution function

As we mentioned above, the data of random error is based on a particular function. This particular function is Gaussian function. This can be illustrated through the data in table 1A.Table [1A] recorded a total measurement time data 26 times. We can calculate the average $\overline{x}$, estimated value and sample the standard deviation (s) by equation (1) and (5).

Table 1A Time measurement data

The___	Data	The___	Data	The___	Data
measurement	(s)	measurement	(s)	measurement	(s)
1	1.50	11	1.50	21	1.65
2	1.65	12	1.60	22	1.50
3	1.45	13	1.50	23	1.55
4	1.50	14	1.45	24	1.45
5	1.30	15	1.55	25	1.60
6	1.40	16	1.40	26	1.50
7	1.60	17	1.80		
8	1.65	18	1.45		
9	1.75	19	1.55		
10	1.55	20	1.65		
average value $\overline{x}$ =estimate value μ' =1.54					
sampling standard deviation s=0.1					

After classifying table [1A] in accordance with the range of values for each number of measurements, we can get table [1B]; different measurement ranges level of distribution data can also be displayed through bar chart, as shown in [1B].

Data Range	Number of measurements
1.16 ~1.26	0
1.27 ~1.37	1
1.38 ~1.48	6
1.49 ~1.59	10
1.60 ~1.70	7
1.71 ~1.81	2
1.82 ~1.92	0

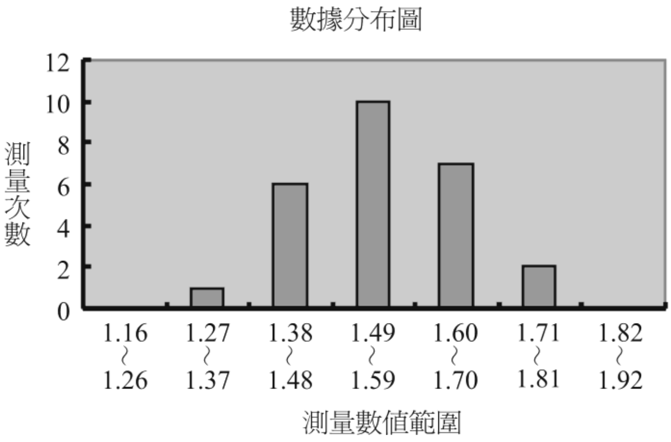

Table 1 B / Figure 1 B Overview of Data Deviation Distribution

The deviation of every data and its average value (i.e. 1.54) in Table[1 B] is not uniform, thus allowing examining the rough data distribution according to the ratio gained from the total measurment numbers (/) divided by the total mearsurement numbers conducted within deviation. The physical meaning reflected by this ratio of measurement number is the probability magnitude of measured value falling in between the deviation range as indicated. For example, 10 measurements in total fall in the deviation range of 0.00±0.05, namely, he data range of 1.49 ~ 1.59, in the data of Table [1 B], and thus the probability magnitude of measured value falling in this deviation range is 10/26 = 0.38.

Range of Deviation	Ratio of Number
-0.33±0.05	0.00
-0.22±0.05	0.04
-0.11±0.05	0.23
0.00±0.05	0.38
+0.11±0.05	0.27
+0.22±0.05	0.08
+0.33±0.05	0.00
Total	1.00

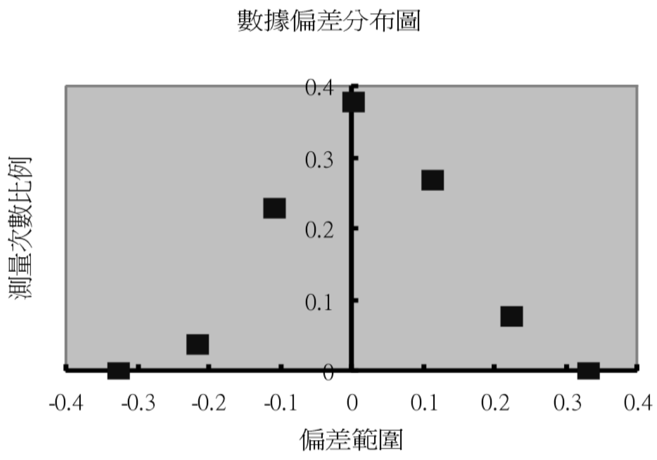

Table 2 A / Figure 2 A Overview of Data Deviation Distribution

From Figure[2 A], we can tell that measurement data and its distribution of random deviation both match the Gaussian function. Of course, if having more measurement data, we can choose a smaller interval of data range or deviation range, so the distribution points can form a smoother curve (Please see Question 1).

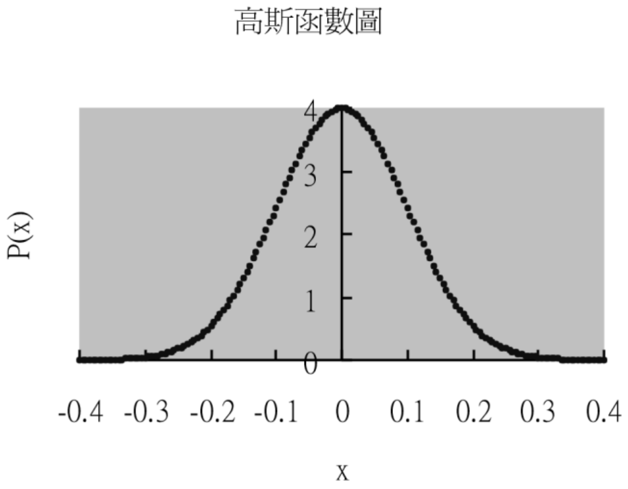

Figure C Gaussian Distribution Function, P(x)

In other words, if the probability of a random measurement result falling right in the data range of x and $x+dx$ is $P(x)dx$, then

$$P(x)=\frac{1}{s\sqrt{2\pi}}\exp[-\frac{1}{2}\frac{(x-\mu')^2}{s^2}] \quad (7)$$

The smooth curve in Figure [3] is a classic Gaussian function curve, and the parameter that it corresponds to is identical to that in Figure [b A], that is $s=0.1, \mu'=0$. Since the result of measurement must fall between $\pm\infty$, the probability of any random measured value falling between $\pm\infty$ will be 100% (or 1); thus,

$$\int_{-\infty}^{+\infty} P(x)dx=1 \qquad (8)$$

And from

$$\int_{-\infty}^{+\infty} x\cdot P(x)dx=\mu' \qquad (9)$$

We can tell that the estimated value of measured physical quantity is also the expected value of measurement data. Of course, the μ' and s in Formula (7), (8), (9) will become μ and σ respectively when the measurement number is big enough (ex. N→∞).

As indicated previously, the area of the Figure [3] curve in the deviation range of 0.00±0.05 is

$$\int_{-0.05}^{+0.05} P(x)dx = erf\,(\frac{0.05}{s\sqrt{2}}) = erf\,(\frac{0.05}{0.1\cdot\sqrt{2}}) = erf\,(\frac{0.5}{\sqrt{2}}) \approx 0.4 \qquad (9')$$

The number represents the "Probability" of measured value falling in this deviation range. And it is very close to the measurement ratio (i.e. 0.38) of this deviation range shown in Figure [2]. The $erf\,(x)$ in Formula (9') is error function.

Representation of Measurement Results

Hence, as the introduction done for aiming the measurement data of Table [A], when having N data, the following can be done:

- Calculate the average value (μ').
- Calculate the sampling standard deviation (*s*).
- Present the measurement results as,

Measured physical quantity＝

$$AverageValue \pm \frac{S\tan dard\ Deviation}{\sqrt{Total\ Measurement\ Number}} \qquad (10)$$

Though the average value (μ') which calculated from these N data has error between the real value (μ) of the measured physical quantity, the magnitude of error is also related to its corresponding magnitude of probability (i.e. "credibility" or "confidence interval"). For example, if we label the error with $\pm\infty$, then from the view of probability, of course its "confidence interval" is 100%; that is, the probability of measured average value(s) falling in the range of $-\infty$ to $+\infty$ will be 100%. The "confidence interval" to which the labeling error of Formula (10) corresponds is 68%. As for the relationship of "confidence interval" and labeling error, please refer to Appendix (B).

(3) Propagation of Errors

Sometimes, the physical quantity we care about will be acquired after operating the data of direct measurement. For example, after measuring the data of time (*T*) and velocity (*V*) directly, we can acquire the magnitude of distance (D) through the operation of ($T\cdot V$). Since the data of direct measurement already has its standard deviation respectively, the physical quantity acquired after operation will also have its standard deviation. The question is, how to estimate the standard deviation of this physical quantity from the standard deviation of measurement data?

The operation formula of the example above is $D = T \cdot V$. If conveyed with the form of function, it will become $D = f(T,V)$; the function *f* within will have two variables *T* and *V*, and $f = T \cdot V$. Actually, any operation can be described by normal function formula. For example, $x = f(p,q,r\cdots)$. The *p, q, r, …* can be seen as the data of direct measurement, and the corresponding standard deviation will be σ_p, σ_q, σ_r, … respectively. And *x* is the physical quantity after function *f* operated, its corresponding standard deviation will be σ_x. The relationships of these variables are presented down below, [Note 5]

$$\sigma_x = \sqrt{\left(\frac{\partial f}{\partial p}\right)^2 \sigma_p^2 + \left(\frac{\partial f}{\partial q}\right)^2 \sigma_q^2 + \left(\frac{\partial f}{\partial r}\right)^2 \sigma_r^2 + \ldots..} \tag{11}$$

$$\bar{x} = f(\bar{p},\bar{q},\bar{r},\ldots.) \tag{12}$$

What should be paid attention is that we can only use Formula (12) to calculate standard deviation when the variables *p, q, r, …* of function *f* are mutually independent. Since the variables *T* and *V* are mutually independent,

$$\sigma_D = \bar{D} \cdot \sqrt{\frac{\sigma_T^2}{\bar{T}^2} + \frac{\sigma_V^2}{\bar{V}^2}} \tag{13}$$

$$\bar{D} = \bar{T} \cdot \bar{V} \tag{14}$$

The σ_D, σ_T, σ_V is the standard deviation of distance, time and velocity respectively; $\bar{D}$ is the estimated value of distance, and $\bar{T}$, $\bar{V}$ is the measured average value of time and velocity respectively.

(D) Method of Least Square Fitting

We often need to analyze the correlations between measured physical quantities in experiments. For example, analyze ones between temperature (*T*) and pressure (*P*) when the volume of perfect gas doesn't change; or analyze those between the elongation of spring (*X*) and its restoring force (*F*). Theoretically, we know the answers to the examples above are both directly proportional, which means $P \propto T$ or $F \propto X$. If showing this correlation on a x-y vertical coordinate diagram, making *P* (or *F*) as *y*-axis, *T* (or *X*) as *x*-axis, then the graph of data will be a straight line; hence, this relationship is also known as the linear relationship. Even to some non-linear relation physical quantity, if conveying with appropriate cartography, it can still show its linear relationship. For example, the relation between the free-fall drop time (t) and falling distance (d) is $d = gt^2/2$. If setting $\log(d)$ as y-axis and $\log(t)$ as x-axis, the relation picture is still a line, that is $\log(d) = 2\log(t) + \log(g/2)$. For general linear relation, we can use linear mathematical formula, $y = mx + b$. However, because of the presence of experimental errors, the actual measured data doesn't perfectly fall on the line but scatter in the vicinity of the line. Scatter density depends on the magnitude of the experimental error. Because the data on the line can be regarded as a "true value", the magnitude of the deviation between "true value" and measured data can be seen as the magnitude of the experimental error. As mentioned above, we can use the "deviance" to quantitatively show the measurement data's deviation from theoretical data on the linear relation. In fact, the problem which we meet more often is how we determine the equation of the "true" equation of line by the scattered distribution of the data. Compared with the "false" equation of line, the so-called "true" equation of line must have the minimum"deviance".Based on the method developed by this is called "Least Square Residual Fitting Method". Take data in Table.4 for example. X-axis and y-axis represent time and measured position respectively, which present that thw data distribution is in a linear relation. Here we assume that the standard deviation of object's position data is s_y and time data (x) has no error. (That is $s_x = 0$.)

Table.4 Measuring the position of objects at different times

Time x(s)	0	1	2	3	4	5	6	7
Position y (m)	1.5	3.0	6.0	5.0	6.0	8.0	8.0	10.0

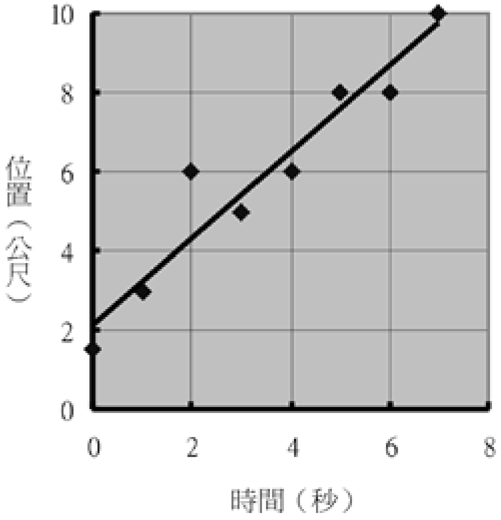

Figure 4 . Relation between time and position

In figure 4 , the data is indeed consistent with a linear relation , $y = mx + b$. But how to calculate m, b and standard deviation s_m and s_b ?

Please refer to the calculation method in Appendix (3).

(5) Relevant Computing Function of Computer and Computer Application Software

Many engineering calculators are equipped with functions of "statistical computing" and "regression calculation". By using "statistical computing", users just need to clip data and will get "mean" and" standard deviation "…etc.Using "regression calculation", users simply clip data and will get "slope, m" and "intercept value, b " …etc. Some applications such as Excel, SigmaPlot, Origin are equipped with more and better functions of "statistical computing" and "regression calculation". Department of Science and Technology students should take the opportunity to learn how to use their own calculators or related applications in the experiment and calculate the value in the formula 2 in appendix (II). (Please see question 5.)

Take Excel for example, when we type data into a workbook, we can use its graphics function to show the relation diagram and also use "linear trend analysis" function in the graphics, which represents the direct linear relation on the graph calculated by "linear regression trend analysis". In "Linear regression trend analysis", using function LINEST can get the statistics generated by the regression model.
For example, INDEX（LINEST（y-axis values，x-axis value），1）can calculate the slope (m) of the linear trend.
INDEX（LINEST（y-axis values，x-axis value），2）can calculate the intercept (b) of the linear trend.See Appendix (3).
Additionally, we can also use built-in statistical functions or handwriting formula on workbooks, calculate mean, average deviation and other parameter... etc.

Take Casio engineering calculator, fx-4500P, for example. We clip data in "statistical model"（SD, mode 3）; follow the key function and then calculate mean, standard deviation, data and other value…etc.
Similarly, we clip data in "regression model"（LR、mode 2）, follow the key function and then calculate constant term A (namely intercept, b) and regression coefficients B (namely slope, m) in "linear regression". See Appendix 4.

3. Instruments Required

An inclined surface slide, a marble, photogate timer (see Appendix 5), phototube×2 (including stand) and steel pipes.

4. Steps to the Experiment
(1) Device as figure 5 indicates. Measure the vertical height of inclined surface.

(2) Place the marble on the top of inclined surface and let it fall along the inclined surface and record the time required to pass two phototubes. Repeat such measurements for 50 times.

(3) Change the vertical height of inclined surface, repeat step2 and record it.

(4) Find arithmetic mean and standard deviation and illustrate the data distribution on experiment report sheet (2). Please refer to the part of principle

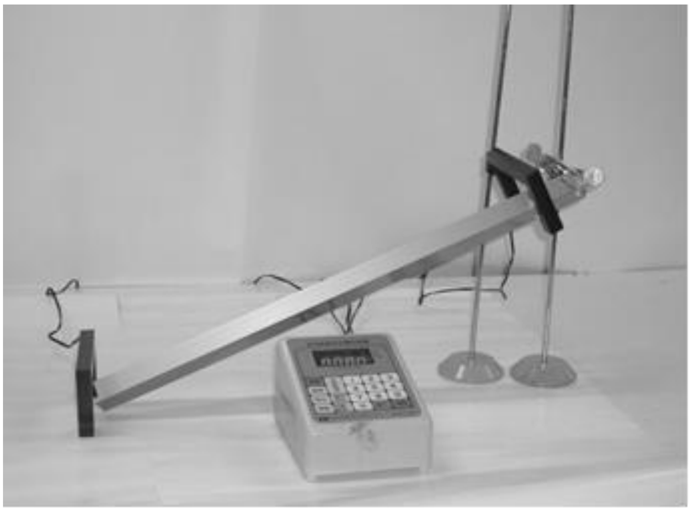

Figure 5

5.Question

(1) After the data points in figure (2, A) increases, will we get smooth Gaussian curve shown in figure 3?

(2) Please prove the area ratio of $\mu' \pm s$ range in Gaussian curve is 68% in figure 3.

(3) Please prove the formula 13 and 14.

(4) There is an instrument used for detecting alcohol concentration in blood whose standard deviation is 0.005%. How many times will take to do measurement if the error is ±0.005% under 95% credibility?

(5) There are three repeatedly-measured data, 0.084, 0.089, 0.079. Please prove that the measurement should be 0.084±0.012 under 95% confidence interval.

Note:

(1) P R. Bevington and D. K. Robinson’s textbook explains why the denominator is N-1 rather than N in chapter 1.

(2) Please see question 5.

(3) Refer to P R. Bevington and D. K. Robinson’s textbook in chapter 4.

(4) Derive and prove it through Binomial Distribution. Refer to chapter 2 in P R. Bevington and D. K. Robinson’s textbook.

(5) Refer to chapter 3 in P R. Bevington and D. K. Robinson’s textbook.