temperature. The data points indicated by the full symbols appear also to hug the trend line. However the data points do not lie on it. This is due to random errors that are always present in any measurement. Actually the standard thermocouple would also have the random errors that are not indicated in the figure. We have deliberately shown only the trend line for the standard thermocouple.

Sub Module 1.3

3. Statistical analysis of experimental data

Statistical analysis and best estimate from replicate data:

- Let a certain quantity X be measured repeatedly to get

$$X_i,\ i=1,n \qquad (1)$$

- Because of random errors these are all *different*.
- How do we find the best estimate X_b for the true value of X?
- It is reasonable to assume that the best value be such that the measurements are as precise as they can be!
- In other words, the experimenter is confident that he has conducted the measurements with the best care and he is like the skilled shooter in the target practice example presented earlier!
- Thus, we minimize the variance with respect to the best estimate X_b of X.
- Thus we minimize:

$$S = \sum_{i=1}^{n} \left[X_i - X_b \right]^2 \qquad (2)$$

- This requires that:

$$\frac{\partial S}{\partial X_b} = 2\sum_{i=1}^{n}[X_i - X_b](-1) = 0$$

$$\text{or } X_b = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (3)$$

- The best estimate is thus nothing but the mean of all the individual measurements!

Error distribution:

When a quantity is measured **repeatedly** it is expected that it will be distributed around the best value according to some **distribution**. Many times the random errors may be distributed as a **normal distribution**. If μ and σ are, respectively, the mean and the standard deviation, then, the probability density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} \qquad (4)$$

The probability that the error around the mean is $(x-\mu)$ is the area under the probability density function between $(x-\mu)+dx$ and $(x-\mu)$ represented by the product of the probability density and dx. The probability that the error is anywhere between $-\infty$ and x is thus given by the following integral:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left[\frac{v-\mu}{\sigma}\right]^2} dv \qquad (5)$$

This is referred to as the **cumulative probability**. It is noted that if $x \to \infty$ the integral tends to 1. Thus the probability that the error is of all possible magnitudes (between $-\infty$ and $+\infty$) is unity! The integral is **symmetrical** with

respect to x=μ as may be easily verified. The above integral is in fact the **error integral** that is a tabulated function. A plot of f(x) and F(x) is given in Figure 5.

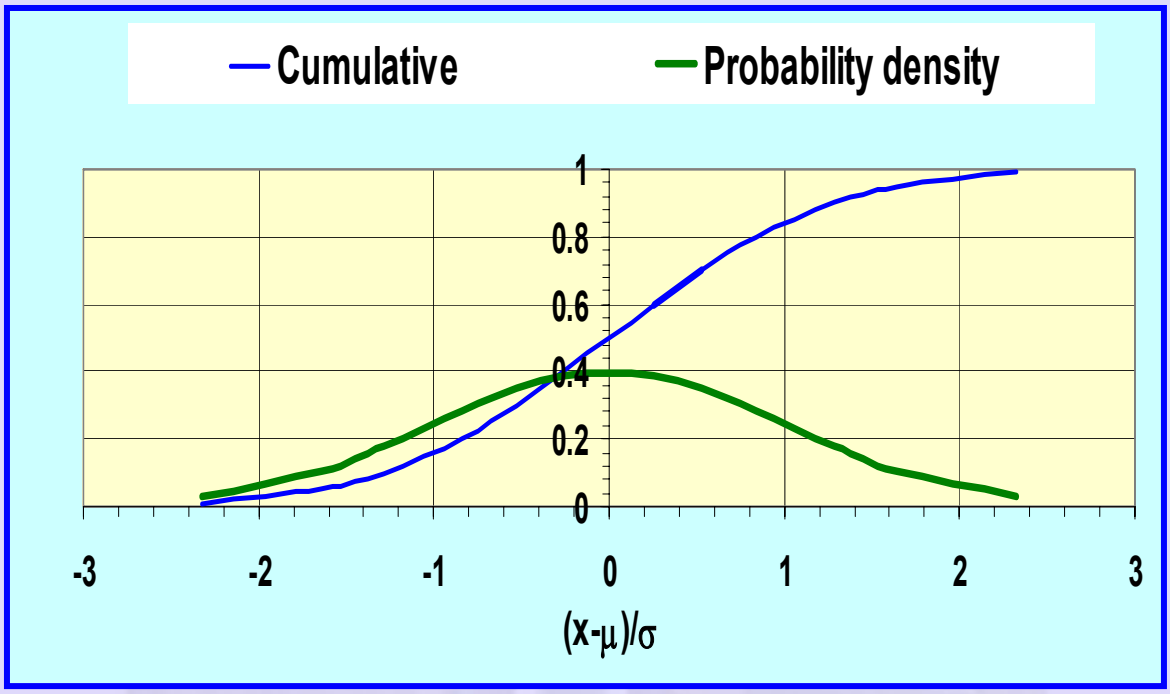

Figure 5 Normal distribution and its integral

Many times we are interested in finding out the chances of error lying between two values in the form ±pσ. This is referred to as the "confidence interval" and the corresponding cumulative probability specifies the chances of the error occurring within the confidence interval. Table 1 gives the confidence intervals that are useful in practice:

Table 1

Confidence intervals according to normal distribution

Cumulative Probability	0	0.95	0.99	0.999
Interval p	0	$\pm$1.96	$\pm$2.58	$\pm$3.29

The table indicates that error of magnitude greater than $\pm 3.29\sigma$ is very unlikely to occur. In most applications we specify $\pm 1.96\sigma$ as the error bounds based on **95%** confidence.

Example 1

- Resistance of a certain resistor is measured repeatedly to obtain the following data.

No.	1	2	3	4	5	6	7	8	9
R, kΩ	1.22	1.23	1.26	1.21	1.22	1.22	1.22	1.24	1.19

- What is the best estimate for the resistance? What is the error with 95% confidence?
- Best estimate is the mean of the data.

$$\bar{R} = \frac{1.22\times 4 + 1.23 + 1.26 + 1.21 + 1.24 + 1.19}{9}$$
$$= 1.223 \approx 1.22 \text{ k}\Omega$$

- Standard deviation of the error σ:

$$\text{Variance} = \frac{1}{9}\sum_{1}^{9}\left[\text{Ri} - \bar{R}\right]^2$$
$$= 3.33\times 10^{-4}$$

Hence:

$$\sigma = \sqrt{3.33\times 10^{-4}}$$
$$= 0.183 \approx 0.02 \text{ k}\Omega$$

- Error with 95% confidence :

$$\text{Error}_{95\%} = 1.96\sigma = 1.96\times 0.0183$$
$$= 0.036 \approx 0.04 \text{ k}\Omega$$

Example 2

Thickness of a metal sheet (in mm) is measured repeatedly to obtain the following replicate data. What is the best estimate for the sheet thickness? What is the variance of the distribution of errors with respect to the best value? Specify an error estimate to the mean value based on 99% confidence.

Experiment No.	1	2	3	4	5	6
t, mm	0.202	0.198	0.197	0.215	0.199	0.194
Experiment No.	7	8	9	10	11	12
t, mm	0.204	0.198	0.194	0.195	0.201	0.202

- The best estimate for the metal sheet thickness is the mean of the 12 measured values. This is given by

$$t_b = \bar{t} = \frac{\sum_1^{12} t_i}{12} = \frac{\left[\begin{array}{l} 0.202+0.198+0.197+0.215+0.199+0.194+0.204 \\ +0.198+0.194+0.195+0.201+0.202 \end{array}\right]}{12} = 0.2 \text{ mm}$$

- The variance with respect to the mean or the best value is given by (on substituting $\bar{t}$ for t_b) as

$$\sigma^2{}_b = \frac{\sum_1^{12}\left[t_i - \bar{t}\right]^2}{12} = \frac{\sum_1^{12} t_i}{12} - \bar{t}^2$$

- $$= \frac{\left[\begin{array}{l} 0.202^2+0.198^2+0.197^2+0.215^2+0.199^2+0.194^2+0.204^2 \\ +0.198^2+0.194^2+0.195^2+0.201^2+0.202^2 \end{array}\right]}{12} - 0.2 \text{ mm}$$

$$= 3.04\times10^{-5}\text{mm}^2$$

- The corresponding standard deviation is given by

$$\sigma_b = \sqrt{3.04\times10^{-5}} = 0.0055 \approx 0.006 \text{ mm}$$

⊙ The corresponding error estimate based on 99% confidence is

$$\text{Error} = \pm 2.58\sigma_b = \pm 2.58\times 0.0055 \approx \pm 0.014 \text{ mm}$$

Principle of Least Squares

Earlier we have dealt with the method of obtaining the best estimate from replicate data based on **minimization of variance**. No mathematical proof was given as a basis for this. We shall now look at the above afresh, in the light of the error distribution that has been presented above.

Consider a set of replicate data x_i. Let the best estimate for the measured quantity be x_b. The probability for a certain value x_i within the interval $x_i, x_i + dx_i$ to occur in the measured data is given by the relation

$$p(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_b - x_i)^2}{2\sigma^2}} dx_i \qquad (6)$$

The probability that the particular values of measured data are obtained in replicate measurements must be given by the compound probability given by

$$p = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \prod_{i=1}^{n} e^{-\frac{(x_b - x_i)^2}{2\sigma^2}} dx_i = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\sum_{i=1}^{n}\frac{(x_b - x_i)^2}{2\sigma^2}} \prod_{i=1}^{n} dx_i \qquad (7)$$

The reason the set of data was obtained as replicate data is that it was the **most probable**! Since the intervals dx_i are arbitrary, the above will have to be maximized by the proper choice of x_b and σ such that the exponential factor is a maximum. Thus we have to choose x_b and σ such that

$$p' = \frac{1}{\sigma^n} e^{-\sum_{i=1}^{n} \frac{(x_b - x_i)^2}{2\sigma^2}} \qquad (8)$$

has the largest possible value. As usual we set the derivatives $\frac{\partial p'}{\partial x_b} = \frac{\partial p'}{\partial \sigma} = 0$ to get the values of the two parameters x_b and σ. We have:

$$\frac{\partial p'}{\partial x_b} = -\frac{1}{2\sigma^{n+2}} e^{\sum_{i=1}^{n} \frac{(x_i - x_b)^2}{2\sigma^2}} \underbrace{\sum_{i=1}^{n} 2(x_i - x_b)(-1)}_{\text{This part should go to zero}} = 0 \qquad (9)$$

Or

$$\sum_{i=1}^{n} (x_i - x_b) = 0 \text{ or } x_b = \sum_{i=n}^{n} x_i = \bar{x} \qquad (10)$$

It is clear thus that the best value is nothing but the mean of the values! We also have:

$$\frac{\partial p'}{\partial \sigma} = \underbrace{\left[-\frac{n}{\sigma^{n+1}} + \frac{1}{\sigma^{n+3}} \sum_{i=0}^{n} (x_i - x_b)^2 \right]}_{\text{This part should go to Zero}} e^{\sum_{i=1}^{n} \frac{(x_i - x_b)^2}{2\sigma^2}} = 0 \qquad (11)$$

Or

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - x_b)^2}{n} \qquad (12)$$

This last expression indicates that the parameter σ^2 is nothing but the variance of the data with respect to the mean! Thus the best values of the measured quantity and its spread is based on the minimization of the squares of errors with respect to the mean. This embodies what is referred to as the **"Principle of Least Squares"**.

Propagation of errors:

Replicate data collected by measuring a single quantity repeatedly enables us to calculate the best value and characterize the spread by the variance with respect to the best value, using the principle of least squares. Now we look at the case of a **derived quantity** that is estimated from the measurement of several **primary** quantities. The question that needs to be answered is the following:

"A derived quantity Q is estimated using a formula that involves the primary quantities. $a_1, a_2, ...a_n$ Each one of these is available in terms of the respective best values $\bar{a}_1, \bar{a}_2, ...\bar{a}_n$ and the respective standard deviations $\sigma_1, \sigma_2 ....\sigma_n$. What is the best estimate for Q and what is the corresponding standard deviation σ_Q ?"

We have, by definition

$$Q = Q(a_1, a_2, ...a_n) \qquad (13)$$

It is obvious that the best value of Q should correspond to that obtained by using the **best values** for the a's. Thus, the best estimate for Q given by $\bar{Q}$ as

$$\bar{Q} = Q(\bar{a}_1, \bar{a}_2, ...\bar{a}_n) \qquad (14)$$

Again, by definition, we should have:

$$\sigma_Q^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Q_i - \bar{Q}\right)^2 \qquad (15)$$

The subscript i indicates the experiment number and the i^{th} estimate of Q is given by

$$Q_i = Q\left(a_{1i}, a_{2i}, ...a_{ni}\right) \qquad (16)$$

If we assume that the spread in values are small compared to the mean or the best values (this is what one would expect from a well conducted experiment), the difference between the i^{th} estimate and the best value may be written using a Taylor expansion around the best value as

$$\sigma_Q^2 = \frac{1}{N}\sum_{i=2}^{N}\left(\frac{\partial Q}{\partial a_1}\Delta a_{1i} + \frac{\partial Q}{\partial a_2}\Delta a_{2i} + ... + \frac{\partial Q}{\partial a_2}\Delta a_{ni}\right)^2 \qquad \textbf{(17)}$$

where the partial derivatives are all evaluated at the best values for the a's. If the a's are all **independent** of one another then the errors in these are unrelated to one another and hence the cross terms. $\sum_{i=1}^{N}\Delta a_{mi}\Delta a_{ki} = 0 \text{ for } m \neq k$ Thus equation (17) may be rewritten as

$$\sigma_Q^2 = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\partial Q}{\partial a_1}\Delta a_{1i}\right)^2 + \left(\frac{\partial Q}{\partial a_2}\Delta a_{2i}\right)^2 + ... + \left(\frac{\partial Q}{\partial a_n}\Delta a_{ni}\right)^2\right] \qquad \textbf{(18)}$$

Noting that $\sum_{i=1}^{N}\left(\Delta a_{ji}\right)^2 = N\sigma_j^2$ we may recast the above equation in the form

$$\sigma_Q^2 = \left(\frac{\partial Q}{\partial a_1}\right)^2\sigma_1^2 + \left(\frac{\partial Q}{\partial a_2}\right)^2\sigma_2^2 + ... + \left(\frac{\partial Q}{\partial a_n}\right)^2\sigma_n^2 \qquad \textbf{(19)}$$

Equation (19) is the error propagation formula. It may also be recast in the form

$$\sigma_Q = \sqrt{\left(\frac{\partial Q}{\partial a_1}\right)^2\sigma_1^2 + \left(\frac{\partial Q}{\partial a_2}\right)^2\sigma_2^2 + ... + \left(\frac{\partial Q}{\partial a_n}\right)^2\sigma_n^2} \qquad \textbf{(20)}$$

Example 3

The volume of a sphere is estimated by measuring its diameter by vernier calipers. In a certain case the diameter has been measured as D = 0.0502 ± 0.00005 m. Determine the volume and specify a suitable uncertainty for the same.

Nominal volume of sphere:

$$V = \pi \frac{D^3}{6} = 3.14159 \times \frac{0.0502^3}{6} = 6.624 \times 10^{-5} \text{m}^3$$

- The error in the measured diameter is specified as:

$$\Delta D = \pm 0.00005 \text{m}$$

- The influence coefficient is defined as

$$I_D = \frac{\partial V}{\partial D} = \pi \frac{D^2}{2} = 3.14159 \times \frac{0.0502^2}{2} = 3.958 \times 10^{-3} \text{ m}^2$$

- Using the error propagation formula, we have

$$\Delta V = I_D \Delta D = 3.958 \times 10^{-3} \times 0.00005 = 1.979 \times 10^{-7} \text{ m}^3$$

- Thus

$$V = 6.624 \times 10^{-5} \pm 1.979 \times 10^{-7} \text{ m}^3$$

Alternate solution to the problem

- By logarithmic differentiation we have

$$\frac{dV}{V} = 3 \frac{dD}{D}$$

- This may be recast as

$$\Delta V = \pm 3V\frac{\Delta D}{D} = \pm 3\times 6.624\times 10^{-5}\times\frac{0.00005}{0.0502} = \pm 0.0198\times 10^{-5}\text{m}^3$$

This is the same as the result obtained earlier.

Example 4

Two resistances R_1 and R_2 are given as 1000 ± 25 Ω and 500 ± 10 Ω . Determine the equivalent resistance when these two are connected in a) series and b) parallel. Also determine the uncertainties in these two cases.

- Given Data:

$$R_1 = 1000, \sigma_1 = 25; R_2 = 500\ \sigma_2 = 10 \rightarrow \text{All Values are in } \Omega$$

Case a) Resistances connected in series:

- Equivalent resistance is

$$R_s = R_1 + R_2 = 1000+500=1500\Omega$$

- Influence coefficients are:

$$I_1 = \frac{\partial R_s}{\partial R_1} = 1\ ; I_2 = \frac{\partial R_s}{\partial R_2} = 1$$

- Hence the uncertainty in the equivalent resistance is

$$\sigma_s = \pm\sqrt{(I_1\sigma_1)^2 + (I_2\sigma_2)^2} = \pm\sqrt{(25)^2 + (10)^2} = \pm 26.93\ \Omega$$

Case b) Resistances connected in parallel:

- Equivalent resistance is given by

$$R_p = \frac{R_1 R_2}{R_1 + R_2} = \frac{1000 \times 500}{1000 + 500} = 333.3\ \Omega$$

$$R_p = \frac{R_1 R_2}{(R_1 + R_2)}$$

- Influence coefficients are:

$$I_1 = \frac{\partial R_p}{\partial R_1} = \frac{R_2}{(R_1 + R_2)} - \frac{R_1 R_2}{(R_1 + R_2)^2} = \frac{500}{1500} - \frac{1000 \times 500}{1500} = 0.111$$

$$I_1 = \frac{\partial R_p}{\partial R_2} = \frac{R_1}{(R_1 + R_2)} - \frac{R_1 R_2}{(R_1 + R_2)^2} = \frac{1000}{1500} - \frac{1000 \times 500}{1500^2} = 0.444$$

- Hence the uncertainty in the equivalent resistance is

$$\sigma_s = \pm\sqrt{(I_1\sigma_1)^2 + (I_2\sigma_2)^2} = \pm\sqrt{(0.111 \times 25)^2 + (0.444 \times 10)^2} = \pm 5.24\ \Omega$$

Thus the equivalent resistance is 1500 $\pm$ 26.9 Ω in the series arrangement and 333.6 $\pm$ 5.24 Ω in the parallel arrangement.