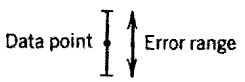
GRAPHICAL ANALYSIS

A purpose of many experiments is to find the relationship between measured variables. A good way to accomplish this task is to plot a graph of the data and then analyze the graph. These guidelines should be followed in plotting your data:

1. Use a sharp pencil or pen. A broad-tipped pencil or pen will introduce unnecessary inaccuracies.
2. Draw your graph on a full page of graph paper. A compressed graph will reduce the accuracy of your graphical analysis.
3. Give the graph a concise title.
4. The dependent variable should be plotted along the vertical (y) axis and the independent variable should be plotted along the horizontal (x) axis.
5. Label axes and include units.
6. Select a scale for each axis and start each axis at zero, if possible.
7. Use error bars to indicate errors in measurements, for example,



8. Draw a smooth curve through the data points. If the errors are random, then about one-third of the points will not lie within their error range of the best curve.

The microcomputer is a powerful tool for data analysis. Commercial software is available that handles data and instructs the microcomputer to carry out graphical analysis. See your instructor about the availability of this software for your laboratory.

As an example consider the study of the speed of an object (dependent variable) as a function of time (independent variable). The data are as follows:

Speed (m/s)	Time (s)
0.45 ± 0.06	1
0.81 ± 0.06	2
0.91 ± 0.06	3
1.01 ± 0.06	4
1.36 ± 0.06	5
1.56 ± 0.06	6
1.65 ± 0.06	7
1.85 ± 0.06	8
2.17 ± 0.06	9

Using the above guidelines, the data are graphed in Figure I.7.

The graphed data show that the speed v is a linear function of the time t. The general equation for a straight line is

$$y = mx + b \tag{22}$$

where m is the slope of the line and b, the vertical intercept, is the value of y when $x = 0$. Let $v = y$,

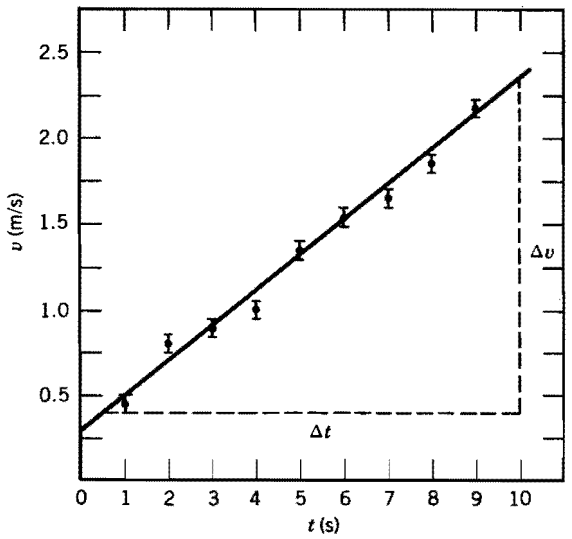


FIGURE I.7 Speed versus time. The graphed data, v versus t, show a linear relation.

$x = t$, $a = m$, and $v_0 = b$; then,

$$v = at + v_0 \qquad \text{(m/s)} \qquad (23)$$

This is the form of the equation for the line drawn through the data, where v_0 is the value of the velocity at $t = 0$ and a is the slope of the line that is the acceleration of the object. From the graph we see that $v_0 = 0.32$ m/s. To determine the slope select two points on the line, but not data points, which are well separated, then

$$a = \text{slope} = \frac{\Delta v}{\Delta t} = \frac{2.35 - 0.40\ \text{(m/s)}}{10.0 - 0.5\ \text{(s)}}$$

$$= \frac{1.95\ \text{(m/s)}}{9.5\ \text{(s)}} = 0.20\ \text{m/s}^2 \qquad (24)$$

The equation for the line is

$$v = 0.20t + 0.32 \qquad \text{(m/s)} \qquad (25)$$

The data plotted in Figure I.7 are analyzed in the section on "Curve Fitting," page 23, as an example of linear regression.

As a second example, let us consider the study of the distance traveled by an object as a function of time. The data are as follows:

Distance (m)	Time (s)
0.20 ± 0.05	1
0.43 ± 0.05	2
0.81 ± 0.05	3
1.57 ± 0.10	4
2.43 ± 0.10	5
3.81 ± 0.10	6
4.80 ± 0.20	7
6.39 ± 0.20	8

The data are graphed, using the above guidelines, in Figure I.8.

In this instance a straight line through the data points would not be acceptable. An inspection of the graph suggests that d is proportional to t^n, where $n > 1$; for example, d may be a quadratic function of time and, hence, $n = 2$.

Suppose that we know the theoretical relation between d and t is

$$d = \tfrac{1}{2}at^2 \qquad \text{(m)} \qquad (26)$$

where a is the object's acceleration. Often it is useful to know if the data agree with the theory. If the data follow the above theoretical relation, then a graph of d versus t^2 should result in a straight line.

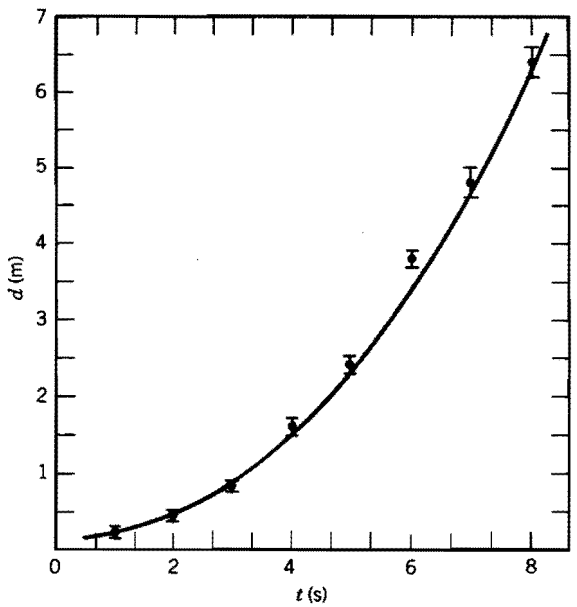


FIGURE I.8 Distance versus time. The graphed data, d versus t, show a nonlinear relation.

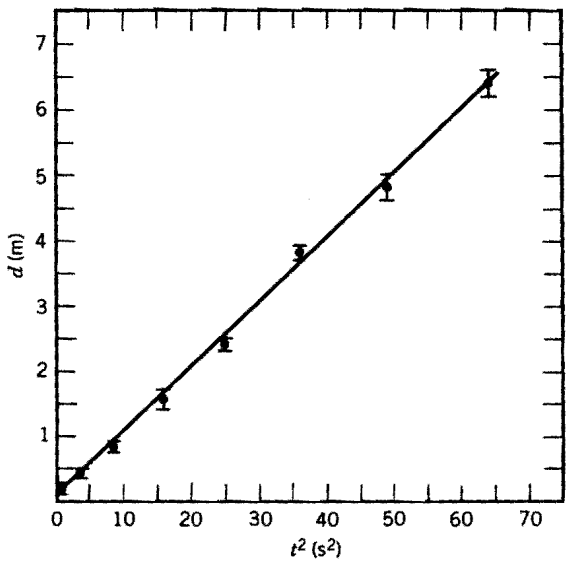


FIGURE I.9 Plotting d versus t^2 yields a linear relation.

The graph in Figure I.9 indicates that d is a linear function of t^2 and, hence, that the data agree with the theoretical relation. The equation for the straight line is

$$d = mt^2 + d_0 \qquad \text{(m)} \qquad (27)$$

where m is the slope and d_0 is the vertical intercept.

PLOTTING DATA ON SEMILOG PAPER

Often the relationship between the measured variables is not linear. For example, consider the intensity of light I transmitted through a sample of thickness x, shown in Figure I.10, where I_0 is the incident intensity of the light.

Lambert's law states the theoretical relationship between the dependent variable I and the independent variable x:

$$I = I_0 e^{-\mu x} \qquad (\text{W/cm}^2) \qquad (28)$$

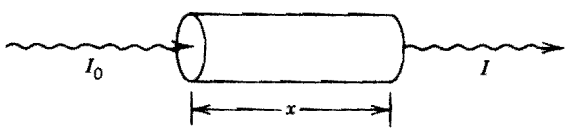


FIGURE I.10 I_0 is the incident light intensity, x is the sample thickness, and I is the transmitted intensity.

where μ is the absorption coefficient, a constant that depends on the wavelength of light and the absorbing properties of the sample. Suppose I is measured as a function of x, and the data are plotted as is shown in Figure I.11.

From the smooth curve it would be difficult to determine the relationship between I and x, that is, it would be difficult to conclude the data obey Lambert's law.

A good way to determine the experimental relationship between I and x is to use semilog paper. Semilog paper has a logarithmic y axis (it automatically takes logarithms of data plotted) and a regularly spaced x axis. The data are plotted on semilog paper in Figure I.12. Note that there is never a zero on the logarithmic axis, and that when reading values off of a logarithmic axis you read the logarithm of the value and not the value, for example, log 9 and not 9.

The smooth curve drawn through the data is a straight line with a negative slope and the intensity at the point on the vertical axis intercepted by the curve is I_0. Lambert's law does agree with this result as can be seen by taking the logarithm of Lambert's law:

$$\begin{aligned} \log I &= \log(I_0 e^{-\mu x}) \\ &= \log e^{-\mu x} + \log I_0 \\ &= -\mu x \log e + \log I_0 \\ &= -0.434\mu x + \log I_0 \qquad \text{(unitless)} \end{aligned} \qquad (29)$$

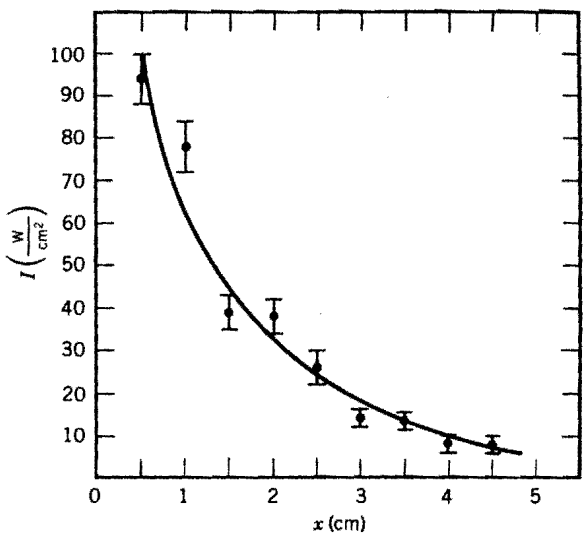


FIGURE I.11 Light intensity versus sample thickness, showing a nonlinear relation. From the graph it is not clear if the data obey Lambert's law or not.

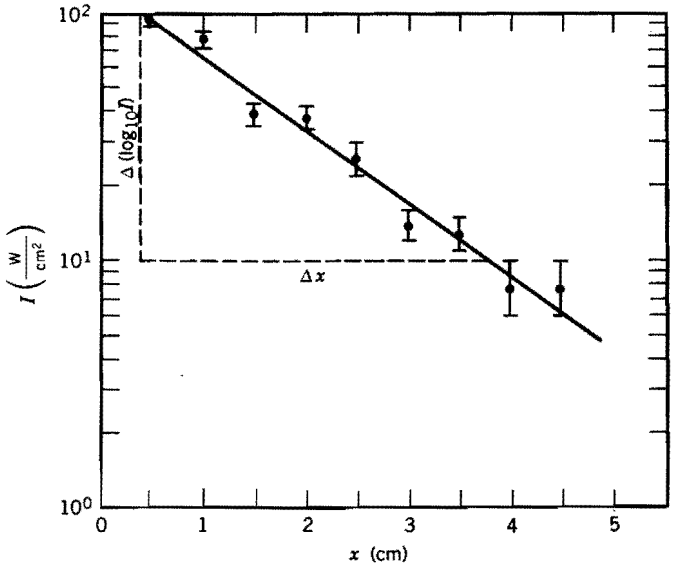


FIGURE 1.12 Light intensity versus sample thickness. The linear relation obtained on semilog paper shows that the data obey Lambert's law.

Again, the general equation of a straight line is of the form:

$$y = mx + b \tag{30}$$

Now let $y = \log I$, $m = -0.434\mu$, and $b = \log I_0$. Then, if $\log I$ is plotted vertically and x is plotted horizontally, the curve will be a straight line with slope -0.434μ and vertical intercept $\log I_0$. Using semilog paper, I is plotted on the logarithmic axis; the vertical intercept on this axis is I_0. Note that the slope of the line drawn through the data points may be used to calculate μ:

$$\text{slope} = \frac{\Delta(\log I)}{\Delta x} = \frac{\log 10 - \log 100}{(3.80 - 0.40)\text{ cm}} = -0.294\text{ cm}^{-1} \tag{31}$$

From Lambert's law the theoretical slope is

$$\text{slope} = -0.434\mu$$

By equating theoretical and experimental slopes, we find that

$$-0.434\mu = -0.294\text{ cm}^{-1}$$

and

$$\mu = +0.678\text{ cm}^{-1}$$

EXERCISE 2

Suppose the functional relation between the dependent variable y and the independent variable x is given by

$$y = a e^{-x} + b$$

where a and b are nonzero constants. Explain why a graph of y versus x on semilog paper would not give a straight line.

PLOTTING DATA ON LOG–LOG PAPER

Log–log paper is used to obtain a straight line plot when y and x satisfy a power-law relation:

$$y = cx^n \tag{32}$$

where c and n are constants. For example, the semimajor axis R of the orbit of a planet is related to its period (time for one revolution around the sun) T:

$$R^3 = KT^2 \quad \text{or} \quad R = K^{1/3}T^{2/3} \tag{33}$$

where K is a constant. R is nonlinearly related to T.

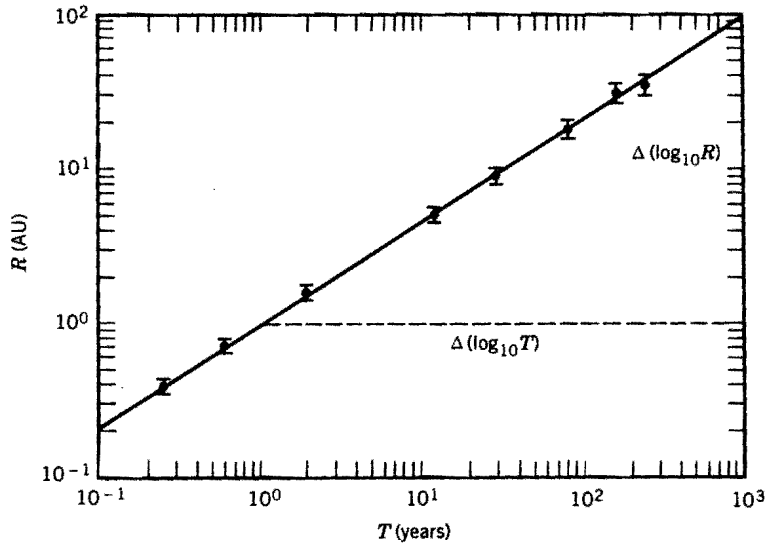


FIGURE I.13 Planets: Semimajor axis versus period. The linear relation on log–log paper indicates R and T obey a power law of the form of equation 32.

A straight-line plot is obtained in the following way. Take logarithms

$$\begin{aligned} \log R &= \log(K^{1/3}T^{2/3}) \\ &= \log T^{2/3} + \log K^{1/3} \\ &= 2/3 \log T + \log K^{1/3} \end{aligned} \tag{34}$$

Let $y = \log R$, $x = \log T$, $m = \frac{2}{3}$, and $b = \log K^{1/3}$. Then a plot of $\log R$ versus $\log T$ would be a straight line. Log–log graph paper automatically takes the logarithm of the plotted data. A log–log graph is shown in Figure I.13.

The units used are years and astronomical units (AU), where 1 AU is the semimajor axis of earth's orbit. (The errors shown in the graph are fictitious.) The slope of the log–log plot is

$$\begin{aligned} \text{slope} &= \frac{\Delta(\log R)}{\Delta(\log T)} = \frac{\log 10^2 - \log 10^0}{\log 10^3 - \log 10^0} \\ &= \frac{2-0}{3-0} = \frac{2}{3} \end{aligned} \tag{35}$$

Note that the slope of the log–log plot is the exponent of the power law relation. For example, the power law relation $y = cx^n$ plotted on log–log paper has a slope equal to n. Hence, a log–log plot is a good way to determine the exponent in a power law relation.

Another way to obtain a straight-line plot is to plot y versus x^n or R versus $T^{2/3}$ on regular graph paper (see Figure I.14).

A problem with plotting R versus $T^{2/3}$ is that values of R less than about 1 AU cannot be plotted with much accuracy.

In units of years and astronomical units the constant K is one, and an inspection of the curve in the figure shows a slope of approximately one.

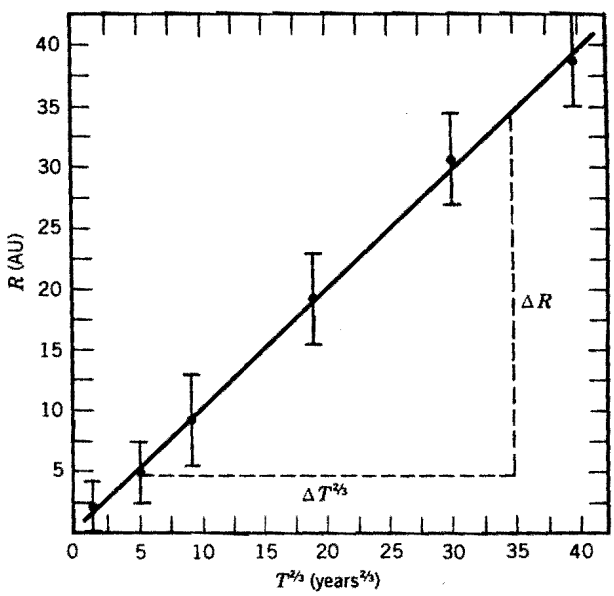


FIGURE I.14 Planets: R versus $T^{2/3}$, showing a linear relation. This graph requires knowing the exponent in the power-law relation.

CURVE FITTING

Just as it is important to learn to use a microcomputer to do graphical analysis, it is also important to use the microcomputer to carry out curve fitting. Commercial software is available for curve fitting. See your instructor about the availability of this software for your use.

Given n data points (x_i, y_i), we want to find the equation for the "best" curve for this set of data. If the data points are linearly related, then the process is called **linear regression**. In general, data points are not linearly related and the process of obtaining the equation for the best curve is called **nonlinear regression**. The technique to be used in determining the best-fitting curve is the **method of least squares**.

Before we consider linear and nonlinear regression, we use the method of least squares to determine the best estimate of a quantity x.

Suppose a physical quantity is measured n times, x_i, $i = 1, 2, \ldots, n$. An example is the measurement of a single period of a pendulum n times, where for each measurement the length, mass, and amplitude are constant. The method of least squares states that the best estimate for the result of the n measurements is that which minimizes the sum of the squares of the deviations of the measurements from their best estimate x, that is, we minimize

$$\sum_{i=1}^{n} (x - x_i)^2 \tag{36}$$

where x is the unknown best estimate. Minimizing expression 36 and solving for x, we find that

$$\begin{aligned} \frac{d}{dx} \sum_{i=1}^{n} (x - x_i)^2 &= 0 \\ 2nx - 2 \sum_{i=1}^{n} x_i &= 0 \\ x &= \frac{1}{n} \sum_{i=1}^{n} x_i \equiv \bar{x} \end{aligned} \tag{37}$$

Hence, the best estimate is the average or mean value, $\bar{x}$.

Note that minimizing the sum of the squared deviations is equivalent to maximizing the probability $P(x_1, x_2, \ldots, x_n)$ of obtaining our set of measurements $x_1, x_2, \ldots, x_n$. We assume that the data points (x_i) are distributed according to the Gauss distribution; then the probability of obtaining a measurement within an interval dx of x_i is

$$P(x_i) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{(x - x_i)^2}{2\sigma^2} \right] dx \tag{38}$$

where

$$x = \text{best estimate for } x_i \tag{39}$$

and σ is the theoretical standard deviation.

The probability of obtaining our set of measurements is

$$\begin{aligned} P(x_1, x_2, \ldots, x_n) &= P(x_1)P(x_2) \cdots P(x_n) \\ &= \left(\frac{dx}{\sigma\sqrt{2\pi}} \right)^n \\ &\quad \times \exp\left[-\sum_{i=1}^{n} \frac{(x - x_i)^2}{2\sigma^2} \right] \end{aligned} \tag{40}$$

If we minimize the exponent in equation 40, then $P(x_1, \ldots, x_n)$ will be a maximum. The sum in the exponent is called the **least-squares sum,**

$$\sum_{i=1}^{n} \frac{(x - x_i)^2}{2\sigma} \tag{41}$$

and minimizing it is equivalent to minimizing $\Sigma (x - x_i)^2$, since σ is (assumed) a constant.

Note: We assume the data points follow the Gauss distribution, and the method of least squares is used to find the most probable value.

METHOD OF LEAST SQUARES AND LINEAR REGRESSION

Given n data points (x_i, y_i) (for example, x_i could be the time and y_i the average speed of a falling object), we would like to find the equation for the best straight line. Typical data points (x_i, y_i) and the equation of the line, which we want to determine, are shown in Figure I.15. We make the following assumptions:

1. The measured values (x_i, y_i) are distributed according to the Gauss distribution (this is usually so if the errors are random).
2. The errors in x_i, δx_i, are negligible in comparison to the errors in y_i, δy_i (then we only consider the distribution of the values y_i).
3. The errors in y are all the same: $\delta y_1 = \delta y_2 = \cdots = \delta y_n$ (then the standard deviation σ_y is constant).

We approximate the set of n measurements (x_i, y_i) by a linear relation:

$$y(x) = a_0 + a_1 x \tag{42}$$

The probability of obtaining the observed value y_i is

$$P(y_i) \propto \frac{1}{\sigma_y} \exp\left[-\frac{[y_i - y(x_i)]^2}{2\sigma_y^2} \right] \tag{43}$$

where

$$y(x_i) = \text{best estimate for } y_i = a_0 + a_1 x_i \tag{44}$$

and σ_y is the theoretical standard deviation. The probability $P(y_1, \ldots, y_n)$ of obtaining the set of measurements is

$$P(y_1, \ldots, y_n) = P(y_1)P(y_2) \cdots P(y_n)$$

$$\propto \frac{1}{(\sigma_y)^n} \exp\left[-\sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i)^2}{2\sigma_y^2} \right] \tag{45}$$

We want this probability to be a maximum; hence, the exponent (least-squares sum) must be a minimum. Minimizing the least-squares sum gives the equation for the best straight line.

In Figure I.15, d_i is the vertical distance from each point (x_i, y_i) to the line $y = a_0 + a_1 x$. We wish to find values of a_0 and a_1 such that we minimize the function $M(a_0, a_1)$ defined to be

$$M(a_0, a_1) = \sum_{i=1}^{n} \frac{d_i^2}{2\sigma_y^2} = \sum_{i=1}^{n} \frac{[y_i - (a_0 + a_1 x_i)]^2}{2\sigma_y^2} \tag{46}$$

which is the exponent in equation 45. Expanding the squared term and ignoring the (assumed) constant σ_y, we find that

$$M = \Sigma (y_i)^2 - 2a_1 \Sigma x_i y_i - 2a_0 \Sigma y_i + a_1^2 \Sigma x_i^2 + 2a_0 a_1 \Sigma x_i + n a_0^2 \tag{47}$$

where Σ is understood as a sum over the index i. Next we set

$$\frac{dM}{da_0} = 0 \quad \text{and} \quad \frac{dM}{da_1} = 0 \tag{48}$$

to find a_0 and a_1 corresponding to the minimum

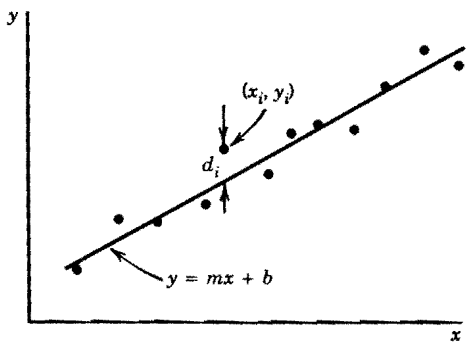


FIGURE I.15 Minimizing the least-squares sum gives the equation for the best straight line.

value of M. This results in two simultaneous equations:

$$\frac{dM}{da_0} = -2\Sigma y_i + 2a_1 \Sigma x_i + 2na_0 = 0$$

$$\frac{dM}{da_1} = -2\Sigma x_i y_i + 2a_1 \Sigma x_i^2 + 2a_0 \Sigma x_i = 0 \quad (49)$$

which when solved for a_0 (intercept) and a_1 (slope) yield

$$a_0 = \frac{(\Sigma x_i^2)\Sigma y_i - (\Sigma x_i)(\Sigma x_i y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad (50)$$

$$a_1 = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2} \quad (51)$$

The equation for the best-fitting line is obtained by substituting equations 50 and 51 into equation 42.

We ask this question: "What are the uncertainties in a_0 and a_1?" Each y_i has an uncertainty (assumed the same for all y_i) and, hence, a_0 and a_1 will both have uncertainties. These uncertainties are the standard deviations of the means, s_{ma_0} and s_{ma_1}. To calculate s_{ma_0} and s_{ma_1}, we need the standard deviation s_y.

We ask the question: "What is the statistical uncertainty in the measurements $y_1, y_2, \ldots, y_n$?" In this case the standard deviation s_y is

$$s_y = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i)^2} \quad (52)$$

The standard deviation of the mean s_{my} is

$$s_{my} = \frac{s_y}{n^{1/2}} \quad (53)$$

For each y_i the result to be reported is

$$y_i \pm s_{my} \qquad i = 1, 2, \ldots, n \quad (54)$$

The reason for the factor of $n-2$ in the denominator of equation 52 is that the calculation of a_0 and a_1 reduces the number of independent data points (x_i, y_i) from n to $n-2$; the denominator in the equation for the standard deviation is the number of independent data points.

Remark: It is important to check whether the estimated errors, δy_i, recorded during data taking are consistent with the calculated statistical error s_{my}. A standard deviation of the mean s_{my}, which is much larger than the estimated errors, δy_i, would indicate estimated errors that are unaccounted for. Experimental errors, δy_i, which are much larger than s_{my} suggest a too conservative error estimate, that is, the δy_i should have been estimated as smaller values.

EXERCISE 3

A physicist plans to calibrate her equipment by determining an average value for some parameter x. She does this by measuring four values of x and estimates the error δx. Suppose that the values of $x \pm \delta x$ are 2.741 ± 0.010, 2.832 ± 0.010, 2.678 ± 0.010, 2.763 ± 0.010. Calculate the mean, $\bar{x}$, and the standard deviation of the mean, s_m. Is her estimated error too large, too small, or reasonable? Explain.

We now consider the errors in a_0 and a_1, s_{ma_0} and s_{ma_1}. Equations 50 and 51 give a_0 and a_1 as functions of the measured values (x_i, y_i) where the statistical error for each y_i is given in equation 53. Since a_0 and a_1 are known functions of y_i and the errors in y_i are known, the errors in a_0 and a_1 may be determined by error propagation. The basic formula for error propagation, equation 12, may be written as

$$\delta Q = \sqrt{\sum_{j=1}^{n}\left(\frac{\partial Q}{\partial b_j}\right)^2 (\delta b_j)^2} \quad (55)$$

where the measured values are $b_j \pm \delta b_j$, $j = 1, 2, \ldots, n$, and δQ is the error in the calculated quantity $Q(b_1, b_2, \ldots, b_n)$. Replacing δQ and δb_j with standard deviations of the mean s_{mQ} and s_{mb_j} and squaring, we have

$$s_{mQ}^2 = \sum_{j=1}^{n}\left(\frac{\partial Q}{\partial b_j}\right)^2 s_{mb_j}^2 \quad (56)$$

Applying equation 56, s_{ma_0} is

$$s_{ma_0}^2 = \sum_{j=1}^{n}\left(\frac{\partial a_0}{\partial y_j}\right)^2 s_{my}^2 \quad (57)$$

where the partial derivative $\partial a_0/\partial y_i$ is calculated by using equation 50:

$$\frac{\partial a_0}{\partial y_j} = \frac{\Sigma_i x_i^2 - \left(\Sigma_i x_i\right)x_j}{n\Sigma_i x_i^2 - \left(\Sigma_i x_i\right)^2} \quad (58)$$

Then, after some algebra, equation 57 becomes

$$s_{ma_0}^2 = \frac{s_{my}^2 \Sigma x_i^2}{n \Sigma x_i^2 - (\Sigma x_i)^2} \tag{59}$$

The result to be reported is

$$a_0 \pm s_{ma_0} \tag{60}$$

The calculation of $s_{ma_1}^2$ is similar to the calculation of $s_{ma_0}^2$. The result is

$$s_{ma_1}^2 = \frac{n s_{my}^2}{n \Sigma x_i^2 - (\Sigma x_i)^2} \tag{61}$$

and we report

$$a_1 \pm s_{ma_1} \tag{62}$$

Example

The method of least squares and linear regression is applied to the speed versus time data given in the section on "Graphical Analysis," p. 18, and plotted in Figure I.7.

The vertical intercept a_0 is calculated using equation 50, where $n = 9$, and the result is

$$a_0 = 0.305 \text{ m/s}$$

The slope a_1 is obtained from equation 51:

$$a_1 = 0.201 \text{ m/s}^2$$

When a_0 and a_1 are known, equations 52 and 53 may be used to calculate s_{my}:

$$s_{my} = 0.025 \text{ m/s}$$

where, in this case, the dependent variable y is the speed v. For each v_i the result to be reported is

$$v_i \pm s_{mv} = v_i \pm 0.025$$

Note that s_{mv} is smaller than the estimated errors $\delta v_i = 0.06$ m/s (see data on p. 18), which suggests the estimated errors were too conservative or too large.

When s_{mv} is known, the uncertainties in a_0 (s_{ma_0}) and a_1 (s_{ma_1}) may be calculated by using equations 59 and 61. The results are

$$s_{ma_0} = 0.018 \text{ m/s}$$

$$s_{ma_1} = 0.003 \text{ m/s}^2$$

Thus,

$$a_0 \pm s_{ma_0} = 0.305 \pm 0.018 \text{ m/s}$$

$$a_1 \pm s_{ma_1} = 0.201 \pm 0.003 \text{ m/s}^2$$

METHOD OF LEAST SQUARES AND NONLINEAR REGRESSION

Given n data points (x_i, y_i), $i = 1, 2, \ldots, n$, that are nonlinearly related, we want to determine the polynomial in x that gives the best fit to the set of n measurements:

$$y(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m \tag{63}$$

If a plot of the data or theoretical considerations suggest a quadratic function of x, then we consider only the first three terms in equation 63. We make the same three assumptions as in the method of least squares and linear regression.

As before, equation 43, the probability of obtaining the observed value y_i is

$$P(y_i) \propto \frac{1}{\sigma_y} \exp\left\{-\frac{[y_i - y(x_i)]^2}{2\sigma_y^2}\right\} \tag{64}$$

where

$$y(x_i) = \text{the best estimate for } y_i = a_0 + a_1 x_i + \cdots + a_m x_i^m \tag{65}$$

and σ_y is the theoretical standard deviation. The probability of obtaining the set of measurements is

$$P(y_1, y_2, \ldots, y_n) = P(y_1)P(y_2) \cdots P(y_n) \propto \frac{1}{(\sigma_y)^n} \exp\left[-\sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2}{2\sigma_y^2}\right] \tag{66}$$

Again we want the probability to be a maximum; hence, the exponent (least-squares sum) must be a minimum and minimizing the least-squares sum gives the equation for the best fitting curve.

We wish to find the values of $a_0, a_1, \ldots, a_m$

such that we minimize the function M defined to be

$$M = \sum_{i=1}^{n} \frac{(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2}{2\sigma_y} \quad (67)$$

Taking the partial derivative of M with respect to a_k and setting it equal to zero yields

$$\frac{\partial M}{\partial a_k} = 2 \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m) x_i^k = 0 \quad (68)$$

where $k = 0, 1, 2, \ldots, m$. Equation 68 is a set of $m+1$ equations in the $m+1$ variables $a_0, a_1, \ldots, a_m$ which determines the best-fitting curve.

CHI-SQUARE TEST OF FIT

If a measurement is repeated many times then the distribution of measured values is expected to follow a theoretical distribution precisely in the limit that the number of measurements approaches infinity. The Gauss and Poisson distributions are two of many theoretical distributions used in physics, corresponding to different kinds of experiments. (The Poisson distribution is discussed in Experiment 6.)

Suppose we have repeated a measurement n times. We ask the question, "How do we determine whether the measurements follow the expected theoretical distribution?" The chi-square, or χ^2, test provides the answer to this question. χ^2 is a number, without units, defined by

$$\chi^2 \equiv \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k} \quad (69)$$

where m is the number of bins, O_k is the number of observed or measured values in the kth bin, and E_k is the number of expected values in the kth bin. The n measured values are divided into bins or ranges of values, where the bins must be chosen so that each bin contains several measured values. By assuming that the measurements follow an expected theoretical distribution, such as Gauss or Poisson distribution, we can calculate the expected number E_k of measurements in each bin k:

$$E_k = nP_k \quad (70)$$

where P_k is the probability that any measurement falls in bin k. Figure I.16 shows a Gauss distribution with 6 bins and probabilities $P_1 - P_6$, where $P_1 = P_6 = 0.02$, $P_2 = P_5 = 0.14$, and $P_3 = P_4 = 0.34$ for the Gauss distribution.

The interpretation of χ^2, calculated from equation 69, is as follows:

1. If $\chi^2 = 0$, then the measured values follow the theoretical distribution exactly.
2. If $\chi^2 \leq m - c$, then the agreement between the distribution of measured values and the theoretical distribution is good, where m is the number of bins and c is the number of parameters that had to be calculated from the data to

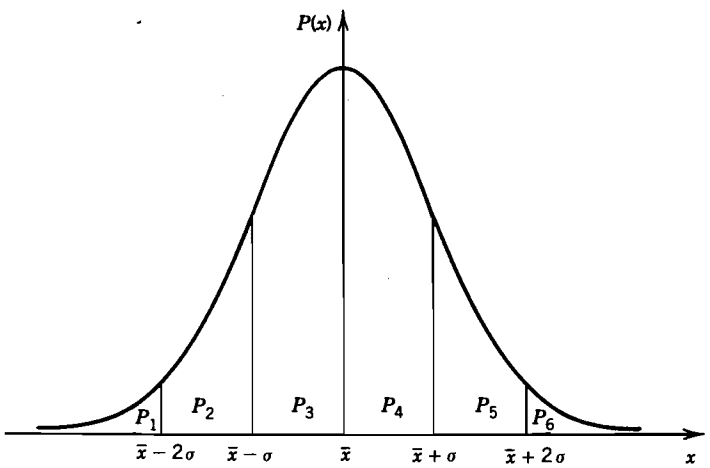


FIGURE I.16 A Gauss distribution with six bins and probabilities P_1 through P_6.

compute the expected number E_k. In statistical calculations $m - c$ is the number of degrees of freedom.

3. If $\chi^2 \gg m - c$, then the agreement is bad.

A more precise interpretation of χ^2 is obtained from a table of values of χ^2.

Example

A distance is measured 20 times. The measured values of x (in cm) are given in Table I.1. The mean value, calculated from equation 1, is $\bar{x} = 16.70$ cm. From equation 2 the standard deviation is $s = 0.16$ cm. To simplify the determination of P_k, we choose the bin boundaries at $\bar{x} - s$, $\bar{x}$, and $\bar{x} + s$, giving four bins as shown in Table I.2. The probability P_k is shown in Figure I.17

TABLE I.1 TWENTY MEASUREMENTS OF THE DISTANCE x

16.7	16.9	16.8	16.7	16.8	16.7	16.6
17.0	16.7	16.7	16.9	16.5	16.3	16.7
16.8	16.7	16.6	16.4	16.7	16.7	

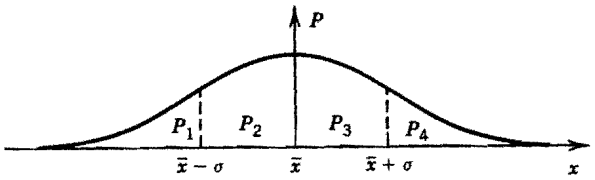


FIGURE I.17 A Gauss distribution with four bins and probabilities P_1 through P_4.

and the expected number E_k is calculated from equation 70 with $n = 20$. If a measured value falls on a bin boundary, then the observed number is determined by alloting 0.5 to each bin. χ^2 is calculated from equation 69, where $m = 4$. The result is $\chi^2 = 0.11$. To calculate E_k, two parameters, $\bar{x}$ and s, had to be determined from the data. In addition,

$$n = \sum_{k=1}^{4} O_k \tag{71}$$

is a constraint. Hence, $c = 3$ and $m - c = 1$. Since $\chi^2 < 1$, the agreement is good.

The probability obtained from a table of χ^2 values is that, on repeating the series of measurements, larger deviations from the expected values would be observed. In this example the probability, obtained from tables (see reference 1), is between 0.90 and 0.95 that a set of measurements with two degrees of freedom will have $\chi^2 > 0.11$. In other words, if the set of measurements was repeated 100 times then we would expect that 90 to 95 cases would yield values of χ^2 greater than 0.11.

In interpreting the value of P obtained from tables, we may say that if

$$0.1 < P < 0.9 \tag{72}$$

then the assumed distribution very probably corresponds to the observed one, while if

$$P < 0.02 \quad \text{or} \quad P > 0.98 \tag{73}$$

then the assumed distribution is very unlikely.

TABLE I.2 DIVIDING THE 20 MEASURED VALUES OF x INTO FOUR BINS FOR A χ^2 CALCULATION

Bin Number, k	1	2	3	4
Range of x in each bin	$x < \bar{x} - s$ or $x < 16.54$	$\bar{x} - s < x < \bar{x}$ or $16.54 < x < 16.70$	$\bar{x} < x < \bar{x} + s$ or $16.70 < x < 16.86$	$\bar{x} + s < x$ or $16.86 < x$
Probability P_k	0.16	0.34	0.34	0.16
Expected number $E_k = nP_k$	3.2	6.8	6.8	3.2
Observed number O_k	3	6.5	7.5	3