DATA REDUCTION AND ERROR ANALYSIS FOR THE PHYSICAL SCIENCES

THIRD EDITION

Philip R. Bevington

Late Associate Professor of Physics
Case Western Reserve University

D. Keith Robinson

Emeritus Professor of Physics
Case Western Reserve University

Mc
Graw
Hill

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York
San Francisco St. Louis Bangkok Bogotá Caracas Kuala Lumpur
Lisbon London Madrid Mexico City Milan Montreal New Delhi
Santiago Seoul Singapore Sydney Taipei Toronto

TABLE 6.2

Number of counts detected in 7½-min intervals as a function of distance from the source

i	Distance d_i (m)	$x_i = 1/d_i^2$ (m^{-2})	Counts C_i	σ_{C_i}	Weight $(1/C_i^2)$ w_i	$w_i x_i$	$w_i C_i$	$w_i x_i^2$	$w_i x_i C_i$	Fitted counts $a + bx_i$
1	0.20	25.00	901	30.0	0.00111	0.0278	1	0.694	25.0	887
2	0.25	16.00	652	25.5	0.00153	0.0254	1	0.393	16.0	610
3	0.30	11.11	443	21.0	0.00226	0.0251	1	0.279	11.1	461
4	0.35	8.16	339	18.4	0.00295	0.0241	1	0.197	8.2	370
5	0.40	6.25	283	16.8	0.00353	0.0221	1	0.138	6.3	311
6	0.45	4.94	281	16.8	0.00356	0.0176	1	0.087	4.9	271
7	0.50	4.00	240	15.5	0.00417	0.0167	1	0.067	4.0	242
8	0.60	2.78	220	14.8	0.00455	0.0126	1	0.035	2.8	205
9	0.75	1.78	180	13.4	0.00556	0.0099	1	0.018	1.8	174
10	1.00	1.00	154	12.4	0.00649	0.0065	1	0.007	1.0	150
Sums					0.03570	0.1868	10	1.912	81.0	

$$\sigma_i = \sqrt{y_i} \quad w_i = 1/\sigma_i^2 = 1/y_i$$

$$\Delta = \Sigma w_i \Sigma w_i x_i^2 - (\Sigma w_i x_i)^2 = 0.03570 \times 1.912 - (0.1868)^2 = 0.0334$$

$$a = [\Sigma w_i C_i \Sigma w_i x_i^2 - \Sigma w_i x_i \Sigma w_i x_i C_i]/\Delta = [10 \times 1.912 - 0.1868 \times 81.0]/\Delta = 119.5$$

$$b = [\Sigma w_i \Sigma w_i x_i C_i - \Sigma w_i x_i \Sigma w_i C_i]/\Delta = [0.03570 \times 81.0 - 0.1868 \times 10]/\Delta = 30.7$$

$$\sigma_a^2 \simeq \Sigma w_i x_i^2/\Delta = 1.912 / 0.0334 = 57.3 \qquad \sigma_a \simeq 7.6$$

$$\sigma_b^2 \simeq \Sigma w_i/\Delta = 0.03570 / 0.0334 = 1.07 \qquad \sigma_b \simeq 1.1$$

Note: A linear fit to the data of the function $C = a + bx$ by the method of determinants gives $a = 119 \pm 8$ and $b = 31 \pm 1$, with $\chi^2 = 11.1$ for 8 degrees of freedom. The χ^2 probability for the fit is about 20%.

We cannot fit a straight line to the data exactly in either example because it is impossible to draw a straight line through all the points. For a set of N arbitrary points, it is always possible to fit a polynomial of degree $N - 1$ exactly, but for our experiments, the coefficients of the higher-order terms would have questionable significance. We assume that the fluctuations of the individual points above and below the solid curves are caused by experimental uncertainties in the individual measurements. In Chapter 11 we shall develop a method for testing whether higher-order terms are significant.

Measuring Uncertainties

If we were to make a series of measurements of the dependent quantity y_i for one particular value x_i of the independent quantity, we would find that the measured values were distributed about a mean in the manner discussed in Chapter 5 with a probability of ~68% that any single measurement of y_i be within 1 standard deviation of the mean. By making a number of measurements for each value of the independent quantity x_i, we could determine mean values $\bar{y}_i$ with any desired precision. Usually, however, we can make only one measurement y_i for each value of $x = x_i$, so that we must determine the value of y corresponding to that value of x with an uncertainty that is characterized by the standard deviation σ_i of the distribution of data for that point.

We shall assume for simplicity in all the following discussions that we can ascribe all the uncertainty in each measurement to the dependent variable. This is equivalent to assuming that the precision of the determination of x is considerably higher than that of y. This difference is illustrated in Figures 6.1 and 6.2 by the fact that the uncertainties are indicated by error bars for the dependent variables but not for the independent variables.

Our condition, that we neglect uncertainties in x and consider just the uncertainties in y, will be valid only if the uncertainties in y that would be produced by variations in x corresponding to the uncertainties in the measurement of x are much smaller than the uncertainties in the measurement of y. This is equivalent, in first order, to the requirement at each measured point that

$$\sigma_x \frac{dy}{dx} \ll \sigma_y$$

where dy/dx is the slope of the function $y = y(x)$.

We are not always justified in ascribing all uncertainties to the dependent parameter. Sometimes the uncertainties in the determination of both quantities x and y are nearly equal. But our fitting procedure will still be fairly accurate if we estimate the indirect contribution σ_{yI} from the uncertainty σ_x in x to the total uncertainty in y by the first-order relation

$$\sigma_{yI} = \sigma_x \frac{dy}{dx} \tag{6.2}$$

and combine this with the direct contribution σ_{yD}, which is the measuring uncertainty in y, to get

$$\sigma_y^2 = \sigma_{yI}^2 + \sigma_{yD}^2 \tag{6.3}$$

For both Examples 6.1 and 6.2 the condition would be reasonable because we predict a linear dependence of y with x. With the linear assumption, we treat the uncertainties in our data as if they were in the dependent variable only, while realizing that the corresponding fluctuations may have been originally derived from uncertainties in the determinations of both dependent and independent variables.

In those cases where the uncertainties in the determination of the independent quantity are considerably greater than those in the dependent quantity, it might be wise to interchange the definition of the two quantities.

6.2 METHOD OF LEAST SQUARES

Our data consist of pairs of measurements (x_i, y_i) of an independent variable x and a dependent variable y. We wish to find values of the parameters a and b that minimize the discrepancy between the measured values y_i and calculated values $y(x)$. We cannot determine the parameters exactly with only a finite number of observations, but can hope to extract the most probable estimates for the coefficients in the same way that we extracted the most probable estimate of the mean in Chapter 4.

Before proceeding, we must define our criteria for minimizing the discrepancy between the measured and predicted values y_i. For any arbitrary values of a

and b, we can calculate the deviations Δy_i between each of the observed values y_i and the corresponding calculated or fitted values

$$\Delta y_i = y_i - y(x_i) = y_i - a - bx_i \tag{6.4}$$

With well chosen parameters, these deviations should be relatively small. However, the sum of these deviations is not a good measure of how well our calculated straight line approximates the data because large positive deviations can be balanced by negative ones to yield a small sum even when the fit of the function $y(x)$ to the data is bad. We might consider instead summing the absolute values of the deviations, but this leads to difficulties in obtaining an analytical solution. Instead we sum the squares of the deviations.

There in no correct unique method for optimizing the parameters valid for all problems. There exists, however, a method that can be fairly well justified, that is simple and straightforward, and that is well established experimentally. This is the *method of least squares,* similar to the method discussed in Chapter 4, but extended to include more than one variable. It may be considered as a special case of the more general *method of maximum likelihood.*

Method of Maximum Likelihood

Our data consist of a sample of observations drawn from a parent distribution that determines the probability of making any particular observation. For the particular problem of an expected linear relationship between dependent and independent variables, we define parent parameters a_0 and b_0 such that the actual relationship between y and x is given by

$$y_0(x) = a_0 + b_0 x \tag{6.5}$$

We shall assume that each individual measured value of y_i is itself drawn from a Gaussian distribution with mean $y_0(x_i)$ and standard deviation σ_i. We should be aware that the Gaussian assumption may not always be exactly true. In Example 6.2 the $y_i = C_i$ were obtained in a counting experiment and therefore follow a Poisson distribution. However, for a sufficiently large number of counts y_i the distribution may be considered to be Gaussian. We shall discuss fitting with Poisson statistics in Section 6.6.

With the Gaussian assumption, the probability P_i for making the observed measurement y_i with standard deviation σ_i for the observations about the actual value $y_0(x_i)$ is

$$P_i = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{y_i - y_0(x_i)}{\sigma_i}\right]^2\right\} \tag{6.6}$$

The probability for making the observed set of measurements of the N values of y_i is the product of the probabilities for each observation:

$$P(a_0, b_0) = \Pi P_i = \prod\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) \exp\left\{-\frac{1}{2}\sum\left[\frac{y_i - y_0(x_i)}{\sigma_i}\right]^2\right\} \tag{6.7}$$

where the product Π is taken with i ranging from 1 to N and the product of the exponentials has been expressed as the exponential of the sum of the arguments. In these products and sums, the quantities $1/\sigma_i^2$ act as weighting factors.

Similarly, for any *estimated* values of the parameters a and b, we can calculate the probability of obtaining the observed set of measurements

$$P(a, b) = \prod\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right)\exp\left\{-\frac{1}{2}\sum\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2\right\} \tag{6.8}$$

with $y(x)$ defined by Equation (6.1) and evaluated at each of the values x_i.

We assume that the observed set of measurements is more likely to have come from the parent distribution of Equation (6.5) than from any other similar distribution with different coefficients and, therefore, the probability of Equation (6.7) is the maximum probability attainable with Equation (6.8). Thus, the maximum-likelihood estimates for a and b are those values that maximize the probability of Equation (6.8).

Because the first factor in the product of Equation (6.8) is a constant, independent of the values of a and b, maximizing the probability $P(a, b)$ is equivalent to minimizing the sum in the exponential. We define this sum to be our goodness-of-fit parameter χ^2:

$$\chi^2 = \sum\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2 = \sum\left[\frac{1}{\sigma_i}(y_i - a - bx_i)\right]^2 \tag{6.9}$$

We use the same symbol χ^2, defined earlier in Equation (4.32), because this is essentially the same definition in a different context.

Our method for finding the optimum fit to the data will be to find values of a and b that minimize this weighted sum of the squares of the deviations χ^2 and hence, to find the fit that produces the smallest sum of the squares or the *least-squares fit.* The magnitude of χ^2 is determined by four factors:

1. Fluctuations in the measured values of the variables y_i, which are random samples from a parent population with expectation values $y_0(x_i)$.
2. The values assigned to the uncertainties σ_i in the measured variables y_i. Incorrect assignment of the uncertainties σ_i will lead to incorrect values of χ^2.
3. The selection of the analytical function $y(x)$ as an approximation to the "true" function $y_0(x)$. It might be necessary to fit several different functions in order to find the appropriate function for a particular set of data.
4. The values of the parameters of the function $y(x)$. Our objective is to find the "best values" of these parameters.

6.3 MINIMIZING χ^2

To find the values of the parameters a and b that yield the minimum value for χ^2, we set to zero the partial derivatives of χ^2 with respect to each of the parameters

$$
\begin{aligned}
\frac{\partial}{\partial a}\chi^2 &= \frac{\partial}{\partial a}\sum\left[\frac{1}{\sigma_i^2}(y_i - a - bx)^2\right] \\
&= -2\sum\left[\frac{1}{\sigma_i^2}(y_i - a - bx_i)\right] = 0 \\
\frac{\partial}{\partial b}\chi^2 &= \frac{\partial}{\partial b}\sum\left[\frac{1}{\sigma_i^2}(y_i - a - bx_i)^2\right] \\
&= -2\sum\left[\frac{1}{\sigma_i^2}(y_i - a - bx_i)\right] = 0
\end{aligned}
\tag{6.10}
$$

These equations can be rearranged as a pair of linear simultaneous equations in the unknown parameters a and b:

$$
\begin{aligned}
\sum\frac{y_i}{\sigma_i^2} &= a\sum\frac{1}{\sigma_i^2} + b\sum\frac{x_i}{\sigma_i^2} \\
\sum\frac{x_iy_i}{\sigma_i^2} &= a\sum\frac{x_i}{\sigma_i^2} + b\sum\frac{x_i^2}{\sigma_i^2}
\end{aligned}
\tag{6.11}
$$

The solutions can be found in any one of a number of different ways, but, for generality we shall use the method of determinants. (See Appendix B.) The solutions are

$$
\begin{aligned}
a &= \frac{1}{\Delta}\begin{vmatrix} \sum\frac{y_i}{\sigma_i^2} & \sum\frac{x_i}{\sigma_i^2} \\ \sum\frac{x_iy_i}{\sigma_i^2} & \sum\frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta}\left(\sum\frac{x_i^2}{\sigma_i^2}\sum\frac{y_i}{\sigma_i^2} - \sum\frac{x_i}{\sigma_i^2}\sum\frac{x_iy_i}{\sigma_i^2}\right) \\
b &= \frac{1}{\Delta}\begin{vmatrix} \sum\frac{1}{\sigma_i^2} & \sum\frac{y_i}{\sigma_i^2} \\ \sum\frac{x_i}{\sigma_i^2} & \sum\frac{x_iy_i}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta}\left(\sum\frac{1}{\sigma_i^2}\sum\frac{x_iy_i}{\sigma_i^2} - \sum\frac{x_i}{\sigma_i^2}\sum\frac{y_i}{\sigma_i^2}\right) \\
\Delta &= \begin{vmatrix} \sum\frac{1}{\sigma_i^2} & \sum\frac{x_i}{\sigma_i^2} \\ \sum\frac{x_i}{\sigma_i^2} & \sum\frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \sum\frac{1}{\sigma_i^2}\sum\frac{x_i^2}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2
\end{aligned}
\tag{6.12}
$$

For the special case in which all the uncertainties are equal ($\sigma = \sigma_i$), they cancel and the solutions may be written

$$
\begin{aligned}
a &= \frac{1}{\Delta'}\begin{vmatrix} \Sigma y_i & \Sigma x_i \\ \Sigma x_iy_i & \Sigma x_i^2 \end{vmatrix} = \frac{1}{\Delta'}(\Sigma x_i^2\Sigma y_i - \Sigma x_i\Sigma x_iy_i) \\
b &= \frac{1}{\Delta'}\begin{vmatrix} N & \Sigma y_i \\ \Sigma x_i & \Sigma x_iy_i \end{vmatrix} = \frac{1}{\Delta'}(N\Sigma x_iy_i - \Sigma x_i\Sigma y_i) \\
\Delta' &= \begin{vmatrix} N & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{vmatrix} = N\Sigma x_i^2 - (\Sigma x_i)^2
\end{aligned}
\tag{6.13}
$$

Examples

For the data of Example 6.1 (Table 6.1), we assume that the uncertainties in the measured voltages V are all equal and that the uncertainties in x_i are negligible. We can therefore use Equation (6.13). We accumulate four sums Σx_i, $\Sigma y_i = \Sigma V_i$, Σx_i^2, and $\Sigma x_i y_i = \Sigma x_i V_i$ and combine them according to Equation (6.13) to find numerical values for a and b. The steps of the calculation are illustrated in Table 6.1, and the resulting fit is shown as a solid line on Figure 6.1.

Determination of the parameters a and b from Equation (6.12) is somewhat more tedious, because the uncertainties σ_i must be included. Table 6.2 shows steps in the calculation of the data of Example 6.2 with the uncertainties σ_i in the numbers of counts C_i determined by Poisson statistics so that $\sigma_i^2 = C_i$. The values of a and b found in this calculation were used to calculate the straight line through the data points in Figure 6.2.

It is important to note that the value of C_i to be used in determining the uncertainty σ_i must be the actual number of events observed. If, for example, the student had decided to improve her statistics by collecting data at the larger distances over longer time periods Δt_i and to normalize all her data to a common time interval Δt_c,

$$C_i' = C_i \times \Delta t_c / \Delta t_i$$

then the statistical uncertainty in C' would be given by

$$\sigma_i' = \sqrt{C_i} \times \Delta t_c / \Delta t_i$$

Program 6.1. FITLINE (Appendix E) Solution of Equations (6.11) by the determinant method of Equation (6.12).

The program uses routines in the programs units FITVARS, FITUTIL, and GENUTIL, which are also used by other fitting programs. The sample programs use single precision variables for simplicity, although double, or higher, precision is highly recommended.

Program 6.1 uses Equation (6.12) to solve both Examples 6.1 and 6.2, although separate routines written for each problem would be slightly more efficient. Because the measurements of Example 6.1 have common errors, we could, for example, increase the fitting speed by using Equations (6.13) rather than Equations (6.12). Similarly, for Example 6.2, we could simplify the fitting routine by replacing the statistical errors SIGY[I] by the explicit expression for $\sqrt{y_i}$. However, in most calculations that involve statistical errors, there are also other errors to be considered, such as those arising from background subtractions, so the loss of generality would more than compensate for any increased efficiency in the calculations.

Program 6.2. FITVARS (website) Include file of constants, variables, and arrays for least-squares fits.

Program 6.3. FITUTIL (website) Utility routines for fitting programs
Input/output routine, χ^2 calculation, χ^2-density, and χ^2-integral probability.

Program 6.4. GENUTIL (website) General Utility Routines
Includes approximate gamma function, Simpson's rule integration.

6.4 ERROR ESTIMATION

Common Uncertainties

If the standard deviations σ_i for the data points y_i are unknown but we can assume that they are all equal, $\sigma_i^2 = \sigma^2$, then we can estimate them from the data and the results of our fit. The requirement of equal errors may be satisfied if the uncertainties are instrumental and all the data are recorded with the same instrument and on the same scale, as was assumed in Example 6.1.

In Chapter 2 we obtained, for our best estimate of the variance of the data sample,

$$\sigma^2 \simeq s^2 \equiv \frac{1}{N-m}\sum (y_i - \bar{y})^2 \tag{6.14}$$

where $N - m$ is the number of degrees of freedom and is equal to the number of measurements minus the number of parameters determined from the fit. In Equation (6.14) we identify y_i with the measured value of the dependent variable, and for $\bar{y}$, the expected mean value of y_i, we use the value calculated from Equation (6.1) for each data point with the fitted parameters a and b. Thus, our estimate $\sigma_i = \sigma$ for the standard deviation of an individual measurement is

$$\sigma^2 \simeq s^2 = \frac{1}{N-2}\sum (y_i - a - bx_i)^2 \tag{6.15}$$

By comparing Equation (6.15) with Equation (6.9), we see that it is just this common uncertainty that we have minimized in the least-squares fitting procedure. Thus, we can obtain the common error in our measurements of y from the fit, although at the expense of any information about the quality of the fit.

Variable Uncertainties

In general the uncertainties σ_i in the dependent variables y_i will not all be the same. If, for example, the quantity y represents the number of counts in a detector per unit time interval (as in Example 6.2), then the errors are statistical and the uncertainty in each measurement y_i is directly related to the magnitude of y (as discussed in Section 4.2), and the standard deviations σ_i associated with these measurements is

$$\sigma_i^2 = C_i \tag{6.16}$$

In principle, the value of y_i, which should be used in calculating the standard deviations σ_i by Equation (6.16), is the value $y_0(x_i)$ of the parent population. In practice we use the measured values that are only samples from that population. In the limit of an infinite number of determinations, the average of all the measurements would very closely approximate the parent value, but generally we cannot make more than one measurement of each value of x, much less an infinite number. We could approximate the parent value $y_0(x_i)$ by using the calculated value $y(x)$ from our fit, but that would complicate the fitting procedure. We shall discuss this possibility further in the following section.

Contributions from instrumental and other uncertainties may modify the simple square root form of the statistical errors. For example, uncertainties in measuring the time interval during which the events of Example 6.2 were recorded might contribute, although statistical fluctuations generally dominate in counting experiments. Background subtractions are another source of uncertainty. In many counting experiments, there is a background under the data that may be removed by subtraction, or may be included in the fit. In Example 6.2, cosmic rays and other backgrounds contribute to a counting rate even when the source is moved far away from the detector, as indicated by the nonzero intercept of the fitted line of Figure 6.2 on the C axis. If the student had chosen to record the radiation background counts C_b in a separate measurement and to subtract C_b from each of her measurements C_i to obtain

$$C_i' = C_i - C_b$$

then the uncertainty in C' would have been given by combining in quadrature the uncertainties in the two measurements:

$$\sigma_i'^2 = \sigma_i^2 + \sigma_b^2$$

χ^2 Probability

For those data for which we know the uncertainties σ_i in the measured values y_i we can calculate the value of χ^2 from Equation (6.9) and test the goodness of our fit. For our two-parameter fit to a straight line, the number of degrees of freedom will be $N - 2$. Then, for the data of Example 6.2, we should hope to obtain $\chi^2 \simeq 10 - 2 = 8$. The actual value, $\chi^2 = 11.1$, is listed in Table 6.2, along with the probability ($p = 20\%$). (See Table C.4.) We interpret this probability in the following way. Suppose that we have obtained a χ^2 probability of $p\%$ for a certain set of data. Then, we should expect that, if we were to repeat the experiment many times, approximately $p\%$ of the experiments would yield χ^2 values as high as the one that we obtained or higher. This subject will be discussed further in Chapter 11.

In Example 6.1, we obtained a value of $\chi^2 = 1.95$ for 7 degrees of freedom, corresponding to a probability of about 96%. Although this probability may seem to be gratifyingly high, the very low value of χ^2 gives a strong indication that the common uncertainty in the data may have been overestimated and it might be wise to use the value of χ^2 to obtain a better estimate of the common uncertainty. From Equations (6.15) and (6.9), we obtain an expression for the revised common uncertainty σ_c' in terms of χ^2 and the original estimate, σ_c:

$$\sigma_c'^2 \simeq \sigma_i^2 \times \chi^2/(N-2) \tag{6.17}$$

or, more generally

$$\sigma_c'^2 \simeq \sigma_i^2 \times \chi_\nu^2 \tag{6.18}$$

where $\chi_\nu^2 = \chi^2/\nu$ and ν is the number of degrees of freedom in the fit. Thus, for Example 6.1, we find $\sigma_c'^2 = 0.05^2 \times 1.95/(9 - 2) = 0.0007$, or $\sigma_c' = {\sim}0.03$ V.

Uncertainties in the Parameters

In order to find the uncertainty in the estimation of the parameters a and b in our fitting procedure, we use the error propagation method discussed in Chapter 3. Each of our data points y_i has been used in the determination of the parameters and each has contributed some fraction of its own uncertainty to the uncertainty in our final determination. Ignoring systematic errors, which would introduce correlations between uncertainties, the variance σ_z^2 of the parameter z is given by Equation (3.14) as the sum of the squares of the products of the standard deviations σ_i of the data points with the effects that the data points have on the determination of z:

$$\sigma_z^2 = \sum\left[\sigma_i^2\left(\frac{\partial z}{\partial y_i}\right)^2\right] \tag{6.19}$$

Thus, to determine the uncertainties in the parameters a and b, we take the partial derivatives of Equation (6.12):

$$\begin{aligned}\frac{\partial a}{\partial y_j} &= \frac{1}{\Delta}\left(\frac{1}{\sigma_j^2}\sum\frac{x_i^2}{\sigma_i^2} - \frac{x_j}{\sigma_j^2}\sum\frac{x_i}{\sigma_i^2}\right)\\ \frac{\partial b}{\partial y_j} &= \frac{1}{\Delta}\left(\frac{x_j}{\sigma_j^2}\sum\frac{1}{\sigma_i^2} - \frac{1}{\sigma_j^2}\sum\frac{x_i}{\sigma_i^2}\right)\end{aligned} \tag{6.20}$$

We note that the derivatives are functions only of the variances and of the independent variables x_i. Combining these equations with the general expression of Equation (6.19) and squaring, we obtain for σ^2,

$$\begin{aligned}\sigma_a^2 &\simeq \sum_{j=1}^{N}\frac{\sigma_j^2}{\Delta^2}\left[\frac{1}{\sigma_j^4}\left(\sum\frac{x_i^2}{\sigma_i^2}\right)^2 - \frac{2x_j}{\sigma_j^4}\sum\frac{x_i^2}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \frac{x_j^2}{\sigma_j^4}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta^2}\left[\sum\frac{1}{\sigma_j^2}\left(\sum\frac{x_i^2}{\sigma_i^2}\right)^2 - 2\sum\frac{x_j}{\sigma_j^2}\sum\frac{x_i^2}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \sum\frac{x_j^2}{\sigma_j^2}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta^2}\left(\sum\frac{x_i^2}{\sigma_i^2}\right)\left[\sum\frac{1}{\sigma_j^2}\sum\frac{x_i^2}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta}\sum\frac{x_i^2}{\sigma_i^2}\end{aligned} \tag{6.21}$$

and for σ_b^2,

$$\begin{aligned}\sigma_b^2 &\simeq \sum_{j=1}^{N}\frac{\sigma_j^2}{\Delta^2}\left[\frac{x_j^2}{\sigma_j^4}\left(\sum\frac{1}{\sigma_i^2}\right)^2 - \frac{2x_j}{\sigma_j^4}\sum\frac{1}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \frac{1}{\sigma_j^4}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta^2}\left[\sum\frac{x_j^2}{\sigma_j^2}\left(\sum\frac{1}{\sigma_i^2}\right)^2 - 2\sum\frac{x_j}{\sigma_j^2}\sum\frac{1}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \sum\frac{1}{\sigma_j^2}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta^2}\left(\sum\frac{x_j^2}{\sigma_i^2}\right)\left[\sum\frac{1}{\sigma_j^2}\sum\frac{1}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]\\ &= \frac{1}{\Delta}\sum\frac{1}{\sigma_i^2}\end{aligned} \tag{6.22}$$

For the special case of common uncertainties in y_i, $\sigma_i = \sigma$, these equations reduce to

$$\sigma_a^2 = \frac{\sigma^2}{\Delta'}\Sigma x_i^2 \qquad \text{and} \qquad \sigma_b^2 = N\frac{\sigma^2}{\Delta'} \tag{6.23}$$

with σ given by Equation (6.15) and Δ' given by Equation (6.13).

The uncertainties in the parameters σ_a and σ_b, calculated from the original error estimates, are listed in Tables 6.1 and 6.2. For Example 6.1, revised uncertainties σ_a' and σ_b', based on the revised common data uncertainty calculated from Equation (6.18), are also listed.

6.5 SOME LIMITATIONS OF THE LEAST-SQUARES METHOD

When a curve is fitted by the least-squares method to a collection of statistical counting data, the data must first be *histogrammed*; that is, a histogram must be formed of the corrected data, either during or after data collection. In Example 6.2, the data were collected over intervals of time Δt, with the size of the interval chosen to assure that a reasonable number of counts would be collected in each time interval. For data that vary linearly with the independent variable, this treatment poses no special problems, but one could imagine a more complex problem in which fine details of the variation of the dependent variable y with the independent variable x are important. Such details might well be lost if the binning were too coarse. On the other hand, if the binning interval were too fine, there might not be enough counts in each bin to justify the Gaussian probability hypothesis. How does one choose the appropriate bin size for the data?

A handy rule of thumb when considering the Poisson distribution is to assume that *large enough* = 10. A comparison of the Gaussian and Poisson distributions for mean $\mu \simeq 10$ and standard deviation $\sigma = \sqrt{\mu}$ (see Figures 2.4 and 2.5) shows very little difference between the two distributions. We might expect this because the mean is more than 3 standard deviations away from the origin. Thus, we may be reasonably confident about the results of a fit if no histogram contains less than ten counts and if we are not placing excessive reliance on the actual value of χ^2 obtained from the fit. If a bin does have fewer than the allowed minimum number of counts, it may be possible to merge that bin with an adjacent one. Note that there is no requirement that intervals on the abscissa be equal, although we must be careful in our choice of the appropriate value of x_i for the merged bin. We should also be aware that such mergers necessarily reduce the resolution of our data and may, when fitting functions more complicated than a straight line, obscure some interesting features.

In general, the choice of bin width will be a compromise between the need for sufficient statistics to maintain a small relative error in the values of y_i and thus in the fitted parameters, and the need to preserve interesting structure in the data. When full details of any structure in the data must be preserved, it might be advisable to apply the maximum-likelihood method directly to the data, event by event, rather than to use the least-squares method with its necessary binning of the data. We return to this subject in Chapter 10.

There is also a question about our use of the experimental errors in the fitting process, rather than the errors predicted by our estimate of the parent distribution. For Example 6.2, this corresponds to our choosing $\sigma_i^2 = y_i$ rather than $\sigma_i^2 = y(x_i) = a + bx_i$. We shall consider the possibility of using errors from our estimate of the parent distribution, as well as the direct application of the Poisson probability function, in the following section.

Another important point to consider when fitting curves to data is the possibility of rounding errors, which can reduce the accuracy of the results. With manual calculations, it is important to avoid rounding the numbers until the very end of the calculation. With computers, problems may arise because of finite computer word length. This problem can be especially severe with matrix and determinant calculations, which often involve taking small differences between large numbers. Depending on the computer and the software, it may be necessary to use double-precision variables in the fitting routine.

We discuss in Chapter 7 the interaction of parameters in a multiparameter fit. For now, it is worth noting that, for a nominally "flat" distribution of data, the intercept obtained from a fit to a straight line may not be identical to the mean value of the data points on the ordinate. See Exercise 6.7 for an example of this effect.

6.6 ALTERNATE FITTING METHODS

In this section we attempt to solve the problem of fitting a straight line to a collection of data points by using errors determined from the estimated parent distribution rather than from the measurements, and by directly applying Poisson statistics, rather than Gaussian statistics. Because it is not possible to derive a set of independent linear equations for the parameters with these conditions, explicit expressions for the parameters a and b cannot be obtained. However, with fast computers, solving coupled, nonlinear equations is not difficult, although the clarity and elegance of the straightforward least-squares method can be lost.

Poisson Uncertainties

Let us consider a collection of purely statistical data that obey Poisson statistics (as in Example 6.2) so that the uncertainties can be expressed by Equation (6.16). We begin by substituting the approximation $\sigma_i^2 = y(x_i) = a + bx_i$ into the definition of χ^2 in Equation (6.9), which is based on Gaussian probability, and minimizing the value of χ^2 as in Equations (6.10). The result is a pair of simultaneous equations that can be solved for a and b:

$$\begin{aligned} N &= \sum \frac{y_i^2}{(a + bx_i)^2} \\ \Sigma x_i &= \sum \frac{x_i y_i^2}{(a + bx_i)^2} \end{aligned} \qquad (6.24)$$

Poisson Probability

Next, let us replace the Gaussian probability $P(a, b)$ of Equation (6.8) by the corresponding probability for observing y_i counts from a Poisson distribution with mean $\mu_i = y(x_i)$,

$$P(a,b) = \prod\left(\frac{[y(x_i)]^{y_i}}{y_i!}e^{-y(x_i)}\right) \tag{6.25}$$

and apply the method of maximum likelihood to this probability. It is easier and equivalent to maximize the natural logarithm of the probability with respect to each of the parameters a and b:

$$\ln P(a,b) = \Sigma[y_i \ln y(x_i)] - \Sigma y(x_i) + \text{constant} \tag{6.26}$$

where the constant term is independent of the parameters a and b. The result of taking partial derivatives of Equation (6.26) is a pair of simultaneous equations similar to those of Equation (6.24),

$$\begin{aligned} N &= \sum \frac{y_i}{a + bx_i} \\ \Sigma x_i &= \sum \frac{x_i y_i}{a + bx_i} \end{aligned} \tag{6.27}$$

but with less emphasis on fitting the larger values of y_i.

Neither the coupled simultaneous Equations (6.24) nor the Equations (6.27) can be solved directly for a and b, but each pair can be solved by an iterative method in which values of a and b are chosen and then adjusted until the two simultaneous equations are satisfied. (See Appendix A.5.)

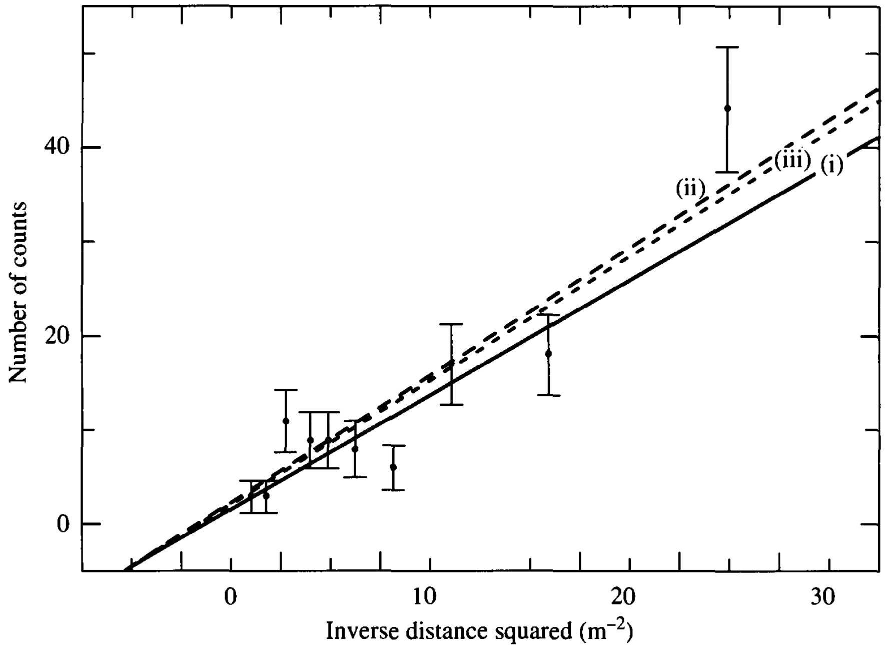

FIGURE 6.3

Least-squares fit of a straight line to the data by three different methods. (i) Standard least-squares method with Gaussian statistics and experimental uncertainties; (ii) Gaussian statistics and analytic uncertainties; (iii) Poisson statistics and analytic uncertainties. The analytic errors are expressed as $\sigma_i^2 = a + bx_i$.

TABLE 6.3

Comparison of fits to a selection of statistical data from Example 6.2 for three different fitting methods

i	Inverse distance squared $x_i\ (m^{-2})$	Number of counts C_i	(1) Standard	(2) Gaussian $\sigma^2 = y(x_i)$	(3) Poisson $\sigma^2 = y(x_i)$
1	25.00	44	32.0	36.3	35.1
2	16.00	18	21.0	24.1	23.2
3	11.11	17	15.1	17.5	16.8
4	8.16	6	11.5	13.5	12.9
5	6.25	8	9.1	10.9	10.4
6	4.94	9	7.5	9.2	8.6
7	4.00	9	6.4	7.9	7.4
8	2.78	11	4.9	6.3	5.8
9	1.78	3	3.7	4.9	4.5
10	1.00	3	2.7	3.9	3.4
Sums		128	114.0	134.4	128.0
		a	1.52	2.50	2.11
		b	1.22	1.35	1.32
		χ^2	13.7	17.6	15.5

Note: (1) Standard least-squares method with Gaussian statistics and experimental uncertainties; (2) Gaussian statistics and analytic uncertainties; (3) Poisson statistics and analytic uncertainties. The analytic uncertainties are expressed as $\sigma^2 = a + bx_i$.

Example 6.3. Because we expect the methods discussed here to be equivalent to the standard method for large data samples, we selected a low statistics sample to emphasize the differences. We chose from the measurements of Example 6.2 only those events collected at each detector position during the first 15-s interval, a total of 128 events at ten different positions. The results of (i) calculations by the standard method, (ii) calculations with Gaussian statistics and with errors given by $\sigma_i = y(x_i) = a + bx_i$, and (iii) calculations with Poisson statistics with errors as in method (ii) are listed in Table 6.3 and illustrated in Figure 6.3. We note that method (i) appears to underestimate the number of events in the sample, whereas method (ii) overestimates the number. Method (iii) with Poisson statistics and errors calculated as in method (ii) finds the exact number.

We can avoid questions of finite binning and the choice of statistics by making direct use of the maximum-likelihood method, treating the fitting function as a probability distribution. This method also allows detailed handling of problems in which the probability associated with individual measurements varies in a complex way from observation to observation. We shall pursue this subject further in Chapter 10.

In general, however, the simplicity of the least-squares method and the difficulty of solving the equations that result from other methods, particularly with more complicated fitting functions, leads us to choose the standard method of least squares for most problems. We make the following two assumptions to simplify the calculation:

1. The shapes of the individual Poisson distributions governing the fluctuations in the observed y_i are nearly Gaussian.
2. The uncertainties σ_i in the observations y_i may be obtained from the uncertainties in the data and may be approximated by $\sigma_i^2 \simeq y_i$ for statistical uncertainties.

SUMMARY

Linear function: $y(x) = a + bx$.
Chi-square:

$$\chi^2 = \sum \left[\frac{1}{\sigma_i}(y_i - a - bx_i) \right]^2$$

Least-squares fitting procedure: Minimize χ^2 with respect to each of the parameters simultaneously.
Solutions for least-squares fit of a straight line:

$$a = \frac{1}{\Delta} \begin{vmatrix} \sum \frac{y_i}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i y_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta} \left(\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} \right)$$

$$b = \frac{1}{\Delta} \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{y_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i y_i}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta} \left(\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} \right)$$

$$\Delta = \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2$$

Estimated uniform variance s^2:

$$\sigma^2 \simeq s^2 = \frac{1}{N-2} \sum (y_i - \bar{y})^2$$

Statistical fluctuations:

$$\sigma_i^2 \simeq y_i \qquad \text{(raw data counts)}$$

Uncertainties in coefficients:

$$\sigma_a^2 = \frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2} \qquad \sigma_b^2 = \frac{1}{\Delta} \sum \frac{1}{\sigma_i^2}$$