1. The shapes of the individual Poisson distributions governing the fluctuations in the observed $y_i$ are nearly Gaussian.
2. The uncertainties $\sigma_i$ 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 $\chi^2$ 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 $s^2$*:

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

*Statistical fluctuations*:

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

*Uncertainties in coefficients*:

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



## EXERCISES

**6.1.** Fit the data of Example 6.2 as if all the data had equal uncertainties $\sigma_i = \bar{\sigma} = 18.5$, where $\bar{\sigma}$ is the average of the given values of $\sigma$. Note that the fitted parameters are independent of the value of $\bar{\sigma}$, but the values of $\chi^2$, $\sigma_a$, and $\sigma_b$ are not.

**6.2.** Derive Equation (6.23) from Equations (6.21) and (6.22).

**6.3.** Show that Equation (6.12) reduces to Equation (6.13) if $\sigma_i = \sigma$.

**6.4.** Derive a formula for making a linear fit to data with an intercept at the origin so that $y = bx$. Apply your method to fit a straight line through the origin to the following coordinate pairs. Assume uniform uncertainties $\sigma_i = 1.5$ in $y_i$. Find $\chi^2$ for the fit and the uncertainty in $b$.

| $x_i$ | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 5.3 | 14.4 | 20.7 | 30.1 | 35.0 | 41.3 | 52.7 | 55.7 | 63.0 | 72.1 | 80.5 | 87.9 |

**6.5.** A student hangs masses on a spring and measures the spring's extension as a function of the applied force in order to find the spring constant $k$. Her measurements are:

| Mass (kg) | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 |
|---|---|---|---|---|---|---|---|---|
| Extension (cm) | 5.1 | 5.5 | 5.9 | 6.8 | 7.4 | 7.5 | 8.6 | 9.4 |

There is an uncertainty of 0.2 in each measurement of the extension. The uncertainty in the masses is negligible. For a perfect spring, the extension $\Delta L$ of the spring will be related to the applied force by the relation $k\Delta L = F$, where $F = mg$, and $\Delta L = L - L_0$, and $L_0$ is the unstretched length of the spring. Use these data and the method of least squares to find the spring constant $k$, the unstretched length of the spring $L_0$, and their uncertainties. Find $\chi^2$ for the fit and the associated probability.

**6.6.** Outline a procedure for solving the simultaneous Equations (6.27). Refer to Appendix A.

**6.7.** A student measures the temperature ($T$) of water in an insulated flask at times ($t$) separated by 1 minute and obtains the following values:

| $t$(s) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $T$(°C) | 98.51 | 98.50 | 98.50 | 98.49 | 98.52 | 98.49 | 98.52 | 98.45 | 98.47 |

(*a*) Calculate the mean temperature and its standard error.
(*b*) To test whether or not the water is cooling, plot a graph of the temperatures versus the time and make a least-squares fit of a straight line to the data. Is there a statistically significant slope to the graph?
(*c*) Note that the intercept is not identical to the mean value of the temperature you calculated in part (*a*). Now, shift the time coordinates by 4 s so that the mean time is 0. Refit the data with the new values of $T$. Is the intercept now identical to the mean value of $T$?
(*d*) Clearly, the results of this experiment cannot depend upon the time at which the measurements were made. Show that, if the mean value of $x$ is equal to zero, then the intercept $b$ calculated from Equation (6.13) is identically equal to the mean value of $y$.

**Computer fits** Routines used for fitting a series of Legendre polynomials to these data are included in Program 7.1. The procedure LEGPOLY in the program unit FITFUNC7 calculates the terms of the Legendre polynomials through tenth order. The procedure is selected through a branch on the variable PAE in the function Funct with PAE = 'A' for all terms to order $n = m - 1$, or PAE = 'E' to fit with just the even terms. Note that the index $k$ of the term in the fitting function, in general, does not correspond to the order $L$ of the Legendre polynomial.

The efficiency of the calculation (and therefore the speed of the linear regression calculation) could be improved in a number of ways. The simplest change would be to calculate the functions once at each value of the independent variable and store the calculated values in an array.

Parameters obtained by fitting a series in Legendre polynomials for terms up to $L = 4$ are listed in Table 7.7. Separate fits were made with all terms and with only the even terms in the series. As expected, the coefficients of terms involving odd orders are comparable to their uncertainties and negligible compared to those involving even poynomials. The full error matrix for the fit with even terms is listed in Table 7.8.

In view of the strong theoretical argument that only even Legendre polynomials are required for this reaction, it would be appropriate to fit a series that includes only the even terms. The parameters obtained in this fit are also displayed in Table 7.7, and the numbers of counts calculated from these parameters are listed. The function calculated with even terms is illustrated as a curve on the data of Figure 7.2.

Because we are fitting with orthogonal functions, we might have expected to obtain identical values for the coefficient $a_0$ from both fits. (We expect the higher-order even coefficients to change because the presence or absence of lower-order coefficients must affect the higher coefficients.) The fact that there is some dependence of $a_0$ on higher-order terms is a result of the fact that a given experiment does not sample uniformly the entire range of the Legendre polynomial, so the orthogonality relation Equation (7.43) is not satisfied by a finite data set. This is in contrast to the situation in the previous section, where we set up orthogonal functions based on the data themselves. Nevertheless, it is generally good practice to use orthogonal

**TABLE 7.7**
**Coefficients and $\chi^2$ from least-squares fit to Legendre polynomial series**

| | $\chi^2$ | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|---|---|
| All terms | 17.2(14%) | 937.4 ± 7.6 | 0.7 ± 12.8 | 259 ± 14 | 10 ± 17 | 158 ± 18 |
| Even terms | 17.6(22%) | 938.1 ± 7.5 | | 261 ± 14 | | 161 ± 16 |

**TABLE 7.8**
**Error matrix for a least-squares fit to even Legendre polynomials**

$$\begin{bmatrix} 56.24 & -5.256 & -6.272 \\ -5.256 & 186.5 & -26.90 \\ -6.272 & -26.90 & 279.8 \end{bmatrix}$$



fitting functions whenever possible to minimize both the correlations between coefficients and the dependence of higher coefficients on the presence of lower ones.

The values of $\chi^2$ and the $\chi^2$-probability for the two fits are also given in Table 7.7. We note that $\chi^2$ for the three-parameter fit is necessarily higher than that for the five-parameter fit, but $\chi^2$ per degree of freedom is smaller and the $\chi^2$-probability is higher.

## 7.4 NONLINEAR FUNCTIONS

In all the procedures developed so far we have assumed that the fitting function was linear in the coefficients. By that we mean that the function can be expressed as a sum of separate terms each multiplied by a single coefficient. How can we fit data with a function that is not linear in the coefficients? For example, suppose we have measured the distribution of decay times of an unstable state and that the distribution can be represented by the normalized function $P(t) = (1/\tau)e^{-t/\tau}$, where $\tau$ is the mean lifetime of the state. Can we find the parameter $\tau$ by the least-squares method? The method of least squares does not yield a straightforward analytical solution for such functions. In Chapter 8 we investigate methods of searching parameter space for values of the coefficients that will minimize the goodness-of-fit criterion $\chi^2$. Here we consider approximate solutions to such problems using linear-regression techniques.

### Linearization

It is possible to transform some functions into linear functions. For example, if we were to fit an exponential decay problem of the form

$$y = ae^{-bx} \tag{7.46}$$

where $a$ and $b$ are the unknown parameters, it would seem reasonable to take logarithms of both sides and to fit the resulting straight line equation

$$\ln y = \ln a - bx \tag{7.47}$$

The method of least squares minimizes the value of $\chi^2$ with respect to each of the coefficients $\ln a$ and $\ln b$ where $\chi^2$ is given by

$$\chi^2 = \sum \left\{ \frac{1}{\sigma_i'^2} [\ln y_i + \ln a - bx_i]^2 \right\} \tag{7.48}$$

where we must use weighted uncertainties $\sigma_i'$ instead of $\sigma_i$ to account for the transformation of the dependent variable:

$$\sigma_i' = \frac{d(\ln y_i)}{dy} \sigma_i = \frac{1}{y_i} \sigma_i \tag{7.49}$$

The importance of weighting the uncertainties is illustrated in Figure 7.3, which shows the function of Equation (7.46) graphed both on a linear and on a logarithmic scale. (For plotting, we use logarithms to base 10 rather than natural logarithms.) The uncertainties are given by $\sigma_i = \sqrt{y_i}$ and therefore increase with increasing $y_i$.

# CHAPTER 8

# LEAST-SQUARES FIT TO AN ARBITRARY FUNCTION

## 8.1 NONLINEAR FITTING

The methods of least squares and multiple regression developed in the previous chapters are restricted to fitting functions that are linear in the parameters as in Equation (7.3):

$$y(x) = \sum_{j=1}^{m} [a_j f_j(x)] \tag{8.1}$$

This limitation is imposed by the fact that, in general, minimizing $\chi^2$ can yield a set of coupled equations that are linear in the $m$ unknown parameters only if the fitting functions $y(x)$ are themselves linear in the parameters. We shall distinguish between the two types of problems by referring to *linear fitting* for problems that involve equations that are linear in the parameters, such as those discussed in Chapters 6 and 7, and *nonlinear fitting* for those problems that are nonlinear in the parameters.

**Example 8.1.** In a popular undergraduate physics laboratory experiment, a real silver quarter is irradiated with thermal neutrons to create two short-lived isotopes of silver, $_{47}Ag^{108}$ and $_{47}Ag^{110}$, that subsequently decay by beta emission. Students count the emitted beta particles in 15-s intervals for about 4 min to obtain a decay curve. Data collected from such an experiment are listed in Table 8.1 and plotted on a semilogarithmic graph in Figure 8.1. The data are reported at the end of each 15-s interval, just as they were recorded by a scaler. The data points do not fall on a straight





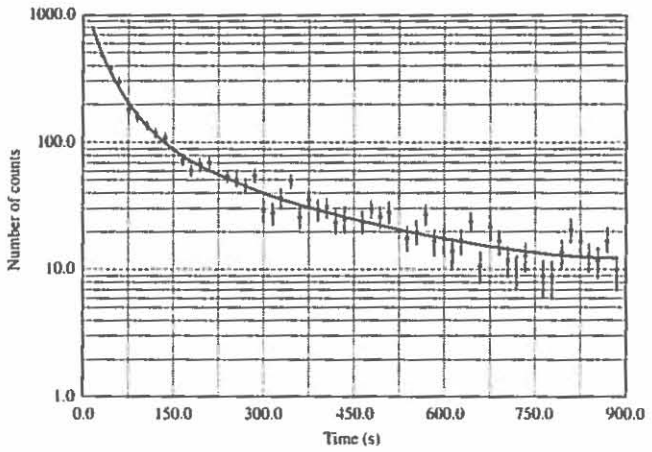


**FIGURE 8.1**
Number of counts detected from the decay of two excited states of silver as a function of time (Example 8.1). Time is reported at the end of each interval. Statistical uncertainties are assumed. The curve was obtained by a nonlinear least-squares fit of Equation (8.2) to the data.

line because the probability function that describes the process is the sum of two exponential functions plus a constant background. We can represent the decay by the fitting function

$$y(x_i) = a_1 + a_2 e^{-t/a_4} + a_3 e^{-t/a_5} \tag{8.2}$$

where the parameter $a_1$ corresponds to the background radiation and $a_2$ and $a_3$ correspond to the amplitudes of the two excited states with mean lives $a_4$ and $a_5$, respectively. Clearly, Equation (8.2) is not linear in the parameters $a_4$ and $a_5$, although it is linear in the parameters $a_1$, $a_2$, and $a_3$.

We can use a graphical analysis method to find the two mean lifetimes by plotting the data on semilogarithmic paper after first subtracting from each data point the constant background contribution, which has been measured separately. (Note that the background counts have not been subtracted in Figure 8.1.) We then consider two regions of the plot: region $a$, at small values of $T$ (e.g., $T < 120$ s) in which the short-lived state dominates the plot, and region $b$, at large values of $T$ (e.g., $T > 200$ s) in which only the long-lived state contributes to the data. We can estimate the mean lifetime of the long-lived state by finding the slope of our best estimate of the straight line that passes through the data points in region $b$. From this result we can estimate the contribution of the long-lived component to region $a$ and subtract that contribution from each of the data points, and thus make a new plot of the number of counts in region $a$, which we attribute to the short-lived state alone. The slope of the line through the corrected points gives us the mean lifetime of the short-lived state. Linear regression techniques discussed in Section 7.4 could be used to find the slope of the graph in each region.

TABLE 8.1
**Geiger counter data from an irradiated silver piece, recorded in 15-s intervals**

| Point number | Time | Measured counts | Calculated counts | Point number | Time | Measured counts | Calculated counts |
|---|---|---|---|---|---|---|---|
| 1 | 15 | 775 | 748.3 | 31 | 465 | 24 | 24.0 |
| 2 | 30 | 479 | 519.8 | 32 | 480 | 30 | 23.0 |
| 3 | 45 | 380 | 370.4 | 33 | 495 | 26 | 22.1 |
| 4 | 60 | 302 | 272.0 | 34 | 510 | 28 | 21.3 |
| 5 | 75 | 185 | 206.7 | 35 | 525 | 21 | 20.5 |
| 6 | 90 | 157 | 162.7 | 36 | 540 | 18 | 19.8 |
| 7 | 105 | 137 | 132.5 | 37 | 555 | 20 | 19.2 |
| 8 | 120 | 119 | 111.5 | 38 | 570 | 27 | 18.5 |
| 9 | 135 | 110 | 96.3 | 39 | 585 | 17 | 18.0 |
| 10 | 150 | 89 | 85.0 | 40 | 600 | 17 | 17.4 |
| 11 | 165 | 74 | 76.5 | 41 | 615 | 14 | 16.9 |
| 12 | 180 | 61 | 69.7 | 42 | 630 | 17 | 16.5 |
| 13 | 195 | 66 | 64.2 | 43 | 645 | 24 | 16.0 |
| 14 | 210 | 68 | 59.5 | 44 | 660 | 11 | 15.6 |
| 15 | 225 | 48 | 55.5 | 45 | 675 | 22 | 15.2 |
| 16 | 240 | 54 | 51.9 | 46 | 690 | 17 | 14.9 |
| 17 | 255 | 51 | 48.8 | 47 | 705 | 12 | 14.6 |
| 18 | 270 | 46 | 45.9 | 48 | 720 | 10 | 14.3 |
| 19 | 285 | 55 | 43.3 | 49 | 735 | 13 | 14.0 |
| 20 | 300 | 29 | 40.9 | 50 | 750 | 16 | 13.8 |
| 21 | 315 | 28 | 38.7 | 51 | 765 | 9 | 13.5 |
| 22 | 330 | 37 | 36.7 | 52 | 780 | 9 | 13.3 |
| 23 | 345 | 49 | 34.8 | 53 | 795 | 14 | 13.1 |
| 24 | 360 | 26 | 33.1 | 54 | 810 | 21 | 12.9 |
| 25 | 375 | 35 | 31.5 | 55 | 825 | 17 | 12.7 |
| 26 | 390 | 29 | 30.0 | 56 | 840 | 13 | 12.6 |
| 27 | 405 | 31 | 28.6 | 57 | 855 | 12 | 12.4 |
| 28 | 420 | 24 | 27.3 | 58 | 870 | 18 | 12.3 |
| 29 | 435 | 25 | 26.1 | 59 | 885 | 10 | 12.1 |
| 30 | 450 | 35 | 25.0 | | | | |

*Note:* The time is reported at the end of each interval. The calculated number of counts was found by method 4.

Because analytic methods of least-squares fitting cannot be used for nonlinear fitting problems, we must consider approximation methods and make searches of parameter space. In the following sections we discuss four nonlinear fitting methods: a simple grid-search method in which we simply calculate $\chi^2$ at trial values of the parameters, and search for those values of the parameters that yield a minimum value of $\chi^2$, a gradient-search method that uses the slope of the function to improve the efficiency of the search, and two semianalytic methods that make use of the matrix method developed in Chapter 7, with a linear approximation to the nonlinear functions. As examples, we shall determine the parameters $(a_1 \ldots a_5)$ by fitting Equation (8.2) to the data of Example 8.1 using each of the four methods. The curve on Figure 8.1 is the result of such a fit.



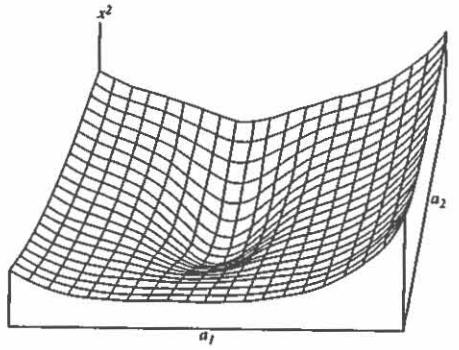


FIGURE 8.2
Chi-square hypersurface as a function of two parameters.

## Method of Least Squares

We can generalize the probability function, or *likelihood function*, of Equation (6.7) to any number of parameters,

$$P(a_1, a_2, \ldots, a_m) = \prod \left[\frac{1}{\sigma_i\sqrt{2\pi}}\right] \exp\left\{-\frac{1}{2}\sum\left[\frac{y_i - y(x_i)}{\sigma_i}\right]^2\right\} \tag{8.3}$$

and, as in the previous chapters, maximize the likelihood with respect to the parameters by minimizing the exponent, or the goodness-of-fit parameter $\chi^2$:

$$\chi^2 \equiv \sum\left\{\frac{1}{\sigma_i^2}[y_i - y(x_i)]^2\right\} \tag{8.4}$$

where $x_i$ and $y_i$ are the measured variables, $\sigma_i$ is the uncertainty in $y_i$, and $y(x_i)$ are values of the function calculated at $x_i$. According to the method of least squares, the optimum values of the parameters $a_j$ are obtained by minimizing $\chi^2$ simultaneously with respect to each parameter,

$$\begin{aligned}\frac{\partial\chi^2}{\partial a_j} &= \frac{\partial}{\partial a_j}\sum\left\{\frac{1}{\sigma_i^2}[y_i - y(x_i)]^2\right\} = 0 \\ &= -2\sum\left\{\frac{1}{\sigma_i^2}[y_i - y(x_i)]\frac{\partial y(x_i)}{\partial a_j}\right\}\end{aligned} \tag{8.5}$$

Taking partial derivatives of $\chi^2$ with respect to each of the $m$ parameters $a_j$ will yield $m$ coupled equations in the $m$ unknown parameters $a_j$ as in Section 7.1. If these equations are not linear in all the parameters, we must, in general, treat $\chi^2$ as a continuous function of the $m$ parameters, describing a hypersurface in an $m$-dimensional space, as expressed by Equation (8.4), and search that space for the appropriate minimum

value of $\chi^2$. Figure 8.2 illustrates such a hyperspace for a function of two parameters. Alternatively, we may apply to the $m$ equations obtained from Equations (8.5) approximation methods developed for finding roots of coupled, nonlinear equations. A combination of both methods is often used.

## Variation of $\chi^2$ Near a Minimum

For a sufficiently large event sample, the likelihood function becomes a Gaussian function of each parameter centered on those values $a_j'$ that minimize $\chi^2$:

$$P(a_j) = Ae^{-(a_j - a_j')^2/2\sigma_j^2} \tag{8.6}$$

where $A$ is a function of the other parameters, but not of $a_j$. Comparing Equation (8.3) for the likelihood function with Equation (8.4) for $\chi^2$, we observe that we can express $\chi^2$ as

$$\chi^2 = -2\ln[P(a_1, a_2, \ldots, a_m)] + 2\sum \ln(\sigma_i\sqrt{2\pi}) \tag{8.7}$$

Then, from Equation (8.6), we can write

$$\chi^2 = \frac{(a_j - a_j')^2}{\sigma_j^2} + C \tag{8.8}$$

to show the variation of $\chi^2$ with any single parameter $a_j$ in the vicinity of a minimum with respect to that parameter. The constant $C$ is a function of the uncertainties $\sigma_i$ and the parameters $a_k$ for $k \neq j$. Thus $\chi^2$ varies as the square of distance from a minimum, and an increase of 1 standard deviation ($\sigma$) in the parameter from the value $a_j'$ at the minimum increases $\chi^2$ by 1. For a more general proof, see Arndt and MacGregor (1966), appendix II.

We can see that this result is consistent with that obtained from a second-order Taylor expansion of $\chi^2$ about the values $a_j'$, where the values of $\chi^2$ and its derivatives at $a = a'$ are written as $\chi_0^2$, $\partial\chi_0^2/\partial a_j$, and so forth:

$$\chi^2 = \chi_0^2 + \sum_{j=1}^{m}\left[\frac{\partial\chi_0^2}{\partial a_j}(a_j - a_j')\right] + \frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{m}\left[\frac{\partial^2\chi_0^2}{\partial a_k \partial a_j}(a_k - a_k')(a_j - a_j')\right] \tag{8.9}$$

Because the condition for minimizing $\chi^2$ is that the first partial derivative with respect to each parameter vanish (i.e., $\partial\chi^2/\partial a_j = 0$), we can expect that near a local minimum in any parameter $a_j$, $\chi^2$ will be a quadratic function of that parameter.

We can obtain another useful relation from Equation (8.8) by taking the second derivative of $\chi^2$ with respect to the parameter $a_j$ to obtain

$$\frac{\partial^2\chi^2}{\partial a_j^2} = \frac{2}{\sigma_j^2} \tag{8.10}$$

We obtain the following expression for the uncertainty in the parameter in terms of the curvature of the $\chi^2$ function in the region of the minimum:

$$\sigma_j^2 = 2\left(\frac{\partial^2\chi^2}{\partial a_j^2}\right)^{-1} \tag{8.11}$$



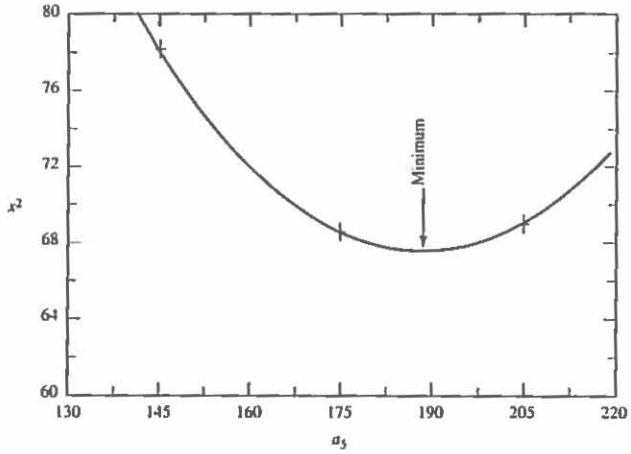


**FIGURE 8.3**
Plot of $\chi^2$ versus a single parameter $a$ in the region of a local minimum. The location of the minimum is calculated by fitting a parabola through the three indicated data points.

We note that for uncorrelated parameters, Equation (8.11) is equivalent to Equation (7.22) with Equation (7.25) for obtaining the uncertainties from the curvature matrix.

We can also use the quadratic relation to find the approximate location of a $\chi^2$ minimum by considering the equation of a parabola that passes through three points that straddle the minimum, and solving for the value of the parameter at the minimum, as illustrated in Figure 8.3. If we have calculated three values of $\chi^2$, $\chi_1^2 = \chi^2(a_{j1})$, $\chi_2^2 = \chi^2(a_{j2})$, and $\chi_3^2 = \chi^2(a_{j3})$, where $a_{j2} = a_{j1} + \Delta a_j$ and $a_{j3} = a_{j2} + \Delta a_j$, then the value $a_j'$ of the parameter at the minimum of the parabola is given by

$$a_j' = a_{j3} - \Delta a_j\left[\frac{\chi_3^2 - \chi_2^2}{\chi_1^2 - 2\chi_2^2 + \chi_3^2} + \frac{1}{2}\right] \tag{8.12}$$

In addition, we can estimate the errors in the fitting parameters $a_j$ by varying each parameter about its minimum to increase $\chi^2$ by 1 from the minimum value. The variation $\sigma_j$ in the parameter $a_j$, which will increase $\chi^2$ by 1 from its value at the minimum of the parabola, is given by

$$\sigma_j = \Delta a_j\sqrt{2(\chi_1^2 - 2\chi_2^2 + \chi_3^2)^{-1}} \tag{8.13}$$

Alternatively, we can attempt to calculate the second derivative of $\chi^2$ at the minimum and find the standard deviation from Equation (8.11).

If the parameters are correlated, the method summarized in Equation (8.13) for determining uncertainties in the parameters is valid only under the condition that, with $a_j = a_j' \pm \sigma_j$, $\chi^2$ be minimized with respect to all other parameters. This condition severely limits the usefulness of this procedure for determining the uncertainties.

we choose initial step sizes to be proportional to the starting values of the parameters and readjust them if necessary after each local minimum is found. In the simple grid-search calculation, we adjust the step sizes to be those values that increase $\chi^2$ by approximately 2 from its value at the local minimum.

### Condition for Convergence

A change in $\chi^2$ per degree of freedom ($\chi^2$/dof) of less than about 1% from one trial set of parameters to the next is probably not significant. However, because of the problems of local minima and very flat valleys in the parameter space, it may not be sufficient to set an arbitrary condition for convergence, start a search, and let it run to completion. If the starting parameters are not chosen very carefully, the search may stop in a flat valley with an inappropriately large value of $\chi^2$. If this happens, there are several possible ways to proceed. We can choose different starting values and retry the fit, as suggested in the previous sections, or we can set tighter convergence requirements (e.g., $\Delta\chi^2/\text{dof} < 0.1\%$) and rerun the search in the hope that the program will escape from the valley and reach the appropriate minimum. A convenient approach for small problems is to observe the process of the search and to cut it off manually when it appears that a stable minimum has been found. If a suitable minimum cannot be found, then different starting values should be tried. When fitting curves to several similar samples of data, we may find it satisfactory to establish suitable starting parameters, step sizes, and a cutoff criterion for the first set, and employ an automatic method for the remaining sets.

### Computer Illustration of Nonlinear Fitting Methods

In the following sections we discuss and illustrate with computer routines four methods of fitting Equation (8.2) to the data of Example 8.1.

**Program 8.0.** NONLINFT (Appendix E) Common calling routine to test the four different fitting methods. Repeats the calculations until a $\chi^2$-minimum is found. Variables are defined in the program until FITVARS and data input and output are handled in the program unit FITUTIL as in the fitting programs of Chapters 6 and 7. FITFUNC8 calculates the fitting function.

Step sizes for the fit are set initially in the routine FETCHPARAMETERS to be a fraction of the starting values of the parameters. (The step sizes must not be scaled to the parameters throughout the calculation, however, lest they become 0 when a parameter is 0, which would halt the search in that parameter.)

Tables 8.2, 8.3, 8.4, and 8.5 show values of $\chi^2$ and the parameters $a_1$ through $a_5$ for several stages of the calculation at the beginning, middle, and end of each of the four types of search. The tables include the time to find the solution relative to the time for the fastest procedure.

**Program 8.1.** GRIDSEAR (Appendix E) Routine GRIDLS illustrates the grid-search method.

**Program 8.2.** GRADSEAR (Appendix E) Routine GRADLS illustrates the gradient-search method.



**Program 8.3.** EXPNDFIT (Appendix E) Routine CHIFIT illustrates fitting by expansion of the fitting function.

**Program 8.4.** MARQFIT (Appendix E) Routine MARQUARDT illustrates fitting by the gradient-expansion algorithm.

**Program 8.5.** FITFUN8 (Appendix E) Fitting function and $\chi^2$-calculation for all fits called from Program 8.0.

**Program 8.6.** MAKEAB8 (Appendix E) Matrix set-up for non-linear fits.

**Program 8.7.** NUMDERIV (Website) Numerical derivatives.

**Program B.1.** MATRIX (Appendix E) Matrix products and inversion.

## 8.3 GRID-SEARCH METHOD

If the variation of $\chi^2$ with each parameter $a_j$ is not very sensitive to the values of the other parameters, then the optimum parameter values can be obtained most simply by minimizing $\chi^2$ with respect to each of the parameters separately. This is the *grid-search* method. The procedure is simply to select starting values of the parameters, find the value of one of the parameters that minimizes $\chi^2$ with respect to that parameter, set the parameter to that value, and repeat the procedure for each parameter in turn. The entire process is then repeated until a stable $\chi^2$ minimum is obtained.

**Grid search.** The procedure for a grid search may be summarized as follows:

1. Select starting values $a_j$ and step or increment sizes $\Delta a_j$ for each parameter and calculate $\chi^2$ with the starting parameters.
2. Increment one parameter $a_j$ by $\pm\Delta a_j$ and calculate $\chi^2$, where the sign is chosen so that $\chi^2$ decreases.
3. Repeat step 2 until $\chi^2$ stops decreasing and begins to increase. The increase in $\chi^2$ indicates that the search has crossed a ravine and started up the other side.
4. Use the last three values of $a_j$ (which bracket the minimum) and the associated values of $\chi^2$ to determine the minimum of the parabola, which passes through the three points as illustrated in Figure 8.3. [See Equation (8.12).]
5. Repeat to minimize $\chi^2$ with respect to each parameter in turn.
6. Continue to repeat the procedure until the last iteration yields a predefined negligibly small decrease in $\chi^2$.

The main advantage of the grid-search method is its simplicity. With successive iterations of the search, the absolute minimum of the $\chi^2$ function in parameter space can be located to any desired precision.

The main disadvantage is that, if the variations of $\chi^2$ with the parameters are strongly correlated, then the approach to the minimum may be very slow. Consider, for example, the contour plot of $\chi^2$ as a function of two parameters in Figure 8.4. The $\chi^2$ contours are generally approximately elliptical near the minimum. The degree of correlation of the parameters is indicated by the tilt of the ellipse. If two

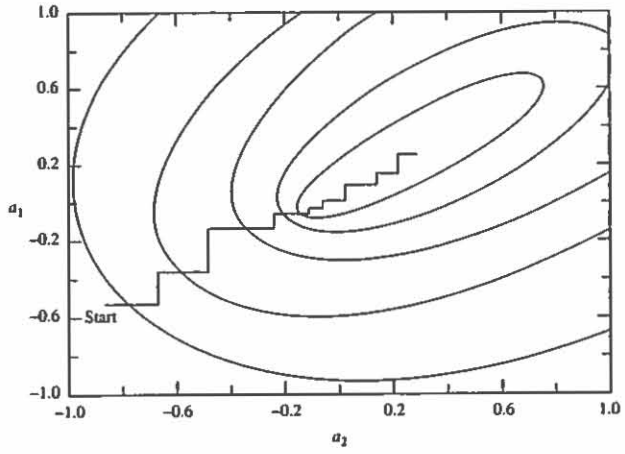


FIGURE 8.4
Contour plot of $\chi^2$ as a function of two highly correlated variables. The zigzag line represents the search path approach to a local minimum by the grid-search method.

parameters are not correlated, so that the variation of $\chi^2$ with each parameter is independent of the variation with the other, then the axes of the ellipse will be parallel to the coordinate axes. Thus, if a grid search is initiated near one end of a tilted ellipse, the search may follow a zigzag path as indicated by the solid line in Figure 8.4 and the search will be very inefficient. Nevertheless, the simplicity of the calculations involved in a grid search often compensates for this inefficiency.

**Program 8.1.** GRIDSEAR (Appendix E) Routine GRIDLS illustrates the grid-search method.

The main search routine, GRIDLS, is entered with the value of $\chi^2$ (CHISQR) as argument. In a loop over each of the $m$ parameters in turn, the value of the parameter is varied until $\chi^2$ has passed through a local minimum in the parameter. The three most recent values of $\chi^2$ that bracket the minimum are stored in the variables CHISQ1, CHISQ2, and CHISQ3. The best estimate of the parameter at this stage of the calculation is determined from the minimum of the parabola that passes through the three points. The step size (DELTAA(J)) is then adjusted to be that value that increases $\chi^2$ by 2 from its value at the local minimum.

One pass through GRIDLS corresponds to a single zigzag along the path of Figure 8.4. The search is repeated until $\chi^2$ does not change by more than the preset level, CHICUT.

A call to the function SIGPARAB in the program unit FITUTIL at the end of the search returns an estimate of the uncertainty in each parameters in turn from a calculation of the independent variation needed to increase $\chi^2$ by 1 from its minimum value.



TABLE 8.2
**Two exponentials plus constant background: grid-search method**

| Trial | $\chi^2$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|---|
| 0 | 406.6 | 10.0 | 900.0 | 80.0 | 27.0 | 225.0 |
| 1 | 143.0 | 14.5 | 1332.3 | 106.8 | 27.7 | 207.2 |
| 2 | 96.9 | 12.6 | 1233.9 | 127.9 | 28.2 | 198.4 |
| 3 | 79.4 | 11.6 | 1155.1 | 140.2 | 28.8 | 192.2 |
| 4 | 72.9 | 11.2 | 1100.3 | 147.0 | 29.3 | 189.2 |
| ⋮ | | | | | | |
| 16 | 66.7 | 11.3 | 963.5 | 148.8 | 32.3 | 185.3 |
| 17 | 66.7 | 11.3 | 962.5 | 148.2 | 32.4 | 185.8 |
| ⋮ | | | | | | |
| 39 | 66.3 | 10.9 | 959.3 | 139.1 | 33.3 | 195.4 |
| 40 | 66.2 | 10.8 | 959.2 | 138.9 | 33.3 | 195.7 |
| Uncertainties | | 0.6 | 28.3 | 4.5 | 0.8 | 5.0 |

$\chi^2$/dof = 1.23; probability = 12.1%; relative time = 9.1

*Note:* Stages in the fit to counts from the decay of excited states of silver. The values of $\chi^2$ and the parameters are listed at the beginning, middle, and end of the search. The uncertainties in the parameters correspond to a change of 1 in $\chi^2$ from its value at the end of the search.

Table 8.2 shows values of $\chi^2$ and the parameters $a_1$ through $a_5$ for several stages of the calculation at the beginning, middle, and end of the search. The search is relatively slow, but eventually a satisfactory solution is found. Note that the calculated uncertainties correspond to the diagonal terms in the error matrix for uncorrelated parameters. If correlations are considered to be important, the matrix inversion methods discussed in the following sections could be used to find better approximations to the uncertainties.

## 8.4 GRADIENT-SEARCH METHOD

The search could be improved if the zigzagging direction of travel in Figure 8.4 were replaced by a more direct vector toward the appropriate minimum. In the *gradient-search* method of least squares, all the parameters $a_j$ are incremented simultaneously, with relative magnitudes adjusted so that the resultant direction of travel in parameter space is along the gradient (or direction of maximum variation) of $\chi^2$.

The gradient $\nabla\chi^2$ is a vector that points in the direction in which $\chi^2$ increases most rapidly and has components in parameter space equal to the rate of change of $\chi^2$ along each axis:

$$\nabla\chi^2 = \sum_{j=1}^{n} \left[ \frac{\partial \chi^2}{\partial a_j} \hat{a}_j \right] \tag{8.14}$$

where $\hat{a}_j$ indicates a unit vector in the direction of the $a_j$ coordinate axis. In order to determine the gradient, we estimate the partial derivatives numerically as discussed in Appendix A:

$$(\nabla\chi^2)_j = \frac{\partial\chi^2}{\partial a_j} \simeq \frac{\chi^2(a_j + f\Delta a_j) - \chi^2(a_j)}{f\Delta a_j} \tag{8.15}$$

where $f$ is a fraction of the step size $\Delta a_j$ by which $a_j$ is changed in order to determine the derivative.

The gradient has both magnitude and dimensions and, if the dimensions of the various parameters $a_j$ are not all the same (which is usually the case), the components of the gradient do not even have the same dimensions. Let us define dimensionless parameters $b_j$ by rescaling each of the parameters $a_j$ to a size that characterizes the variation of $\chi^2$ with $a_j$ rather roughly. We shall use the step sizes $\Delta a_j$ as the scaling constants, so that

$$b_j = \frac{a_j}{\Delta a_j} \tag{8.16}$$

The derivative with respect to $b_j$ then becomes

$$\frac{\partial\chi^2}{\partial b_j} = \frac{\partial\chi^2}{\partial a_j}\Delta a_j \tag{8.17}$$

which may be calculated numerically as

$$\frac{\partial\chi^2}{\partial b_j} \simeq \frac{\chi^2(a_j + f\Delta a_j) - \chi^2(a_j)}{f\Delta a_j}\Delta a_j = \frac{\chi^2(a_j + f\Delta a_j) - \chi^2(a_j)}{f} \tag{8.18}$$

We can then define a dimensionless gradient $\gamma$, with unit magnitude and components

$$\gamma_j = \frac{\partial\chi^2/\partial b_j}{\sqrt{\sum_{j=1}^{m}(\partial\chi^2/\partial b_j)^2}} \tag{8.19}$$

In the numerical calculation of Equation (8.18), the quantities $\Delta a_j$ and $f$ occur only in the argument of $\chi^2$ and not as scale factors.

The direction that the gradient-search method follows is the *direction of steepest descent*, which is opposite of the gradient $\gamma$. The search begins by incrementing all parameters simultaneously by an amount $\Delta a_j$, with relative value given by the corresponding component $\gamma_j$ of the dimensionless gradient and absolute magnitude given by the size constant $\Delta a_j$:

$$\delta a_j = -\gamma_j \Delta a_j \tag{8.20}$$

The minus sign ensures that the value of $\chi^2$ decreases. The size constant $\Delta a_j$ of Equation (8.20) is the same as that of Equation (8.16).

There are several possible methods of continuing the gradient search after a first step. The most straightforward is to recompute the gradient after each change in the parameters. One disadvantage of this method is that it is difficult to approach the bottom of the minimum asymptotically because the gradient tends to 0 at the minimum. Another disadvantage is that recomputation of the gradient at each step for small step sizes results in an inefficient search, but the use of larger step sizes makes location of the minimum less precise.

A reasonable variation on the method is to search along one direction of the original gradient in small steps, calculating only the value of $\chi^2$ until $\chi^2$ begins to rise again. At this point, the gradient is recomputed and the search continues in the new direction. Whenever the search straddles a minimum, a parabolic interpretation of $\chi^2$ is used to improve the determination of the minimum.



A more sophisticated approach would be to use second partial derivatives of $\chi^2$ to determine changes in the gradient along the search path:

$$\left.\frac{\partial\chi^2}{\partial a_j}\right|_{a_i+\delta a_i} \simeq \left.\frac{\partial\chi^2}{\partial a_j}\right|_{a_i} + \sum_{k=1}^{m}\left(\frac{\partial^2\chi^2}{\partial a_j \partial a_k}\delta a_k\right) \tag{8.21}$$

If the search is already fairly near the minimum, this method does decrease the number of steps needed, but at the expense of more elaborate computation. If the search is not near enough to the minimum, this method can actually increase the number of steps required when first-order perturbations on the gradient are not valid.

The efficiency of the gradient search decreases markedly as the search approaches a minimum because the evaluation of the derivative according to the method of Equation (8.18) involves taking differences between nearly equal numbers. In fact, at the minimum of $\chi^2$, these differences should vanish. For this reason, one of the methods discussed in the following sections may be used to locate the actual minimum once the gradient search has approached it fairly closely.

**Program 8.2.** GRADSEAR (Appendix E) Routine GRADLS illustrates the gradient-search method.

On each entry to the main search routine, GRADLS, the components of the gradient GRADLS(J) are calculated numerically from Equation (8.18) in the procedure CALCGRAD. The argument FRACT of this routine, corresponding to the variable $f$ of Equation (8.18), determines the fraction of the step size (DELTAA) used in the numerical calculation of the partial derivative. Each parameter A(J) is then changed by the amount STEPDOWN*DELTAA(J)*GRAD(J), where STEPDOWN is a scaling factor that is set initially in the main program and readjusted after each stage to the size needed to locate the minimum.

The initial values of DELTAA(J) determines to some extent the execution speed of each pass through the routine GRADLS, and the value of CHICUT determines when the search will stop. Because of the small gradient near the $\chi^2$ minimum, it may take many steps to reach a reasonable value of $\chi^2$, and the cutoff, CHICUT, may have to be set to a very low value. For such cases, user intervention can be provided as an alternate method of stopping the search.

At the conclusion of the search, the uncertainties in the parameters are estimated in the function SIGPARAB as in the routine GRADLS.

Table 8.3 shows values of $\chi^2$ and the parameters $a_1$ through $a_5$ for several stages of the calculation at the beginning, middle, and end of the search. For Example 8.1, the gradient search is considerably faster than the grid-search approach because all the parameters are varied together at each step. However, the gradient-search method has one disadvantage that is not illustrated. If the starting values of the parameters are too far from the final values, the grid search has a good chance of plodding along until it reaches the correct solution. The gradient search, on the other hand, may tend to get bogged down in local minima that correspond to a long, flat valley in the parameter space.

TABLE 8.3
**Two exponentials plus constant background: gradient-search method**

| Trial | $\chi^2$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|---|
| 0 | 406.6 | 10.0 | 900.0 | 80.0 | 27.0 | 225.0 |
| 1 | 82.3 | 10.6 | 1061.0 | 94.0 | 34.4 | 254.2 |
| 2 | 72.6 | 9.8 | 984.0 | 98.8 | 36.8 | 237.4 |
| 3 | 69.8 | 9.9 | 966.9 | 100.9 | 36.8 | 244.6 |
| 4 | 69.3 | 9.8 | 953.7 | 101.6 | 36.7 | 242.1 |
| ⋮ | | | | | | |
| 19 | 66.6 | 8.9 | 952.2 | 114.7 | 35.5 | 233.6 |
| 20 | 66.5 | 8.9 | 954.8 | 114.9 | 35.6 | 233.9 |
| Uncertainties | | 0.6 | 26.5 | 3.8 | 0.8 | 7.0 |

$\chi^2$/dof = 1.23; probability = 11.8%; relative time = 4.0

*Note:* Stages in the fit to counts from the decay of excited states of silver. The values of $\chi^2$ and the parameters are listed at the beginning, middle, and end of the search. The uncertainties in the parameters corresponding to a change of 1 in $\chi^2$ from its value at the end of the search.

## 8.5 EXPANSION METHODS

Instead of searching the $\chi^2$ hypersurface to map the variation of $\chi^2$ with parameters, we should be able to find an approximate analytical function that describes the $\chi^2$ hypersurface and use this function to locate the minimum, with methods developed for linear least-squares fitting. The approximations will introduce errors into the calculated values of the parameters, but successive iterations of the analytical method should approach the $\chi^2$ minimum with increasing accuracy. The main advantage of such an approach is that the number of points on the $\chi^2$ hypersurface at which computations must be made will be fewer than for a grid or gradient search. This advantage is somewhat offset by the fact that the computations at each point are considerably more complicated. However, the analytical solution essentially chooses its own step size and, thus, the user is spared the problem of trying to optimize the step size for speed and precision.

### Parabolic Expansion of $\chi^2$

In Equation (8.9) we expanded $\chi^2$ to second order in the parameters about a local minimum $\chi_0^2$ where $a_j = a_j'$:

$$\chi^2 \simeq \chi_0^2 + \sum_{j=1}^{m}\left[\frac{\partial \chi_0^2}{\partial a_j}\delta a_j\right] + \frac{1}{2}\sum_{k=1}^{m}\sum_{j=1}^{m}\left[\frac{\partial^2 \chi_0^2}{\partial a_j \partial a_k}\delta a_j \delta a_k\right] \tag{8.22}$$

which is equivalent to approximating the $\chi^2$ hypersurface by a parabolic surface. Here we define $\delta a_j \equiv a_j - a_j'$, and $\chi_0^2$ is given by

$$\chi_0^2 = \sum\left\{\frac{1}{\sigma_i^2}[y_i - y'(x_i)]^2\right\} \tag{8.23}$$

where $y'(x_i)$ is the value of the function when $\delta a_j = 0$.



Applying the method of least squares, we minimize $\chi^2$ as expressed in Equation (8.22) with respect to the *increments* ($\delta a_j$) in the parameters, and solve for the optimum values of these increments to obtain

$$\frac{\partial \chi^2}{\partial(\delta a_k)} = \frac{\partial \chi_0^2}{\partial a_k} + \sum_{j=1}^{m}\left[\frac{\partial^2 \chi_0^2}{\partial a_k \partial a_j}\delta a_j\right] = 0 \quad k = 1, m \tag{8.24}$$

The result is a set of $m$ linear equations in $\delta a_j$ that we can write as

$$\beta_k - \sum_{j=1}^{m}(\delta a_j \alpha_{jk}) = 0 \qquad k = 1, m \tag{8.25}$$

with

$$\beta_k \equiv -\frac{1}{2}\frac{\partial \chi_0^2}{\partial a_k} \quad \text{and} \quad \alpha_{jk} \equiv \frac{1}{2}\frac{\partial^2 \chi_0^2}{\partial a_j \partial a_k} \tag{8.26}$$

The factors $\pm\frac{1}{2}$ are included for agreement with the conventional definitions of these quantities.

As in Chapter 7, we can treat Equation (8.25) as a matrix equation:

$$\boldsymbol{\beta} = \delta \mathbf{a}\, \boldsymbol{\alpha} \tag{8.27}$$

where $\boldsymbol{\beta}$ and $\delta\mathbf{a}$ are row matrices and $\boldsymbol{\alpha}$ is a symmetric matrix of order $m$. We shall find that $\boldsymbol{\alpha}$ is the *curvature matrix* discussed in Section 7.2, so named because it measures the curvature of the $\chi^2$ hypersurface.

### Method of Computation

The solution of Equation (8.27) can be obtained by matrix inversion as in Section 7.2:

$$\delta \mathbf{a} = \boldsymbol{\beta}\boldsymbol{\epsilon} \qquad \delta a_k = \sum_{j=1}^{m}(\epsilon_{kj}\beta_j) \tag{8.28}$$

where the error matrix $\boldsymbol{\epsilon} = \boldsymbol{\alpha}^{-1}$ is the inverse of the curvature matrix.

If the parameters are independent of one another, that is, if the variation of $\chi^2$ with respect to each parameter is independent of the values of the other parameters, then the cross-partial derivatives $a_{jk}$ ($j \neq k$) will be 0 in the limit of a very large data sample and the matrix $\boldsymbol{\alpha}$ will be diagonal. The inverse matrix $\boldsymbol{\epsilon}$ will also be diagonal and Equation (8.27) will degenerate into $m$ separate equations:

$$\delta\alpha_j \simeq \frac{\beta_j}{\alpha_{jj}} = \frac{\partial x_0^2}{\partial a_j} \div \frac{\partial^2 x_0^2}{\partial a_j^2} \tag{8.29}$$

Computation of the matrix elements by Equation (8.26) requires knowledge of the first and second derivatives of $\chi^2$ evaluated at the current values of the parameters. Analytic forms of the derivatives are generally quickest to compute, but may be difficult or cumbersome to derive. If it is not convenient or possible to provide analytic forms of the derivatives, then they can be computed by the method of finite differences (see Appendix A). In the following expressions, we use forward differences

### Gradient-Expansion Algorithm

A convenient algorithm (see Marquardt 1963), which combines the best features of the gradient search with the method of linearizing the fitting function, can be obtained by increasing the diagonal terms of the curvature matrix $\alpha$ by a factor $1 + \lambda$ that controls the interpolation of the algorithm between the two extremes. Equation (8.34) becomes

$$\beta = \delta a\, \alpha' \quad \text{with} \quad \alpha'_{jk} = \begin{cases} \alpha_{jk}(1+\lambda) & \text{for } j = k \\ \alpha_{jk} & \text{for } j \neq k \end{cases} \tag{8.39}$$

If $\lambda$ is very small, Equations (8.39) are similar to the solution of Equation (8.34) developed from the Taylor expansion. If $\lambda$ is very large, the diagonal terms of the curvature matrix dominate and the matrix equation degenerates into $m$ separate equations

$$\beta_j \simeq \lambda \delta a_j \alpha_{jj} \tag{8.40}$$

which yield the vector increment $\delta a$ in the same direction as the vector $\beta$ of Equation (8.37) (or opposite to the gradient of $\chi^2$).

The solution for the parameter increments $\delta a_j$ follows from Equations (8.39) after matrix inversion

$$\delta a_j = \sum_{k=1}^{m} (\beta_k \epsilon'_{jk}) \tag{8.41}$$

where the $\beta_k$ are given by Equation (8.37) and the matrix $\epsilon'$ is the inverse of the matrix $\alpha'$ with elements given by Equations (8.39).

The initial value of the constant factor $\lambda$ should be chosen small enough to take advantage of the analytical solution, but large enough that $\chi^2$ decreases. Because this algorithm approaches the gradient-search method with small steps for large $\lambda$, there should exist a value of $\lambda$ such that $\chi^2(a + \delta a) < \chi^2(a)$. The recipe given by Marquardt is:

1. Compute $\chi^2(a)$.
2. Start initially with $\lambda = 0.001$.
3. Compute $\delta a$ and $\chi^2(a + \delta a)$ with this choice of $\lambda$.
4. If $\chi^2(a + \delta a) > \chi^2(a)$, increase $\lambda$ by a factor of 10 and repeat step 3.
5. If $\chi^2(a + \delta a) < \chi^2(a)$, decrease $\lambda$ by a factor of 10, consider $a' = a + \delta a$ to be the new starting point, and return to step 3, substituting $a'$ for $a$.

For each iteration it may be necessary to recompute the parameter increments $\delta a_j$ from Equation (8.41), and the elements $\alpha_{jk}$ and $\beta_j$ of the matrices, several times to optimize $\lambda$. As the solution approaches the minimum, the value of $\lambda$ will decrease and the program should locate the minimum with a few iterations. A lower limit may be set for the value $\lambda$, but in practice this limit will seldom be reached.



TABLE 8.5
**Two exponentials plus constant background: Marquardt method**

| Trial | $\chi^2$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|---|
| 0 | 406.6 | 10.0 | 900.0 | 80.0 | 27.0 | 225.0 |
| 1 | 82.9 | 11.0 | 933.5 | 139.3 | 33.9 | 173.9 |
| 2 | 66.4 | 10.8 | 960.1 | 130.6 | 33.8 | 201.2 |
| 3 | 66.1 | 10.4 | 958.3 | 131.4 | 33.9 | 205.0 |
| Uncertainties | | 1.8 | 49.9 | 21.7 | 2.5 | 30.5 |

$\chi^2$/dof = 1.22; probability = 12.4%; relative time = 1.0

*Note:* All stages in the fit to counts from the decay of excited states of silver. The uncertainties in the parameters correspond to the square roots of the diagonal terms in the error matrix.

TABLE 8.6
**Elements of the error matrix (Marquardt method)**

| I/k | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 3.38 | −3.69 | 27.98 | −2.34 | −49.24 |
| 2 | −3.69 | 2492.26 | 81.89 | −69.21 | −3.90 |
| 3 | 27.98 | 81.89 | 468.99 | −44.22 | −615.44 |
| 4 | −2.34 | −69.21 | −44.22 | 6.39 | 53.80 |
| 5 | −49.24 | −3.90 | −615.44 | 53.80 | 929.45 |

*Note:* Error matrix from a fit to the radioactive silver data. The diagonal terms are the variances $\sigma_k^2$ and the off-diagonal terms are the covariances $\sigma_{jk}^2$ of the parameters $a_k$.

**Program 8.4.** MARQFIT (Appendix E) Routine MARQUARDT illustrates fitting by the gradient-expansion algorithm.
The procedure uses the same program units as those in Program 8.3, and is identical to that program except for the adjustment of the diagonal elements $\alpha_{jj}$ of the matrix $\alpha$ by the variable LAMBDA according to Equation (8.39).

At the conclusion of the search, the inverse $\epsilon$ of the final value of the curvature matrix $\alpha$ is treated as the error matrix, and the errors in the parameters are obtained from the square roots of the diagonal terms by calls to the function SIGMATRX in the unit FitFunc8. Table 8.5 shows values of $\chi^2$ and the parameters $a_1$ through $a_5$ for all stages of the calculation. Table 8.6 shows the error matrix from the fit.

## 8.7 COMMENTS

Although the Marquardt method is the most complex of the four fitting routines, it is also the clear winner for finding fits most directly and efficiently. It has the strong advantage of being reasonably insensitive to the starting values of the parameters, although in a peak-over-background example (Chapter 9), it does have difficulty when the starting parameters of the function for the peak are outside reasonable

with insufficient data to satisfy the requirement of Gaussian statistics for individual histogram bins and (2) experiments in which the fitting function corresponds to a different probability density function for each measured event so that binning the data leads to a reduction in information and a loss of sensitivity in determining the parameters. If the data set is sufficiently large, then the least-squares method can be applied to problems of either type, and that method is generally preferred in view of its smaller computing requirement. At any rate, it is not possible to extract more than minimal information from a very small data set, so we should expect the direct maximum-likelihood method to be most useful for intermediate problems with modest data samples.

## 10.1 INTRODUCTION TO MAXIMUM LIKELIHOOD

The basic maximum-likelihood procedure is relatively simple. Assume that we have a collection of $N$ events corresponding to the measurement of an independent variable $x_i$ and a dependent variable $y_i$, where $i$ runs from 1 to $N$. We wish to obtain the parameters, $a_1, a_2, \ldots, a_m$, of a fitting function $y(x_i) = y(x_i; a_1, a_2, \ldots, a_m)$ from these data. For each event, we convert $y(x_i)$ to a normalized probability density function

$$P_i = P(x_i; a_1, a_2, \ldots, a_m) \tag{10.1}$$

evaluated at the observed value $x_i$. The likelihood function $\mathcal{L}(a_1, a_2, \ldots, a_m)$ is the product of the individual probability densities

$$\mathcal{L}(a_1, a_2, \ldots, a_m) = \prod_{i=1}^{N} P_i \tag{10.2}$$

and the maximum-likelihood values of the parameters are obtained by maximizing $\mathcal{L}(a_1, a_2, \ldots, a_m)$ with respect to the parameters.

In many experiments, the probability density function $P_i$ will be made up of two components: a theoretical factor corresponding to the underlying principle being tested and an experimental factor corresponding to the biases introduced by experimental conditions.

**EXAMPLE 10.1** In Example 5.7 we presented a Monte Carlo program for studying biases that could arise in an experiment to measure the mean life of the short-lived $K_1^0$ meson (or kaon). The example includes details of the experiment and Figure 5.4 illustrates schematically the experimental apparatus.

In brief, the experiment involves measuring the distance between the point of production and point of decay of the kaon, determining the meson's velocity, and calculating the meson's time of flight from production to decay. After correction for bias introduced by the finite size of the experimental apparatus, the mean life of the kaon could be determined from measurements of many such events.

The dashed rectangle on Figure 5.4 indicates the region in which events are collected, the fiducial region for the experiment. We select decay vertices only within this region to assure precise measurements of both the separation of the two vertices and the trajectories of secondary particles from decay of the kaon. These latter measurements determine the momentum, and thus the velocity, of the kaon. Loss of events that do not fall within the fiducial region bias the final calculation of the mean life and therefore we must understand the biases and make corrections.

In the following examples, we assume that the coordinates of the two vertices and the magnitude of the momentum of the decaying kaon have been determined.

We used the Monte Carlo program of Example 5.7, with the mean life of the kaon set to its nominal value of $\tau_K = 0.894 \times 10^{-10}$s, to generate 40,000 events in order to study the efficiency of the detector with reasonably high precision. It is important that the statistical uncertainties introduced in the determination of the efficiency function be negligible compared to the statistical and other uncertainties in the actual experiment. The distribution of the 23,565 generated events that survived fiducial cuts is shown as crosses in Figure 10.1 with the expected exponential distribution of the total 40,000-event sample shown as a smooth curve.

In Figure 10.2 we have plotted the resulting efficiency as a function of the times of flight of the kaons (the *proper time*) in their individual rest frames, with the efficiency function defined as the ratio of observed to expected events [or the point-by-point ratio $\epsilon(T) = N'(T)/N(T)$] from Equation (5.31). The dotted line in Figure 10.2 illustrates the region over which the efficiency reasonably may be assumed to be 100%.



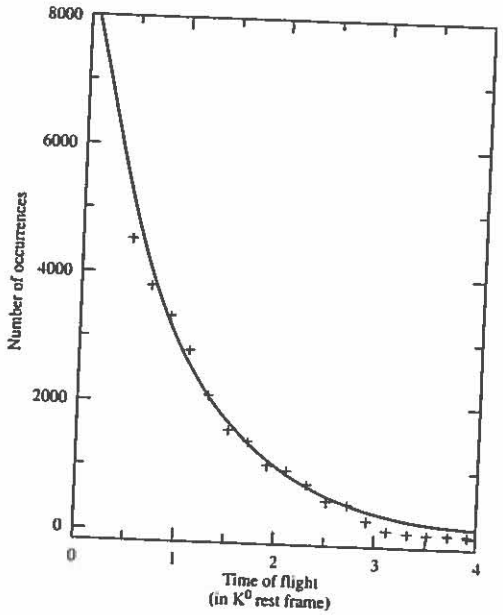


**FIGURE 10.1**
Frequency distribution of times of flight for 23,565 events that survived fiducial cuts in a 40,000-event Monte Carlo generation, as a function of the proper time (in units of $10^{-10}$ s). The exponential curve was calculated from the nominal value $\tau_K = 0.894 \times 10^{-10}$ s to represent the expected distribution of the 40,000 generated events.

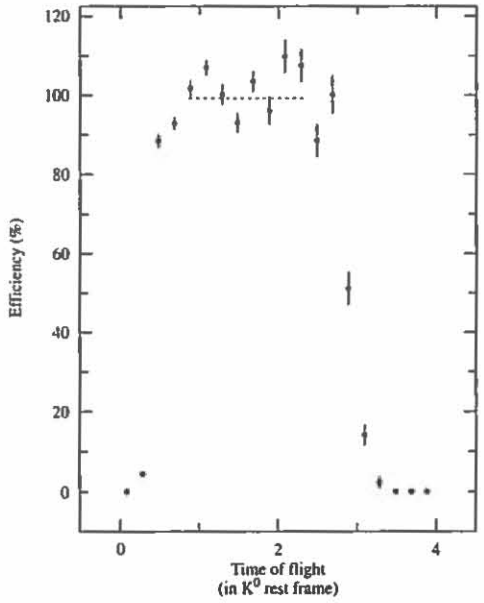


**FIGURE 10.2**
Efficiency function $\epsilon(T) = N(T)/N(T)$, calculated from the ratio of observed events (crosses) to expected events (smooth curve in Figure 10.1). The dotted line illustrates the region over which the efficiency reasonably may be assumed to be 100%.

We also used the Monte Carlo program, with different random-number seeds and the same nominal value of $\tau_K$, to generate a small "data set" of 1000 events, of which 598 survived the fiducial cut, to use in testing our analysis procedures.

We shall discuss several aspects of the analysis of such data in the following examples.

**EXAMPLE 10.1a: Least-squares Method** Figure 10.3 shows on a semilogarithmic plot the distribution, as crosses (x), of the 598 events that survived the fiducial cuts from the total sample of 1000 events generated in Example 10.1. The straight line shows the expected distribution if there had been no efficiency losses. In order to extract the mean life of the kaon from these data, we apply the efficiency function illustrated in Figure 10.2 to correct for losses. The corrected data points are plotted in Figure 10.3 as data points with vertical error bars corresponding to the statistical uncertainties in the data, scaled by the efficiency factor. (Uncertainties in the correction factor were negligible.) The efficiency was assumed to be 100% in the region indicated by the horizontal dotted line in Figure 10.2. The very large error bars on "corrected" points at the two ends of the plot result from scaling low-statistics data points and illustrate the problem of using data in regions of low efficiency. Generally, it is



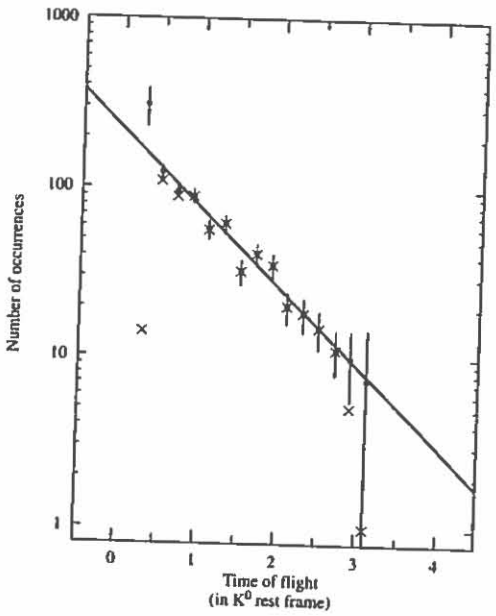


**FIGURE 10.3**
Semilogarithmic plot of the frequency distribution of 598 events that survived fiducial cuts from a 1000-event (Monte Carlo) data sample. The uncorrected data are shown as crosses; the data corrected for efficiency losses are shown as data points with error bars. The straight line shows the result of a linear least-squares fit to the corrected semilogarithmic data.

wise to eliminate points that require such large corrections from the sample, because they contribute little to the overall result and depend heavily on the corrections.

From the linear slope of the logarithmic plot, illustrated by the straight line through the data points, we obtain an "experimental" mean life $\tau = (0.925 \pm 0.058)$. Alternatively, we could have used a nonlinear least-squares fitting technique to determine $\tau$ directly from a linear plot of the data.

## Direct Maximum Likelihood

Most actual experiments are more complex and have efficiency functions that are considerably more complicated than the one illustrated by our example. For such problems, application of direct maximum likelihood may be the preferable method for finding the best estimate of the parameters. To apply this method, we must define a probability function for each recorded event.

The probability of observing a single event that survives for a time $t_i$ is

$$P_i = A_i p(t_i; \tau) \tag{10.3}$$

The first factor $A_i$ represents the *detection efficiency*, or probability that the particle will decay within a predefined *fiducial volume* within our apparatus, so that a satisfactory measurement can be made of its flight time. This factor depends upon the coordinates of the production and decay vertices of the decaying particle, its momentum vector, and the geometry of the fiducial volume. The second factor $p(t_i; \tau)$ is proportional to the probability that a particle of mean lifetime $\tau$ will decay between time $t_i$ and $t_i + dt$ and is therefore proportional to $e^{-t_i/\tau}$. Equation (10.3) becomes

$$P_i = A_i e^{-t_i/\tau} \tag{10.4}$$

It might appear that the two factors in Equation (10.3) are independent, so that the detection efficiency factor is independent of the decay probability, but, as we have observed in the previous example, this is not generally true. Because of the finite size of our measuring apparatus, we may preferentially lose events that survive for very short times so that we can't make precise measurements of their flight paths, as well as those that survive for very long times and therefore decay outside the acceptable limits of our detectors. Losses of both types depend upon the mean life that we are attempting to determine, the "$\tau$" in the second factor of Equation (10.3). For each particle that is observed to decay within the apparatus, we can define a *potential path length* as the distance it would travel if it had not decayed. Because each decaying particle has a different potential path length, we must calculate geometric factors to correct for those particles that decay outside the detector. The correction factors will depend on the parameters and will be a function of the production and decay coordinates and the momentum vectors of each decaying particle. Clearly, one element of good experiment design should be to minimize the dependence of these geometric correction factors on the parameters sought in the experiment.

## Normalization for Maximum Likelihood

The factor $A_i$ in Equation (10.4) corresponds to a normalization for each measurement to assure unit probability for observing in this experiment *any* event that has the mean life, coordinates, and kinematics of the observed decaying particle. To determine the normalizing factor $A_i$ we refer to Figure 5.4 and consider the fiducial volume of our apparatus, indicated by the dashed rectangle. From each particle's production coordinates and momentum vector, we can determine the minimum distance $d_1$ that the particle must travel to enter the region and the maximum distance $d_2$ it can travel before leaving the region. (We can, of course, observe some events outside the fiducial volume, but we reject them because they cannot be measured precisely.) These minimum and maximum distances $d_1$ and $d_2$ must be converted to times of flight $t_1$ and $t_2$ in the rest frame of the decaying particles, and the normalizing factors $A_i$ can then be determined from the condition

$$\int_{t_1}^{t_2} P_i dt_i = A_i \int_{t_1}^{t_2} e^{-t_i/\tau} dt_i = 1 \tag{10.5}$$

With this normalization, the individual event probability $P_i$ of Equation (10.4) becomes the probability density for observing a single event. The normalized joint



probability or the likelihood function for observing $N$ such events in our experiment is just the product of the individual probability functions:

$$\mathcal{L}(\tau) = \prod_{i=1}^{N} P_i = \prod_{i=1}^{N} A_i e^{-t_i/\tau} \tag{10.6}$$

## Parameter Search

Our object is to find the value of the parameter $\tau$ that maximizes this likelihood function. Because the probability of observing any particular event is less than 1, the product of a large number of such probabilities (one for each measured event) may be a very small number, and may, in fact, be too small for the computer to handle. To avoid problems, it is usually preferable to maximize the logarithm of the likelihood function

$$M = \ln \mathcal{L} \tag{10.7}$$

rather than the likelihood function itself, so that the product of Equation (10.6) becomes a sum. The logarithms should be reasonable, negative numbers. For our particular example, the logarithm of the likelihood function of Equation (10.6) is given by

$$M(\tau) = \ln[\mathcal{L}(\tau)] = \sum \left[ \ln A_i - \frac{t_i}{\tau} \right] \tag{10.8}$$

with $A_i$ defined by Equation (10.5). Note that $A_i$ is a function of the unknown parameter $\tau$, as well as of the production coordinates, momentum vector, and fiducial volume, and must be calculated separately for each event, *and for every trial value of* $\tau$.

In general, this problem, like the corresponding nonlinear least-squares fitting problem, cannot be solved in closed form. However, either the grid- or gradient-search method of minimizing the $\chi^2$ function discussed in Chapter 8 can be adopted directly. It is only necessary to search for a maximum of $M$ (or a minimum value of $-M$) with the same routines we used in Chapter 8 to find a minimum of $\chi^2$.

We may note a correspondence between the quantity $M(\tau)$, determined in Equation (10.7) from the likelihood function for *individual events*, and the goodness-of-fit parameter $\chi^2$, determined by Equation (8.7) from the likelihood function $P(a)$ for *binned data:*

$$\chi^2 = -2\ln[\mathcal{L}(\tau)] + \text{constant} \tag{10.9}$$

In the limit of a large number of events, the two methods must yield the same value $\tau'$ for the maximum-likelihood estimate of the parameter $\tau$. In both cases the likelihood function will be a Gaussian function of the parameter near the optimum value

$$\mathcal{L}(\tau) \propto \exp\left(-\frac{(\tau - \tau')^2}{2\sigma^2}\right) \tag{10.10}$$

so we can expect $M(\tau)$, like $\chi^2(\tau)$, to vary quadratically with the parameter $\tau$ in the vicinity of $\tau'$.

**EXAMPLE 10.1b** Let us consider the simplest form of this problem. Assume that the unknown mean lifetime is sufficiently short so that our apparatus is large enough to include many lifetimes and, therefore, the loss of particles that decay at very long times is negligible. Let us also assume that our equipment can detect particles at very short as well as very long times. Then the limits on the normalization integral of Equation (10.5) become $t_1 = 0$ and $t_2 = \infty$ and $A_i$ is the same for every event and is given by $A_i = 1/t$. The likelihood function becomes

$$\mathcal{L}(\tau) = \prod A_i e^{-t_i/\tau} = \prod \frac{e^{-t_i/\tau}}{\tau} \tag{10.11}$$

with logarithm

$$M(\tau) = \ln[\mathcal{L}(\tau)] = -\frac{1}{\tau}\sum t_i - N \ln \tau \tag{10.12}$$

We can obtain the maximum of Equation (10.12) by taking the derivative of $M(\tau)$ with respect to $\tau$ and setting it to 0:

$$\begin{aligned} \frac{dM(t)}{d\tau} &= \frac{d}{d\tau}\left\{-\frac{1}{\tau}\sum t_i - N \ln \tau\right\} \\ &= \frac{1}{\tau^2}\sum t_i - \frac{N}{\tau} = 0 \end{aligned} \tag{10.13}$$

The solution is $\tau = \Sigma t_i/N$; that is, the maximum-likelihood estimate of the mean life is just the mean of the individual lifetime measurements. We should have reached the same result if we had found the maximum of $\mathcal{L}(t)$ from Equation (10.11).

**EXAMPLE 10.1c** Suppose that we repeat the experiment, but with poorer experimental resolution so that we cannot distinguish the decay vertex $(x_2, y_2, z_2)$ from the creation vertex $(x_1, y_1, z_1)$ unless they are separated by a distance $d_1$. For simplicity, we assume that the decaying particles are all produced with the same velocity, so that the lower cutoff distance $d_1$ translates into the same lower cutoff in time $t_1$ for all events. (In an actual experiment, of course, the decaying particles would be produced with various velocities, so that the calculated lower cutoff time $t_1$ would vary from event to event.)

For this example, the normalization integral of Equation (10.5) becomes

$$A_i \int_{t_1}^{\infty} e^{-t_i/\tau} dt_i = 1 \tag{10.14}$$

which gives

$$A_i = \frac{e^{t_1/\tau}}{\tau} \tag{10.15}$$

The likelihood function becomes

$$\mathcal{L}(\tau) = \prod_{i=1}^{N} A_i e^{-t_i/\tau} = \prod_{i=1}^{N} \frac{e^{t_1/\tau}}{\tau} e^{-t_i/\tau} = \prod_{i=1}^{N} \frac{e^{(t_1 - t_i)/\tau}}{\tau} \tag{10.16}$$



so that

$$M = \ln \mathcal{L} = \sum \frac{(t_1 - t_i)}{\tau} - \sum \ln \tau \tag{10.17}$$

Setting

$$\frac{dM(\tau)}{d\tau} = 0 \tag{10.18}$$

gives

$$\frac{d}{d\tau}\sum\left\{\frac{(t_1 - t_i)}{\tau} - \ln \tau\right\} = -\sum\left\{\frac{t_1 - t_i}{\tau^2}\right\} - \frac{N}{\tau} = 0 \tag{10.19}$$

or

$$\tau = \frac{\Sigma(t_i - t_1)}{N} = \frac{\Sigma t_i}{N} - t_1 \tag{10.20}$$

As we should expect, the lifetime $\tau$ would have been overestimated if we had neglected to take account of the cutoff at short times.

**EXAMPLE 10.1d** Let us consider a more realistic problem in which we have both short and long cutoffs on the observable path. We also assume that the unstable particles are produced at various locations within the target and with various momentum vectors p.

For this example, we must calculate the normalization integral, Equation (10.5), separately for each event with individual values for $t_1$ and $t_2$ determined from the minimum and maximum distance cutoffs, $d_1$ and $d_2$, respectively. The resulting expression for the likelihood function is

$$\mathcal{L}(\tau) = \prod_{i=1}^{N} A_i e^{-t_i/\tau} = \prod_{i=1}^{N}\left[\frac{e^{-t_i/\tau}}{\tau\left[e^{-t_1/\tau} - e^{-t_2/\tau}\right]}\right] \tag{10.21}$$

with

$$M(\tau) = \ln[\mathcal{L}(\tau)]$$

Setting to zero the derivative of $M(\tau)$ with respect to $\tau$ gives us the equation for the maximum-likelihood value of $\tau$. However, the resulting equation cannot be solved analytically for $\tau$ although it could be solved by interpolation (see Appendix A). We choose, rather, to maximize $M(\tau)$ by a one-dimensional grid-search method because search methods are more generally applicable to maximum-likelihood problems and can readily be extended to multiple parameter problems.

## 10.2 COMPUTER EXAMPLE

### Sample Maximum Likelihood Fit

We use the program MAXLIKE to select and analyze the 598 events that survived the fiducial area cuts, from the 1000-event uncorrected data sample generated in

Example 10.1a. The events were generated with $\tau_K = 0.894 \times 10^{-10}$ s and the distribution of the selected events is illustrated by the crosses in Figure 10.3.

**Program 10.1** MAXLIKE (Appendix E) A grid-search method to maximize the logarithm of the likelihood function of Equation (10.21). The routines have been written specifically for Example 10.1d.
STARTUP sets the range of the parameter TAU for the search.
FETCHDATA assigns the input data file, reads the limits of the fiducial region ($d_1$ and $d_2$), reads data for individual events.
SEARCH sets and increments TAU and calls LOGLIKE, which returns the logarithm of the likelihood function $M$. Compares each calculated value of $M$ to the preceding value. Terminates the search when $M$ stops increasing and starts to decrease, indicating that $M$ has passed through a local maximum. At termination, fits a parabola to the last three points to find a better estimate of TAU at the maximum.
LOGLIKE calls LOGPROB to find the logarithm of the probability density for each event; sums to calculate the logarithm of the likelihood function.
LOGPROB calculates the logarithm of the probability density for an event.
ERROR calculates the uncertainty SIGTAU in TAUATMIN, the maximum likelihood value of the parameter TAU, by finding the change in TAU needed to decrease $M$ by $\Delta M = 1/2$.
PLOTLIKECURVE (Not listed) calculates and plots the shape of the likelihood function in the region of the maximum. Plots a Gaussian curve with mean and standard deviation equal to TAUMIN and DTAU.

## Grid-Search Solution

At each step the program increments $\tau$ by a preset amount $\Delta\tau$ and repeats the calculation until $M(\tau)$ has passed through a maximum and has started to decrease. The program fits a parabola to the three points that bracket the maximum to find the value $\tau'$ at the maximum of $M(\tau)$. For a more detailed problem, the program could be written to repeat the calculation with smaller values of $\Delta\tau$ to find a better estimate of $\tau'$, as in the fitting examples in Chapter 8. Either the grid- or gradient-search method of Chapter 8 could be adapted to solve multiparameter problems.

## Results of the Fit

We analyzed the data set twice: first with data selected in the nominal fiducial region (10 cm to 40 cm), which gave $\tau' = (0.943 \pm 0.059) \times 10^{-10}$s for the 598 events that survived the cut, and then, to test the sensitivity of the calculation to our choice of fiducial region, with data selected in the less-appropriate fiducial region with $d_1 = 10$ cm and $d_2 = 20$ cm, which gave $\tau' = (0.78 \pm 0.14) \times 10^{-10}$s for the 373 events that survived this cut. Plots of the relative values of the likelihood function versus trial values of the parameter $\tau$ are shown as crosses in Figure 10.4a for the data selected in the nominal fiducial region and in Figure 10.4b for data selected in the less-appropriate fiducial region. As expected, the incorrect fiducial region clearly selects



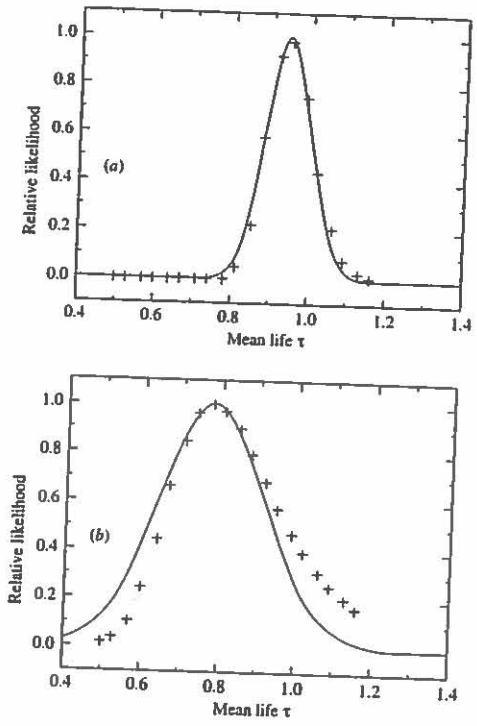


**FIGURE 10.4**
Relative values of the likelihood function versus trial values of the parameter for events that passed the fiducial cuts for the decay vertex. The data points are indicated by crosses; the smooth Gaussian curves were calculated from Equation (10.10) with the values of the means and standard deviations obtained in the two fits. (*a*) Nominal fiducial cuts: 10 – 40 cm; 598 events survived; $\tau' = 0.943 \times 10^{-10}$ s, $\sigma = 0.059 \times 10^{-10}$ s. (*b*) Incorrect fiducial cuts: 10 – 20 cm; 373 events survived; $\tau' = 0.78 \times 10^{-10}$ s, $\sigma = 0.14 \times 10^{-10}$ s.

fewer events and, therefore, gives a less-precise result. In an actual experiment, we should have to consider a trade-off between the number of surviving events in the sample, and the precision with which those surviving events could be measured, and choose our fiducial region to maximize the overall quality of the result.

We observed that, for a sufficiently large event sample, the likelihood function should become Gaussian in the parameters in the vicinity of a $\chi^2$ minimum (or a

maximum of the likelihood function) according to Equation (10.10), where $\tau'$ is the value of the parameter $\tau$ that maximizes the likelihood function. We show on Figures 10.4a and 10.4b Gaussian curves calculated from Equation (10.10), with $\tau'$ and $\sigma$ determined by the respective fits. Both the data points and the Gaussian curves have been scaled to unit height at $\tau = \tau'$. The data points of Figure 10.4a closely follow the curve; in the lower statistics example in Figure 10.4b, the data points depart from the curve considerably.

## Uncertainties

To estimate the uncertainty $\sigma$ in our determination of $\tau'$, we found the change in $\tau$ necessary to decrease $M$ by $\Delta M = 1/2$ from its value at the maximum $\tau'$ (corresponding to an increase of $\chi^2$ by 1 or a change of $e^{-1/2}$ in the likelihood function $\mathcal{L}$). Because the likelihood function for the larger sample (Figure 10.4a) closely followed the Gaussian form, our estimate of the uncertainty should be satisfactory. However, the smaller sample (Figure 10.4b) was skewed from the Gaussian, so that our estimate of the standard deviation might be somewhat low. For multiparameter fits it is often useful to plot contours of $\chi^2$ (or of $M$) as a function of pairs of the parameters to study the uncertainties. (See Chapter 11.)

There are several other ways to estimate the uncertainty in a parameter after performing a maximum-likelihood fit. If the distribution of the likelihood function is sufficiently close to a Gaussian, we can find $\sigma_\tau$ from Equation (8.11):

$$\sigma_\tau^2 = \left(\frac{\partial^2 M(\tau)}{\partial \tau^2}\right)^{-1} \tag{10.22}$$

If it is not possible to calculate Equation (10.22) exactly (although it is possible for our example), we can find the second derivative by taking finite differences as discussed in Appendix A.

If the likelihood function does not follow the Gaussian distribution, we can try a numerical integration of the likelihood function to find limiting values that include ~68.3% of the total area, corresponding to the 1 standard deviation limit. Alternatively, we may use a method suggested by Orear (1958) who points out that, for small event samples, where the likelihood function may not be very Gaussianlike, it may be preferable to calculate an average value of the second derivative through the equation

$$\overline{\frac{\partial^2 M}{\partial a^2}} = \frac{\int [\partial^2 M/\partial a^2]\, \mathcal{L}(a)\, da}{\int \mathcal{L}(a)\, da} \tag{10.23}$$

where $a$ is the unknown parameter and the integrals are over the allowable range of the parameter. This procedure has the advantage over the method of Equation (10.22) of giving more weight to the tails of the distribution in cases where they drop off more slowly than those of a Gaussian curve.

Another method of determining the uncertainties in the parameters is to use a Monte Carlo calculation to produce simulated data sets, comparable to our measured data, and to use the method outlined in Chapter 11 for determining confidence levels for our results. This method has the advantage that it depends only on the assumptions made in the Monte Carlo generation, and not on any statistical expectations about the shape of the likelihood function. In many experiments, especially those with low statistics, it provides the most reliable estimate of parameter uncertainties.



## Goodness of Fit

One disadvantage of the direct maximum-likelihood method is that it does not provide a convenient test of the quality of the fit. The value at the peak of the likelihood function itself is not useful because it represents only the maximized probability for obtaining our particular experimental result and we have no way of predicting the expected probability.

An estimate of the goodness of fit can be obtained by making a histogram of the data and comparing it to a prediction based on our best estimate of the parameters. A Monte Carlo simulation of the experiment may be required to calculate the predicted distribution, with a $\chi^2$ test to compare the data to the prediction.

It is not always clear just which data variable should be histogrammed for this purpose. We would like to find that variable on which the parameters depend most strongly. For our sample problem, the lifetime $\tau$ in the rest frames of the particles is an obvious choice, because that is the variable we would choose if we were to solve the problem by the least-squares method. However, it might be wise to try plots of several variables to be sure that the fit is satisfactory. To test, we could generate with our Monte Carlo program a large sample of events based on the parameters discovered in each search, apply the fiducial cuts, and calculate $\chi^2$ from the agreement between the Monte Carlo results and our data sample. We should be aware that, because we did not actually minimize $\chi^2$ for the experimental distribution with respect to the parameters, a satisfactory value of $\chi^2$ may be at best an indication that nothing is drastically wrong with the solution.

## SUMMARY

*Normalized probability density function:*

$$P_i = P(x_i, a_1, a_2, \ldots, a_m)$$

*Likelihood function:*

$$\mathcal{L}(a_1, a_2, \ldots, a_m) = \prod_{i=1}^{N} P_i$$

*Single-event probability density:* $P_i = A_i \cdot p(x_i; a)$ where $A_i$ is the detection efficiency and $p(x_i; a)$ is proportional to the interaction probability
*Logarithm of likelihood function:* $M = \ln \mathcal{L} = \Sigma \ln P_i$
*Maximization of $\mathcal{L}$ or of $M$:* $\partial \mathcal{L}/\partial a_j = 0$ or $\partial M/\partial a_j = 0$ for all $a_j$

*Gaussian form of likelihood function for large data sample:*

$$\mathcal{L}(a_j) \propto \exp\left(-\frac{(a_j - a_j')^2}{2\sigma^2}\right)$$

*Uncertainties in parameters:*

$$\sigma_j^2 = \left(\frac{\partial^2 M(a_j)}{\partial a_j^2}\right)^{-1}$$

*Method for low statistics:*

$$\overline{\frac{\partial^2 M}{\partial a^2}} = \frac{\int [\partial^2 M/\partial a^2]\, \mathcal{L}(a)\, da}{\int \mathcal{L}(a)\, da}$$

## EXERCISES

**10.1.** In a scattering experiment, the angles of the scattered particles are measured and the cosines of the angles in the center-of-mass rest frame of the incident and target particles are calculated and recorded. Fifty such measurements, drawn from the distribution $y(x) = a_1 + a_2 \cos^2\theta$, are listed in the table. Use the direct maximum-likelihood method to determine the values of the parameters $a_1$ and $a_2$. Note that it is necessary to convert the distribution function $y(x_i)$ to a normalized probability function and that the normalization constant will be different for each pair of trial values of $a_1$ and $a_2$.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| −0.999 | −0.983 | −0.956 | −0.946 | −0.933 | −0.925 | −0.916 | −0.910 |
| −0.881 | −0.739 | −0.734 | −0.717 | −0.715 | −0.675 | −0.665 | −0.649 |
| −0.621 | −0.537 | −0.522 | −0.508 | −0.499 | −0.471 | −0.460 | −0.419 |
| −0.403 | −0.311 | −0.305 | −0.281 | −0.170 | −0.162 | −0.063 | 0.214 |
| 0.438 | 0.444 | 0.508 | 0.586 | 0.638 | 0.677 | 0.721 | 0.730 |
| 0.768 | 0.785 | 0.790 | 0.793 | 0.877 | 0.896 | 0.931 | 0.938 |
| 0.948 | 0.993 | | | | | | |

Because of the small amount of data, the uncertainties in the parameters $a_1$ and $a_2$ are so large that the values of the parameters are not very meaningful. Therefore, to complete the problem, you should use the Monte Carlo program written for Exercise 5.8 to generate 500 events and use your calculation to find the parameters from those data.

**10.2.** Students in an undergraduate physics laboratory determined the mass of the $\Lambda$ hyperon by measuring graphically the energies and the momentum vectors of the proton and $\pi$ meson into which the $\Lambda$ hyperons decayed. Because of the large uncertainties in the measurements, the calculated square of the masses of the decaying particles forms a truncated Gaussian distribution that is limited on the low-mass side by $(M_p + M_\pi)^2 = 1.1617\ (\text{GeV}/c^2)^2$, but is not limited on the high-mass side. The following 50 numbers represent squares of the calculated masses in units of $(\text{GeV}/c^2)^2$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.2981 | 1.2618 | 1.2145 | 1.2539 | 1.4230 | 1.3963 | 1.3701 | 1.2303 | 1.3655 | 1.2042 |
| 1.3190 | 1.2086 | 1.2118 | 1.2078 | 1.2726 | 1.2438 | 1.1838 | 1.1666 | 1.1908 | 1.1922 |
| 1.2525 | 1.3615 | 1.1855 | 1.2697 | 1.2044 | 1.3397 | 1.4317 | 1.2713 | 1.2203 | 1.2817 |
| 1.2046 | 1.2856 | 1.1980 | 1.2595 | 1.1721 | 1.2608 | 1.1689 | 1.4838 | 1.1743 | 1.2954 |
| 1.2586 | 1.2655 | 1.2316 | 1.2372 | 1.2969 | 1.2015 | 1.2000 | 1.1677 | 1.2080 | 1.1893 |



Use the direct maximum-likelihood method to fit a truncated Gaussian to these data to determine the maximum-likelihood value of the mass of the squared particle. A search in two-parameter space will be required since neither the mean nor the width of the distribution is known.

Note that it is necessary to calculate numerically the normalization of the truncated Gaussian for each pair of trial values of the mean and standard deviation of the Gaussian function. It is advisable to set up a table of the integral of the standard Gaussian and to use interpolation to find the desired normalizations. A simple automatic or manual grid search will suffice for maximizing the likelihood function.

**10.3.** Use Program 5.4 (available on the website) to generate 1000 sample kaon decay events with nominal mean life $\tau = 0.894 \times 10^{-10}$ s.

(*a*) Plot a histogram of the times of flight of all the generated kaons in their own rest frames (proper times).

(*b*) Use Program 10.1 (available on the website), with nominal fiducial cuts on your data ($d_1 = 10.0$ cm and $d_2 = 40$ cm) to repeat the analysis of Example 10.1d to find the maximum likelihood solution $\tau'$ for the kaon mean life. Plot a histogram of the events that survive the cuts.

(*c*) With the value of $\tau'$, which you determined in part (*b*), and random number seeds that are different from those used in part (*a*), generate 20,000 events to serve as your estimate of the parent distribution. Apply the nominal fiducial cuts to these data and plot a histogram of the data in the same bins as you used in part (*b*).

(*d*) Calculate $\chi^2$ for the agreement between your "experimental" histogram and the surviving events from the "parent" distribution. If the numbers of events in your bins of the parent distribution are large enough, their uncertainties can be ignored in this calculation. If they are not, you must use the combined statistical errors of the two distributions when calculating $\chi^2$.

# CHAPTER 11

# TESTING THE FIT

## 11.1 $\chi^2$ TEST FOR GOODNESS OF FIT

The method of least squares is based on the hypothesis that the optimum description of a set of data is one that minimizes the weighted sum of the squares of the deviation of the data $y_i$ from the fitting function $y(x_i)$. The sum is characterized by the variance of the fit $s^2$, which is an *estimate* of the variance of the data $\sigma^2$. For a function $y(x_i)$, which is linear in $m$ parameters and is fitted to $N$ data points, we have

$$s^2 = \frac{1}{N-m}\frac{\Sigma[(1/\sigma_i^2)[y_i - y(x_i)]^2]}{(1/N)\Sigma(1/\sigma_i^2)} = \frac{1}{N-m}\sum w_i[y_i - y(x_i)]^2 \qquad (11.1)$$

where the factor $\nu = N - m$ is the number of degrees of freedom for fitting $N$ data points (implied in the unlabeled sums) with $m$ parameters and the weighting factor for each measurement is given by

$$w_i = \frac{1/\sigma_i^2}{(1/N)\Sigma(1/\sigma_i^2)}, \qquad (11.2)$$

the inverse of the variance $\sigma_i^2$ that describes the uncertainties in each point, normalized to the average of all the weighting factors.

The variance of the fit $s^2$ is also characterized by the statistic $\chi^2$ defined in Equation (7.5) for polynomials:

$$\chi^2 \equiv \sum\left\{\frac{1}{\sigma_i^2}[y_i - y(x_i)]^2\right\} \qquad (11.3)$$





with

$$y(x_i) = \sum_{k=1}^{m} a_k f_k(x_i)$$

The relationship between $s^2$ and $\chi^2$ can be seen most easily by comparing $s^2$ with the reduced chi-square $\chi_\nu^2$,

$$\chi_\nu^2 = \frac{\chi^2}{\nu} = \frac{s^2}{\langle\sigma_i^2\rangle} \qquad (11.4)$$

where $\langle\sigma_i^2\rangle$ is the weighted average of the individual variances

$$\langle\sigma_i^2\rangle = \frac{(1/N)\Sigma((1/\sigma_i^2)\sigma_i^2)}{(1/N)\Sigma(1/\sigma_i^2)} = \left[\frac{1}{N}\sum\frac{1}{\sigma_i^2}\right]^{-1} \qquad (11.5)$$

and is equivalent to $\sigma^2$ if the uncertainties are all equal, $\sigma_i = \sigma$.

The parent variance of the data $\sigma^2$ is a characteristic of the dispersion of the data about the parent distribution and is not descriptive of the fit. The estimated variance of the fit $s^2$, however, is characteristic of both the spread of the data and the accuracy of the fit. The definition of $\chi^2$, as the ratio of the estimated variance $s^2$ to the parent variance $\sigma^2$ times the number of degrees of freedom $\nu$, makes it a convenient measure of the goodness of fit.

If the fitting function is a good approximation to the parent function, then the estimated variance $s^2$ should agree well with the parent variance $\sigma^2$, and the value of the reduced chi-square should be approximately unity, $\chi_\nu^2 = 1$. If the fitting function is not appropriate for describing the data, the deviations will be larger and the estimated variance will be too large, yielding a value of $\chi_\nu^2$ greater than 1. A value of $\chi_\nu^2$ less than 1 does not necessarily indicate a better fit, however; it is simply a consequence of the fact that there exists an uncertainty in the determination of $s^2$, and the observed values of $\chi_\nu^2$ will fluctuate from experiment to experiment. A value of $\chi_\nu^2$ that is very small may indicate an error in the assignment of the uncertainties in the measured variables.

### Distribution of $\chi^2$

The probability distribution function for $\chi^2$ with $\nu$ degrees of freedom is given by

$$p_\chi(x^2; \nu) = \frac{(x^2)^{1/2(\nu-2)}e^{-x^2/2}}{2^{\nu/2}\Gamma(\nu/2)} \qquad (11.6)$$

The chi-square distribution of Equation (11.6) is derived in many texts on statistics[1] but we shall simply quote the results here.

The gamma function $\Gamma(n)$ is equivalent to the factorial function $n!$ extended to nonintegral arguments. It is defined for integral and half-integral arguments by the values at arguments of 1 and ½ and a recursion relation:

[1]See Pugh and Winslow (1966), Section 12-5.

$$\left.\begin{aligned} &\Gamma(1)=1 \qquad \Gamma(\tfrac{1}{2})=\sqrt{\pi} \qquad \Gamma(n-1)=n\Gamma(n) \\ &\text{For integral values of } n \\ &\qquad \Gamma(n+1)=n! \qquad n=0,1,\ldots \\ &\text{For half-integral values of } n \\ &\qquad \Gamma(n+1)=n(n-1)(n-2)\cdots(\tfrac{3}{2})(\tfrac{1}{2}\sqrt{\pi}) \\ &\qquad\qquad n=\tfrac{1}{2},\tfrac{3}{2},\tfrac{5}{2},\ldots \end{aligned}\right\} \quad ((11.7)$$

Calculating factorial functions can lead to computer overflow problems. For computational purposes it is convenient to replace the factorial form of the gamma function by a form of Stirling's approximation[2]:

$$\Gamma[n]=\sqrt{2\pi}e^{-n}n^{(n-1/2)}(1+0.0833/n) \tag{11.8}$$

This approximation, which is accurate to ~0.1% for all $n \geq \frac{1}{2}$, avoids both the problems of overflow in calculating factorials and the necessity of testing and choosing the appropriate form for integral or half-integral argument. The trade-off is computer speed. Calculating exponentials may be slower than calculating factorials, but high speed usually is not required for nonrepetitive calculations.

If the function of the parent population is denoted by $y_0(x)$, the value of $\chi_0^2$ determined from the parameters of the parent function

$$\chi_0^2=\sum\left\{\frac{1}{\sigma_i^2}[y_i-y_0(x_i)]^2\right\} \tag{11.9}$$

is distributed according to Equation (11.6) with $\nu = N$ degrees of freedom. If the function $y(x)$ used in the determination of $\chi^2$ contains $m$ parameters, the value of $\chi^2$ calculated from Equation (11.3) is distributed according to Equation (11.6) with $\nu = N - m$ degrees of freedom.

More useful for our purposes than the probability density distribution $p_\chi(x^2; \nu)$ of Equation (11.6) is the integral probability $P_\chi(\chi^2; \nu)$ between $x^2 = \chi^2$ and $\chi^2 = \infty$:

$$P_\chi(\chi^2;\nu)=\int_{\chi^2}^{\infty}P_\chi(x^2;\nu)\,dx^2 \tag{11.10}$$

Equation (11.10) describes the probability that a random set of $n$ data points drawn from the parent distribution would yield a value of $\chi^2$ equal to or greater than the tabulated value.

**Program 11.1.** CHI2PROB (Appendix E) $\chi^2$-probability.
CHIPROBDENS computation of the function $p_\chi(\chi^2; \nu)$ [Equation (11.6)] using function GAMMA to approximate the gamma function.
CHIPROB Numerical calculation of the integral, Equation (11.10), by Simpson's rule. If variable overflow is a problem, double-precision variables could be employed.

[2] "Review of Particle Properties" (1986), p. 53.



The calculation returns the integral to an accuracy of about ±0.1%. The trade-off on accuracy versus speed of computation is controlled by the value of the constant DX, the integration step.

For the special case of 1 degree of freedom, $\nu = 1$, the $\chi^2$-probability density function of Equation (11.6) takes the form

$$p_\chi(x^2;\nu)=e^{-x^2/2}/(2\pi x^2)^{1/2}$$

which is difficult to integrate numerically near $x = 0$. However, the integral is finite, and the function can be expanded in a Taylor series about $x = 0$ and integrated analytically. We use that technique for $\nu = 1$ and $\chi^2 < 2$.

Similarly, for $\nu = 2$, where the function takes the form

$$p_\chi(x^2;\nu)=e^{-x^2/2}/2$$

the analytic form of the integral is used.

For a fitting function that is a good approximation to the parent function, the experimental value of $\chi_\nu^2$ should be close to one and the probability from Equation (11.10) should be approximately 0.5. For poorer fits, the values of $\chi_\nu^2$ will be larger and the associated probability will be smaller. There is an ambiguity in interpreting the probability because $\chi_\nu^2$ is a function of the quality of the data as well as the choice of parent function, so that even correct fitting functions occasionally yield large values of $\chi_\nu^2$. However, the probability of Equation (11.10) is generally either reasonably close to 0.5, indicating a reasonable fit, or unreasonably small, indicating a bad fit. In fact, for most purposes, the reduced chi-square $\chi_\nu^2$ is an adequate measure of the probability directly. The probability will be reasonably close to 0.5 so long as $\chi_\nu^2$ is reasonably close to 1; that is, less than about 1.5.

**Example 11.1.** Consider the solution of the problem of fitting two exponential curves plus a linear background to the data from the radioactive silver decay of Example 8.1. The fit (see Table 8.5) gave $\chi^2 = 66.1$ for 54 degrees of freedom, or $\chi_\nu^2 = 1.22$, with $P_\chi(\chi^2; \nu) = 12.4\%$. We can interpret this result in the following way. Assume that the parameters we found are, indeed, the parameters of the parent distribution. Then, suppose that we were to repeat our experiment many times, drawing many different data samples from that parent distribution. Our result indicates that in 12.4% of those experiments we should expect to obtain fits that are no better than that listed in Table 8.5.

## 11.2 LINEAR-CORRELATION COEFFICIENT

Let us assume that we have made measurements of pairs of quantities $x_i$ and $y_i$. We know from the previous chapters how to fit a function to these data by the least-squares method, but we should stop and ask whether the fitting procedure is justified and whether, indeed, there *exists* a physical relationship between the variables $x$ and $y$. What we are asking here is whether or not the variations in the observed values of one quantity $y$ are *correlated* with the variations in the measured values of the other quantity $x$.

For example, if, as in Example 6.1, we were to measure the potential difference across segments of a current-carrying wire as a function of the segment length, we should find a definite and reproducible correlation between the two quantities. But if we were to measure the potential of the wire as a function of time, even though there might be fluctuations in the observations, we should not find any significant reproducible long-term relationship between the pairs of measurements.

On the basis of our discussion in Chapter 6, we can develop a quantitative measure of the degree of correlation or the probability that a linear relationship exists between two observed quantities. We can construct a linear-correlation coefficient $r$ that will indicate quantitatively whether or not we are justified in determining even the simplest linear correspondence between the two quantities.

### Reciprocity in Fitting $x$ Versus $y$

Our data consist of pairs of measurements $(x_i, y_i)$. If we consider the quantity $y$ to be the dependent variable, then we want to know if the data correspond to a straight line of the form

$$y = a + bx \tag{11.11}$$

We have already developed the analytical solution for the coefficient $b$, which represents the slope of the fitted line given in Equation (6.12):

$$b = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma x_i^2 - (\Sigma x_i)^2} \tag{11.12}$$

where the weighting factors in $\sigma_i$ have been omitted for clarity. If there is no correlation between the quantities $x$ and $y$, then there will be no tendency for the values of $y$ to increase or decrease with increasing $x$, and, therefore, the least-squares fit must yield a horizontal straight line with a slope $b = 0$. But the value of $b$ by itself cannot be a good measure of the degree of correlation because a relationship might exist that included a very small slope.

Because we are discussing the interrelationship between the variables $x$ and $y$, we can equally well consider $x$ as a function of $y$ and ask if the data correspond to a straight-line form

$$x = a' + b'y \tag{11.13}$$

The values of the coefficients $a'$ and $b'$ will be different from the values of the coefficients $a$ and $b$ in Equation (11.11), but they are related if the variables $x$ and $y$ are correlated.

The analytical solution for the inverse slope $b'$ is similar to that for $b$ in Equation (11.12):

$$b' = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma y_i^2 - (\Sigma y_i)^2} \tag{11.14}$$

If there is no correlation between the quantities $x$ and $y$, then the least-squares fit must yield a horizontal straight line with a slope $b' = 0$.



If there is a complete correlation between $x$ and $y$, then there exists a relationship between the coefficients $a$ and $b$ of Equation (11.11) and between $a'$ and $b'$ of Equation (11.13). To see what this relationship is, we rewrite Equation (11.13):

$$y = -\frac{a'}{b'} + \frac{1}{b'}x = a + bx \tag{11.15}$$

and equate coefficients

$$a = -\frac{a'}{b'} \qquad b = \frac{1}{b'} \tag{11.16}$$

We see from Equation (11.16) that $bb' = 1$ for complete correlation. If there is no correlation, both $b$ and $b'$ are 0 and Equations (11.16) do not apply. We, therefore define, as a measure of the degree of linear correlation, the experimental linear-correlation coefficient $r \equiv \sqrt{bb'}$:

$$r \equiv \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[N\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2}[N\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}} \tag{11.17}$$

The value of $r$ ranges from 0, when there is no correlation, to $\pm 1$, when there is complete correlation. The sign of $r$ is the same as that of $b$ (and $b'$), but only the absolute magnitude is important.

The correlation coefficient $r$ cannot be used directly to indicate the degree of correlation. A probability distribution for $r$ can be derived from the two-dimensional Gaussian distribution, but its evaluation requires a knowledge of the correlation coefficient $\rho$ of the parent population. A more common test of $r$ is to compare its value with the probability distribution for the parent population that is completely uncorrelated; that is, for which $\rho = 0$. Such a comparison will indicate whether or not it is probable that the data points could represent a sample derived from an uncorrelated parent population. If this probability is small, then it is more probable that the data points represent a sample from a parent population where the variables are correlated.

For a parent population with $\rho = 0$, the probability that any random sample of uncorrelated experimental data points would yield an experimental linear-correlation coefficient equal to $r$ is given by[3]

$$p_r(r; \nu) = \frac{1}{\sqrt{\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} (1 - r^2)^{(\nu-2)/2} \tag{11.18}$$

where $\nu = N - 2$ is the number of degrees of freedom for an experimental sample of $N$ data points. The gamma function for integral and half-integral values was defined in Equation (11.7).

### Integral Probability

A more useful distribution than that of Equation (11.18) is the probability $P_c(r; N)$ that a random sample of $N$ uncorrelated experimental data points would yield an

[3]For a derivation see Pugh and Winslow (1966), Section 12-8.

experimental linear-correlation coefficient as large as or larger than the observed value of $|r|$. This probability is the integral of $p_c(r; \nu)$ for $\nu = N - 2$:

$$P_c(r; N) = 2\int_{|r|}^{1} p_c(r; \nu)\, dx \qquad \nu = N - 2 \tag{11.19}$$

With this definition, $P_c(r; N)$ indicates the probability that the observed data could have come from an uncorrelated ($\rho = 0$) parent population. A small value of $P_c(r; N)$ implies that the observed variables are probably correlated.

Because Equation (11.19) cannot be integrated analytically, the function must be integrated either by making a series expansion of the argument and integrating term by term or by performing a numerical integration. With fast computers, the latter method is more convenient and generally applicable to such problems.

**Program 11.2** LCORLATE (Appendix E) Correlation probability computations. LCORPROB computes the probability of Equation (11.19) by numerical integration. Input variables RCORR and NOBSERV correspond to the value of the experimental linear-correlation coefficient and the number of observations, respectively. (The number of degrees of freedom is the number of observations minus 2.) The program uses the following routines: LINCORREL computes the function $p_c(r; \nu)$ of Equation (11.18) using the approximation of Equation (11.8) for the gamma function (calculated by the function GAMMA in the program unit GENUTIL). Because LINCORREL is intended to be used as an argument to the integration routine SIMPSON, it can have only one argument. The parameter $\nu$ is passed in the global variable PSIMPS by the calling routine.
LINCORPROB computes $P_c(r; \nu)$ of Equation (11.19) by numerically integrating LINCORREL by Simpson's rule. The calculation returns the integral to an accuracy of about $\pm 0.01$. The trade-off on accuracy versus speed of computation is controlled by the value of the constant DX, the integration step.

**Example 11.2.** For the data of Example 6.1, the linear-correlation coefficient $r$ can be calculated from Equation (11.17) with the data of Table 6.1:

$$r = \frac{9 \times 779.3 - 450.0 \times 12.44}{\sqrt{(9 \times 28{,}500 - 450.0^2) \times (9 \times 21.32 - 12.44^2)}} = 0.9998$$

The probability for determining, from an uncorrelated population with $9 - 2 = 7$ degrees of freedom, a value of $r$ equal to or larger than the observed value, can be calculated from Equation (11.19) (see Table C.3). The result $P_c(r; N) < 0.001\%$ indicates that it is extremely improbable that the variables $x$ and $V$ are linearly uncorrelated. Thus, the probability is high that the variables are correlated and the linear fit is justified.

Similarly, in the experiment of Example 6.2, the linear-correlation coefficient can be calculated from Equation (11.17) by including the weighting factors $\sigma_i^2 = y_i$ as in Table 6.2, so that, for example, $N$ is replaced by $\Sigma w_i$ and $\Sigma x_i$ is replaced by $\Sigma w_i x_i$, and so forth:

$$r = \frac{0.03570 \times 81.02 - 0.1868 \times 10}{\sqrt{(0.03570 \times 1.912 - 0.1868^2) \times (0.03570 \times 3693 - 10^2)}} = 0.9939$$



Again, the probability $P_c(r; N)$ for $r = +0.9938$ with $\nu = 10 - 2 = 8$ degrees of freedom is very small ($< 0.001\%$), indicating that the change in counting rate $C$ is linearly correlated to a high degree of probability with $x = 1/r^2$, the inverse square of the distance between the source and counter.

## 11.3 MULTIVARIABLE CORRELATIONS

If the dependent variable $y_i$ is a function of more than one variable,

$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots \tag{11.20}$$

we might investigate the correlation between $y_i$ and each of the independent variables $x_{ij}$ or we might also enquire into the possibility of correlation between different variables $x_{ij}$. Here, we use the first subscript $i$ to represent the observation, as in the previous discussions, and the second subscript $j$ to represent the particular variable under investigation. The variables $x_{ij}$ could be different variables, or they could be functions of $x_i$, $f(x_i)$, as in Chapter 7. We shall rewrite Equation (11.17) for the linear-correlation coefficient $r$ in terms of another quantity $s_{jk}^2$.

We define the *sample covariance* $s_{jk}^2$:

$$s_{jk}^2 \equiv \frac{1}{N-1}\sum[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)] \tag{11.21}$$

where the means $\bar{x}_j$ and $\bar{x}_k$ are given by

$$\bar{x}_j \equiv \frac{1}{N}\sum x_{ij} \quad \text{and} \quad \bar{x}_k = \frac{1}{N}\sum x_{ik} \tag{11.22}$$

and the sums are taken over the range of the subscript $i$ from 1 to $N$. The weights have been omitted for clarity. With this definition, the sample variance for one variable $s_j^2$,

$$s_j^2 \equiv s_{jj}^2 = \frac{1}{N-1}\sum(x_{ij} - \bar{x}_j)^2 \tag{11.23}$$

is analogous to the sample variance $s^2$ defined in Equation (1.9):

$$s^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2 \tag{11.24}$$

It is important to note that the sample variances $s_j^2$ defined by Equation (11.23) are measures of the ranges of variation of the variables and not of the uncertainties in the variables.

Equation (11.21) can be rewritten for comparison with Equation (11.17) by substituting the definitions of Equation (11.22):

$$\begin{aligned} s_{jk}^2 &\equiv \frac{1}{N-1}\sum[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)] \\ &= \frac{1}{N-1}\sum(x_{ij}x_{ik} - \bar{x}_j\bar{x}_k) \\ &= \frac{1}{N-1}\sum\left(x_{ij}x_{ik} - \frac{1}{N}\sum x_{ij}\sum x_{ik}\right) \end{aligned} \tag{11.25}$$

If we substitute $x_{ij}$ for $x_i$ and $x_{ik}$ for $y_i$ in Equation (11.17), we can define the *sample linear-correlation coefficient* between any two variables $x_j$ and $x_k$ as

$$r_{jk} = \frac{s_{jk}^2}{s_j s_k} \tag{11.26}$$

with the covariances and variances $s_{jk}^2$, $s_j^2$, and $s_k^2$ given by Equations (11.23) and (11.25). Thus, the linear-correlation coefficient between the $j$th variable $x_j$ and the dependent variable $y$ is given by

$$r_{jy} = \frac{s_{jy}^2}{s_j s_y} \tag{11.27}$$

Similarly, the linear-correlation coefficient of the parent population of which the data are a sample is defined as

$$\rho_{jk} = \frac{\sigma_{jk}^2}{\sigma_j \sigma_k} \tag{11.28}$$

where $\sigma_j^2$, $\sigma_k^2$, and $\sigma_{jk}^2$ are the true variances and covariances of the parent population. These linear-correlation coefficients are also known as product-moment correlation coefficients.

With these definitions we can consider either the correlation between the dependent variable and any other variable $r_{jy}$ or the correlation between any two variables $r_{jk}$.

## Polynomials

In Chapter 7 we investigated functional relationships between $y$ and $x$ of the form

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots \tag{11.29}$$

In a sense, this is a variation on the linear relationship of Equation (11.20) where the powers of the single independent variable $x$ are considered to be various variables $x_j = x^j$. The correlation between the independent variable $y$ and the $m$th term in the power series of Equation (11.29), therefore, can be expressed in terms of Equations (11.23) through (11.27):

$$\begin{aligned} r_{my} &= \frac{s_{my}^2}{s_m s_y} \\ s_m^2 &= \frac{1}{N-1}\left[\Sigma x_i^{2m} - \frac{1}{N}(\Sigma x_i^m)^2\right] \\ s_y^2 &= \frac{1}{N-1}\left[\Sigma y_i^2 - \frac{1}{N}(\Sigma y_i)^2\right] \\ s_{my}^2 &= \frac{1}{N-1}\left[\Sigma x_i^m y_i - \frac{1}{N}\Sigma x_i^m \Sigma y_i\right] \end{aligned} \tag{11.30}$$



## Weighted Fit

If the uncertainties in the data points are not all equal ($\sigma_i \neq \sigma$), we must include the individual standard deviations $\sigma_i$ as weighting factors in the definition of variances, covariances, and correlation coefficients. From Chapter 6 the prescription for introducing weighting is to multiply each term in the sum by $1/\sigma_i^2$.

The formula for the correlation remains the same as Equations (11.26) and (11.27), but the formulas of Equations (11.21) and (11.23) for calculating the variances and covariances must be modified:

$$\begin{aligned} s_{jk}^2 &\equiv \frac{1/(N-1)\Sigma[(1/\sigma_i^2)(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)]}{(1/N)\Sigma(1/\sigma_i^2)} \\ s_j^2 \equiv s_{jj}^2 &= \frac{1/(N-1)\Sigma[(1/\sigma_i^2)(x_{ij} - \bar{x}_j)^2]}{(1/N)\Sigma(1/\sigma_i^2)} \end{aligned} \tag{11.31}$$

where the means $\bar{x}_j$ and $\bar{x}_k$ are also weighted means

$$\bar{x}_j = \frac{\Sigma x_{ij} w_i}{N} = \frac{\Sigma(x_{ij}/\sigma_i^2)}{\Sigma(1/\sigma_i^2)}$$

The weighting factors

$$w_i = \frac{1/\sigma_i^2}{(1/N)\Sigma(1/\sigma_i^2)} \tag{11.32}$$

for each data point are the inverse of the variances $\sigma_i^2$ that describe the uncertainties in each point, normalized to the average of all the weighting factors.

## Multiple-Correlation Coefficient

We can extrapolate the concept of the linear-correlation coefficient, which characterizes the correlation between two variables at a time, to include multiple correlations between groups of variables taken simultaneously. The linear-correlation coefficient $r$ of Equation (11.17) between $y$ and $x$ can be expressed in terms of the variances and covariances of Equation (11.31) and the slope $b$ of a straight-line fit given in Equation (11.12):

$$r^2 = \frac{s_{xy}^4}{s_x^2 s_y^2} = b\frac{s_{xy}^2}{s_y^2} \tag{11.33}$$

In analogy with this definition of the linear-correlation coefficient, we define the *multiple-correlation coefficient* $R$ to be the sum over similar terms for the variables of Equation (11.20):

$$R^2 \equiv \sum_{j=1}^{n}\left(b_j \frac{s_{jy}^2}{s_y^2}\right) = \sum_{j=1}^{n}\left(b_j \frac{s_j}{s_y} r_{jy}\right) \tag{11.34}$$

The linear-correlation coefficient $r$ is useful for testing whether one particular variable should be included in the theoretical function that is fitted to the data. The

multiple-correlation coefficient $R$ characterizes the fit of the data to the entire function. A comparison of the multiple-correlation coefficient for different functions is therefore useful in optimizing the theoretical functional form.

We shall discuss in the following sections how to use these correlation coefficients to determine the validity of including each term in the polynomial of Equation (11.29) or the series of arbitrary functions of Equation (11.20).

## 11.4 $F$ TEST

As noted in Section 11.1, the $\chi^2$ test is somewhat ambiguous unless the form of the parent function is known, because the statistic $\chi^2$ measures not only the discrepancy between the estimated function and the parent function, but also the deviations between the data and the parent function simultaneously. We would prefer a test that separates these two types of information so that we can concentrate on the former type. One such test is the $F$ test, which combines two different methods of determining a $\chi^2$ statistic and compares the results to see if their relation is reasonable.

### $F$ Distribution

If two statistic $\chi_1^2$ and $\chi_2^2$, which follow the $\chi^2$ distribution, have been determined, the ratio of the reduced chi-squareds, $\chi_{\nu_1}^2$ and $\chi_{\nu_2}^2$, is distributed according to the $F$ distribution[4]

$$f = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \tag{11.35}$$

with probability density function

$$P_f(f; \nu_1, \nu_2) = \frac{\Gamma[(\nu_1 + \nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{f^{1/2(\nu_1 - 2)}}{(1 + f\nu_1/\nu_2)^{1/2(\nu_1+\nu_2)}} \tag{11.36}$$

where $\nu_1$ and $\nu_2$ are the numbers of degrees of freedom corresponding to $\chi_1^2$ and $\chi_2^2$. By the definition of $\chi_\nu^2$ [see Equation (11.4)], a ratio of ratios of variances

$$\frac{\chi_{\nu_1}^2}{\chi_{\nu_2}^2} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{11.37}$$

is also distributed as $F$, where $s_1$ and $s_2$ are experimental estimates of standard deviations $\sigma_1$ and $\sigma_2$ pertaining to some characteristic of the same or different distributions.

As with our tests of $\chi^2$ and the linear-correlation coefficient $r$, we shall be more interested in the integral probability

$$P_F(F; \nu_1, \nu_2) = \int_F^\infty P_f(f; \nu_1, \nu_2)\, df \tag{11.38}$$

which describes the probability of observing such a large value of $F$ from a random set of data when compared to the correct fitting function. The integral function $P_F(F; \nu_1, \nu_2)$ is tabulated and graphed in Table C.5 for a wide range of $F$, $\nu_1$, and $\nu_2$.

[4]See Pugh and Winslow (1966), Section 12-7, for a derivation.



A word of caution is in order concerning the use of these tables. Because the statistic $F$ in Equation (11.35) is defined as the ratio of two determinations of $\chi^2$ without specifying which must be in the numerator, we can define two statistics $F_{12}$ and $F_{21}$,

$$F_{12} = \frac{\chi_{\nu 1}^2}{\chi_{\nu 2}^2} \qquad F_{21} = \frac{\chi_{\nu 2}^2}{\chi_{\nu 1}^2} = \frac{1}{F_{12}} \tag{11.39}$$

which must both be distributed according to the $F$ distribution.

If in some experiment our calculations yield a particular value of $F_{12}$, we can use Table C.5 to determine whether such a large value is less than 5% probable (Table C.6 and Figure C.6) or less than 1% probable (Table C.7 and Figure C.7). If the test value is less than the tabulated values, we must also make sure that it is not too small. To do this, we compare the value

$$F_{21} = 1/F_{12} \tag{11.40}$$

to the same tables and graphs, noting that the values of $\nu_1$ and $\nu_2$ are reversed. The values of $\nu_1$ and $\nu_2$ specified in Table C.5 correspond to the degrees of freedom for the numerator and denominator of Equation (11.39), respectively.

**Example 11.3.** Suppose that $F_{12} = 0.2$ with $\nu_1 = 2$ and $\nu_2 = 10$. For Table C.6, the observed value of $F_{12}$ may be as high as 4.10 and still be exceeded by about 5% of random observations. Similarly, we compare $F_{21} = 1/F_{12} = 5.0$ with the 5% point for $\nu_1 = 10$ and $\nu_2 = 2$, which has a value of 19.4. Because the values of $F_{12}$ and $F_{21}$ are well within the 5% limits, we can have confidence in the fit.

What we are estimating in this example is the probability $P_F(F_{12}; \nu_1, \nu_2)$ that $F_{12}$ is not too large and the probability $P_F(1/F_{12}; \nu_2, \nu_1)$ that $F_{12}$ is not too small. It is tempting to simplify this procedure by assuming that

$$P_F(1/F_{12}; \nu_2, \nu_1) = P_F(F_{12}; \nu_1, \nu_2) \tag{11.41}$$

so that our test consists of determining $F$ such that

$$P_F(F; \nu_1, \nu_2) = 0.05$$

with the requirement that

$$F > F_{12} > 1/F$$

This approximation is valid for reasonably large values of $\nu_1$ and $\nu_2$ but not for small values of either, as in the preceding example, where we have $4.10 > F_{12} > 1/19.4$.

### Multiple-Correlation Coefficient

There are two types of $F$ tests that are normally performed on least-squares fitting procedures. One is designed to test the entire fit and can be related to the multiple-correlation coefficient $R$. The other, to be discussed later, tests the inclusion of an additional term in the fitting function.

If we consider the sum of squares of deviations $S_Y^2$ associated with the spread of the data points around their mean (omitting factors of $1/\sigma_i^2$ for clarity),

$$S_y^2 = \Sigma(y_i - \bar{y})^2 \tag{11.42}$$

this is a statistic that follows the $\chi^2$ distribution with $N - 1$ degrees of freedom (only one parameter $\bar{y}$ must be determined from the $N$ data points). It is a characteristic of quantities that follow the $\chi^2$ distribution that they may be expressed as the sum of other quantities that also follow the $\chi^2$ distribution such that the number of degrees of freedom of the original statistic is the sum of the numbers of degrees of freedom of the terms in the sum.

By suitable manipulation and rearrangement, it can be shown that $S_y^2$ can be expressed as the sum of the two terms,

$$\begin{aligned} S_y^2 = \sum(y_i - \bar{y})^2 &= \sum_{j=1}^{m}\left[(y_i - \bar{y})\sum_{j=1}^{m} a_j(f_j - \bar{f}_j)\right] + \sum_{j=1}^{m}(y_i - \sum a_j f_j)^2 \\ &= \sum_{j=1}^{m}[a_j\sum[(y_i - \bar{y})(f_j - \bar{f}_j)]] + \sum[y_i - y(x_i)]^2 \end{aligned} \tag{11.43}$$

where the fitting function is of the form

$$y(x_i) = \sum_{j=1}^{m} a_j f_j(x_i) \tag{11.44}$$

and we have

$$\bar{f}_j = \frac{1}{N}\sum f_j(x_i) \tag{11.45}$$

The left-hand side of Equation (11.43) is distributed as $\chi^2$ with $N - 1$ degrees of freedom. The right-hand term is our definition of $\chi^2$ from the Equation (11.3) and has $N - m$ degrees of freedom. Consequently, the middle term must be distributed according to the $\chi^2$ distribution with $m - 1$ degrees of freedom.

By comparison with our definition of the multiple-correlation coefficient $R$ in Equation (11.34), we can express this middle tern as a fraction $R^2$ of the statistic $S_y^2$:

$$\sum_{j=1}^{m} a_j\sum[(y_i - \bar{y})(f_j - \bar{f}_j)] = R^2\sum(y_i - \bar{y})^2 \tag{11.46}$$

Equation (11.43) becomes

$$\sum(y_i - \bar{y})^2 = R^2\sum(y_i - \bar{y})^2 + (1 - R^2)\sum(y_i - \bar{y})^2 \tag{11.47}$$

or

$$S_y^2 = R^2 S_y^2 + (1 - R^2)S_y^2 \tag{11.48}$$

where, as before, both terms on the right-hand side are distributed as $\chi^2$, the first with $m - 1$ degrees of freedom and the second with $N - m$ degrees of freedom.

Thus, the physical meaning of the multiple-correlation coefficient becomes evident. It divides the total sum of squares of deviations $S_y^2$ into two parts. The first fraction $R^2S_y^2$ is a measure of the spread of the dependent and independent variable data space. The second fraction, $(1 - R^2)S_y^2$, is the sum of squares of the deviations about the regression and represents the agreement between the fit and the data.



From the definition of Equation (11.35), we can define a ratio $F_R$ of the two terms in the right-hand side of Equation (11.47) that follow the $F$ distribution with $\nu_1 = m - 1$ and with $\nu_2 = N - m$ degrees of freedom,

$$F_R = \frac{R^2/(m-1)}{(1-R^2)/(N-m)} = \frac{R^2}{(1-R^2)} \times \frac{(N-m)}{(m-1)} \tag{11.49}$$

From this definition of $F_R$ in terms of the multiple-correlation coefficient $R$, it is clear that a large value of $F_R$ corresponds to a good fit, where the multiple correlation is good and $R \simeq 1$. The $F$ test for this statistic is actually a test that the coefficients are 0 ($a_j = 0$). So long as $F_R$ exceeds the test value for $F$, we can be fairly confident that our coefficients are nonzero. If, on the other hand, $F_R < F$, we may conclude that at least one of the terms in the fitting function is not valid, is decreasing the multiple correlation by its inclusion, and should have a coefficient of 0.

### Test of Additional Term

Because of the additive nature of functions that obey the $\chi^2$ statistics, we can form a new $\chi^2$ statistic by taking the difference of two other statistics that are distributed as $\chi^2$. In particular, if we fit a set of data with a fitting function with $m$ terms, the resulting value of chi-square associated with the deviations about the regression $\chi^2(m)$ has $N - m$ degrees of freedom. If we add another term to the fitting function, the corresponding value of chi-square $\chi^2(m + 1)$ has $N - m - 1$ degrees of freedom. The difference between these two must follow the $\chi^2$ distribution for 1 degree of freedom.

If we form the ratio of the difference $\chi^2(m) - \chi^2(m + 1)$ to the new value $\chi_\nu^2(m + 1)$, we can form a statistic $F_\chi$ that follows the $F$ distribution with $\nu_1 = 1$ and $\nu_2 = N - m - 1$:

$$F_\chi = \frac{\chi^2(m) - \chi^2(m+1)}{\chi^2(m+1)/(N-m-1)} = \frac{\Delta\chi^2}{\chi_\nu^2} \tag{11.50}$$

This ratio is a measure of how much the additional term has improved the value of the reduced chi-square and should be small when the function with $m + 1$ terms does not significantly improve the fit over the function with $m$ terms. Thus, we can be confident in the relative merit of the new terms if the value of $F_\chi$ is large. As for $F_R$, this is really a test of whether the coefficient for the new term is 0 ($a_{m+1} = 0$). If $F_\chi$ exceeds the test value for $F$, we can be fairly confident that the coefficient should not be 0 and the term, therefore, should be included. Table C.5 and Figure C.5 are useful for testing $F_\chi$. They give the value of $F$ corresponding to various values of the probability $P_F(F; 1, \nu_2)$ and various values of $\nu_2$ for the case where $\nu_1 = 1$. Thus, rather than evaluating $F$ for critical values of the probability (for example, 5% or 1%), we can evaluate the probability corresponding to the observed value of $F_\chi$.

A calculation of $F_\chi$ could be built into a linear regression program and the resulting value compared to a supplied test value $F$, to indicate whether or not the last term in the series is justified, and therefore, to determine how many terms in the series should be included in the fit. However, it is probably safer, except possibly in a large, well debugged production run involving fitting polynomials to many similar data sets, to examine the individual values of $\chi^2$ along with $F_\chi$ and to adjust the

number of terms in the calculation manually. One should, however, be aware that the important figure of merit for added terms is the difference of the two values of $\chi^2$ divided by the new value $\chi^2_\nu$ of the *reduced* chi-square.

## 11.5 CONFIDENCE INTERVALS

The object of data fitting is to obtain values for the parameters of the fitted function, and the uncertainties in the parameters. The quality of the fit is indicated by $\chi^2$ and its associated probability, and the uncertainties give the probabilities that our values of the fitted parameters are good estimates of the parent parameters. Whether we estimate our parameters by the least-squares method or by direct application of the maximum-likelihood method, as discussed in Chapter 10, we must always estimate the uncertainty in our parameters to indicate numerically our confidence in our results.

Generally, we assume Gaussian statistics and quote the standard deviation $\sigma$ in a result, where $\sigma$ appears in the Gaussian probability density function

$$p_G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{11.51}$$

and determines the width of the distribution. As noted in Chapter 2, approximately 68.3% of the events of the Gaussian distribution fall within $\pm\sigma$ of the mean $\mu$ and approximately 95.4% fall within $\pm 2\sigma$.

### Confidence Level for One-Parameter Fit

One way of looking at the 1 standard deviation limit is to consider that, in a series of repeated experiments, there is approximately a 68% chance of obtaining values within $\pm\sigma$ of the mean $\mu$. Of course, we usually do not know $\mu$, and perhaps not $\sigma$ either, but have determined experimentally only $\bar{x}$ and $s$, our estimate of the parameters. However, as long as our experimental estimates $\bar{x}$ and $s$ are reasonably close to the true values $\mu$ and $\sigma$, we can state that there is approximately a 68% probability that the true value of the measured parameter lies between $\bar{x} - s$ and $\bar{x} + s$, or that at the 68.3% *confidence level*, the true value of the parameter lies between these two limits.

We may wish to quote results in terms of other confidence levels. For example, we refer to the $\pm 2\sigma$ limit as the 95.4% confidence interval, or we may quote a 99% or 99.9% confidence level for a high-precision experiment. The conventional $1\sigma$ and $2\sigma$ limits are based on the Gaussian distribution, which may or may not apply to the data in question, and even an experimental distribution that nominally follows Gaussian statistics is apt to deviate in the tails.

For any distribution, represented by the normalized probability density function, $p_x(x;\mu)$, we determine the probability that a measurement of the parameter will fall between $\bar{x} - a$ and $\bar{x} + b$ by the integral

$$P_x = \int_{\bar{x}-a}^{\bar{x}+b} p_x(\bar{x};x)\,dx \tag{11.52}$$

and could quote a confidence level of $P_x$ that the "true" value of the measured parameter is between these two values. Note that we have not specified a region that is



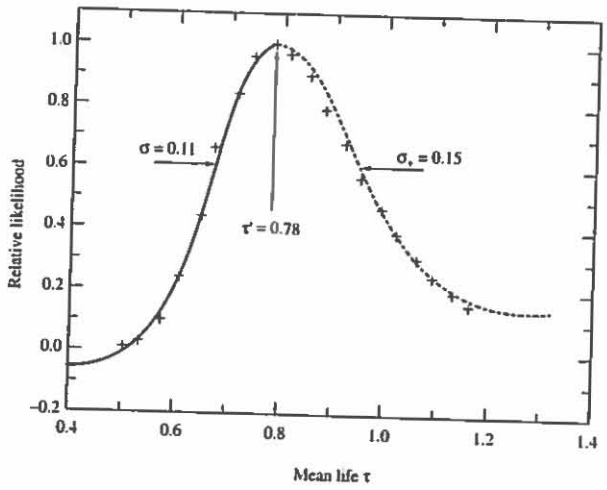


**FIGURE 11.1**
Relative values of the likelihood function versus trial values of the parameter for the 373-event sample of Example 10.1d. The data points (from Figure 10.4b) are indicated by crosses; the solid and dashed curves represent the results of fitting Gaussian curves separately to the two sides of the distribution. Parameters determined in the two fits are indicated on the graph. All measurements are in units of $10^{-10}$ s.

symmetrical about the mean. The uncertainties in our measurements may not be symmetrical, although the asymmetry may be hidden if we assume Gaussian statistics in our calculations. For example, the routines for finding uncertainties in parameters found by least-squares fitting (Chapters 7 and 8) generally assume a Gaussian distribution of the parameters and hence produce a single number for the uncertainties.

**Example 11.4.** As an example of an asymmetrical probability distribution, consider the 373-event data sample of Example 10.1d. In Figure 10.4b we plot as crosses the scaled values of the likelihood function for these data as a function of trial values of the parameter $\tau$. The data points exhibit a marked asymmetry about the mean $\tau'$. The dashed curve was calculated from Equation (10.10) with parameters obtained from the fit.

To make a better determination of $\sigma$ from this curve, we considered the regions on each side of the mean separately and estimated two separate standard deviations, $\sigma_L$ and $\sigma_R$, with the aid of Equation (1.11). To reduce the effect of the right-hand side tail on the value of $\sigma_r$, we imposed a cutoff at $\tau = 1.6$ and used only those data points below the cutoff in this calculation.

A composite curve formed of two Gaussians with the same mean $\tau$ but different values of $\sigma$ is shown as the solid curve in Figure 11.1. It would be reasonable to consider the two values of $\sigma$ obtained in this way as appropriate estimates of

the uncertainty in $\tau$, so that we could report $\tau' = 0.78^{+0.15}_{-0.11}$, as indicated by the arrows on Figure 11.1 rather than $\tau' = 0.78 \pm 0.14$ as we did in Chapter 10. This is equivalent to finding the two positions at which the logarithm of the likelihood function has decreased by $\Delta M = \frac{1}{2}$ as discussed in Section 10.2. Clearly this result is somewhat subjective if either side of the curve does not follow the Gaussian form. For this example, the value of $\sigma_R$ depends on how much of the tail is included in the calculation.

## Confidence Levels for Multiparameter Fits

The definition of the confidence level in a one-parameter experiment is generally straightforward. We can plot our data and observe if the distribution is Gaussian and estimate directly from the distribution of the probability that the true result lies between two specified values. When two or more variables have been determined and those variables exhibit some correlation, the definition of the confidence level becomes a little more difficult. Consider, for example, the determination of the mean lifetimes $\tau_1$ and $\tau_2$ of two unstable silver isotopes of Example 8.1. The problem was treated in Chapter 8 as a five-parameter problem, with parameters $a_4$ and $a_5$ corresponding to the two mean lifetimes, $\tau_1$ and $\tau_2$, respectively, and parameters $a_1$, $a_2$, and $a_3$ corresponding to the amplitudes of a uniform background and the two decaying states. The parameters of most interest in the experiment are $a_4$ and $a_5$, and we want to define a joint confidence interval for those two variables.

Figure 11.2 shows two sets of contours for the variation of $\chi^2$ as a function of $a_4$ and $a_5$ from the least-squares fit by the Marquardt method discussed in Chapter 8. The small contours, drawn with solid lines, were calculated by holding the parameters $a_1$, $a_2$, and $a_3$ fixed at their optimum values (see Table 8.5) and varying $a_4$ and $a_5$ to obtain increases in $\chi^2$ of 1, 2, and 3 from the minimum value. The large contours, shown as dashed lines, were calculated by allowing $a_1$, $a_2$, and $a_3$ to vary to minimize $\chi^2$ for each pair of values of $a_4$ and $a_5$. The contour plots cover very different ranges because of the correlations of the displayed parameters, $a_4$ and $a_5$, with the remaining parameters $a_1$ through $a_3$. The tilt of the closed figures on each plot indicates the degree of correlation of parameters $a_4$ and $a_5$ with each other. In an ideal experiment, the contours are ellipses in the region of the $\chi^2$ minimum and if $a_4$ and $a_5$ are not correlated, then, with suitable scaling of the axes, the ellipses are circles.

Which plot should we use? Additionally, how do we determine a confidence interval; that is, a region of the $a_4$-$a_5$ space in which we estimate there is, for example, a ~68% probability of finding the true values of the two parameters?

First, we should note that, because the fitting function, Equation (8.2), is not linear in the parameters, the methods of testing described in the previous sections strictly do not apply. However, we are much more likely to run into nonlinear fitting problems than the easier linear problems, so we shall continue with this example. At any rate, the function is linear in parameters $a_1$ through $a_3$, and we could make a linear expansion of it, over a limited region, in the parameters $a_4$ and $a_5$. In fact, this was the basis of a method of fitting nonlinear functions in Chapter 8.

Then, we should use the larger of the two contour diagrams to define our confidence intervals. That implies that if we wish to find the standard deviation of $a_4$



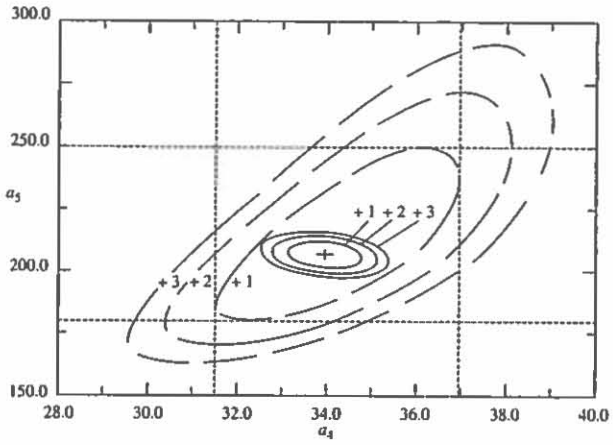


**FIGURE 11.2**
Two sets of contours for the variation of $\chi^2$ with parameters $a_4$ and $a_5$ in the region of the $\chi^2$ minimum. Data are from the least-squares fit by the Marquardt method discussed in Chapter 8. The small contours, drawn with solid lines, were calculated by holding parameters $a_1$ through $a_3$ fixed at their optimum while varying $a_4$ and $a_5$ to obtain increases in $\chi^2$ of 1, 2, and 3 from the minimum values. The large contours, shown as dashed lines, were calculated by allowing $a_1$, $a_2$, and $a_3$ to vary to minimize $\chi^2$ for each pair of values of $a_4$ and $a_5$.

from the contour plot, we should consider the full range of the outer limit of the $\Delta\chi^2 = 1$ contour, and not the intersection of that contour with the $a_4$ axis. This is equivalent to allowing $a_5$ to assume its best values for each chosen value of $a_4$, as we have already assumed for the parameters $a_1$ through $a_3$. The two dashed vertical lines indicate the two limits on $a_4$ that include the 1 standard deviation, or 68.3% of the probability, and the two horizontal lines indicate the 1 standard deviation limits for $a_5$.

How do we know that the vertical lines enclose 68.3% of the probability? By allowing the four parameters $a_1$, $a_2$, $a_3$, and $a_5$ to find their optimum values for each chosen value of $a_4$ and varying $a_4$, we have separated our $\chi^2$ fitting problem into two parts: a fit of $N$ data points to $m - 1$ parameters with $N - m - 1$ degrees of freedom and a variation of $\Delta\chi^2$ with $a_4$ about the minimum $\chi^2$, with 1 degree of freedom. As we observed in the previous section, the two variations separately must follow their appropriate $\chi^2$ distributions, so our variation of $\Delta\chi^2$ obeys the $\chi^2$ probability distribution for 1 degree of freedom. If we look at the integrated probability distribution $P_\chi$ for 1 degree of freedom [Table C.4, or calculated from Equation (11.10)], we see that $\chi^2 \geq 1$ corresponds to 31.7% of the probability, or $\Delta\chi^2 < 1$ corresponds to 68.3%. Similarly, if we wish to find the limits for 2 standard deviations, we should find the limits of $a_4$ on the $\Delta\chi^2 = 4$ contour, with all other parameters optimized.

To find the 1 standard deviation region encompassed by the *joint variation* of two parameters, $a_4$ and $a_5$, with all other parameters optimized, we must draw the contour corresponding to that value of $\Delta\chi^2$ for 2 degrees of freedom that includes 68.3% of the probability. Referring again to Table C.4 or Equation (11.10), we find that we should draw the contour for $\Delta\chi^2 = 2.30$, and for the 2 standard deviation contour, we should choose $\Delta\chi^2 = 6.14$. Joint confidence intervals with more than two parameters are often of interest, but are difficult to display and are represented best by two-dimensional projections of contours for pairs of variables.

### Confidence Level for a Predicted Value

Suppose the predicted value of a physical quantity is $\mu = 1000.0$, and we have made a measurement and obtained the value $\bar{x} = 999.4 \pm 2.0$. At what confidence level is the predicted value consistent with our measurement? The question could be rephrased as, "What is the probability of obtaining from the predicted parent distribution a distribution that is as bad as the one we got, or worse?" Because the shape of the parent distribution was not predicted, but only the value of the mean, we must use our value of the standard deviation, $\sigma \simeq 2.0$, as an estimate of that of the parent distribution. If the distribution is known to follow Gaussian statistics, then the required confidence is twice the integral of the standard Gaussian probability function from $x = \delta$ to $\infty$, where $\delta = |\mu - \bar{x}|/\sigma = |1000.0 - 999.4|/2.0$.

Now, suppose that the predicted value was necessarily positive—an intensity, for example. Then, we might again assume a Gaussian distribution, but only for positive values of the variable $x$, and therefore our confidence integral becomes the integral of the standard Gaussian from $\delta$ to $\infty$. However, because the total probability must be normalized to 1, we again multiply the integral by 2 so that the probability or confidence level is the same for both problems.

The method of determining the confidence level thus depends on the type of problem as well as the probability function that is applicable to the problem. For distributions that are symmetrical about their means, such as the Gaussian distribution, we generally consider the probability of obtaining a result that is the specified number of standard deviations from the mean, without regard to sign, unless a particular sign is excluded by the physical problem. For distributions such as the chi-square and Poisson distributions, which are only defined for positive values of their arguments, it is conventional to find a "one-sided" probability as in the case of the $\chi^2$ distribution where we quote the probability of obtaining a value as large as or larger than the value we obtained for a given number of degrees of freedom.

## 11.6 MONTE CARLO TESTS

A Monte Carlo calculation can help us understand the statistical significance of our results and possibly obtain a better estimate of some of the parameters of the experiment. As a by-product, the Monte Carlo program may also help us identify biases in our analysis procedure.

Suppose, for example, that we have measured a quantity $x$ that is predicted to have a value $\mu$. From our experiment we obtain the value $\bar{x}$ for our estimate of $\mu$.



We want to find the probability of obtaining from a series of similar experiments a value $\bar{x}$ that differs from the predicted value $\mu$ by

$$\Delta x \geq |\mu - \bar{x}| \tag{11.53}$$

We can set up a Monte Carlo program to simulate our experiment and to generate events with the parameters predicted by the theoretical principle that we are testing and with the same cuts as those imposed by our experimental apparatus. Such a program can be quite complex, but it may already exist at the time of analysis, if, for example, a Monte Carlo program was written to help plan the experiment. Or it might be possible to use some geometric and kinematic quantities from the actual experiment and only generate those parts of each event that are affected by the parameters in question.

After the Monte Carlo program has been written and debugged, we can simulate repeated experiments with the same parent parameters and the same number of final measurements as in our real experiment. The data from each of these simulated experiments can be processed by our regular analysis program to obtain a group of "experimental" values of $\bar{x}$, and from the distribution of these values we can estimate the required probability.

**Example 11.5.** Let us use the Monte Carlo method to try to learn more about the significance of the small peak in our data of Example 9.2. Examination of Figure 9.2 leaves no doubt about the existence of a large peak at ~1.0 GeV. Without the fitted curve, the smaller peak near 0.8 GeV would be considerably less striking and further analysis might be helpful. (We note that, if the small peak were indeed spurious, we should have to refit the large peak to obtain a better estimate of its mean energy and width.) In Chapter 9, we estimated the probability to be about 0.01% that the smaller peak is just a fluctuation in a single bin above the single-peak background, with a probability of about 0.6% of such a fluctuation occurring in any one of the 60 bins into which the data were sorted. These are quite compelling numbers. Can we support them with a more detailed calculation by the Monte Carlo method?

We adapted to the study of this problem the Monte Carlo program and the least-squares fitting program, which were used to generate and analyze the data in Chapter 9. With the Monte Carlo program, we simulated the experiment according to Equation (9.1) to generate 4000 single-peak events in each of 1000 trial "experiments." The mean energy ($E_0$), half-width ($\Gamma$), and amplitude of the larger peak, and the amplitudes ($a_1$ through $a_3$) of the quadratic background, were set to the values obtained in the six-parameter fit, listed in Table 9.1.

To each set of trial data we fitted Equation (9.13), using identical procedures to those used in Chapter 9, with the exception that, starting values for the parameters of the smaller peak ($a_7$, $a_8$, and $a_9$) were set to the values obtained in the nine-parameter fits of Chapter 9, listed in column 6 of Table 9.1. We selected those fits that yielded parameters of the lower peak consistent with the values determined in Chapter 9 by imposing the following conditions: (1) We required that both the chi-square probability and the amplitude of the smaller peak ($a_7$) be equal to or greater than the corresponding fitted values listed for the nine-parameter fit in Table 9.1; (2) We required that the central energy of that peak be within plus or minus one histogram bin (0.05 GeV) of the values obtained in that fit.

From the 1000 generated experiments, 5 survived these cuts, or 0.5% of the total trials. This number considerably exceeds the rough estimate of 0.01% made in

**TABLE 11.1**
**Results of generating 4000-event "experiments" from Equations (9.1) and (9.13) with parameters from fits listed in Table 9.1. We used several values of the amplitude $A_1$ of the smaller peak to test the sensitivities of our analysis to small and possibly spurious peaks.**

| $A_1$ | Equation | Number of experiments | Number of successes |
|---|---|---|---|
| 3.50 | 9.13 | 100 | 61 |
| 1.75 | 9.13 | 100 | 18 |
| 0.875 | 9.13 | 100 | 5 |
| 0.000 | 9.1 | 1000 | −5 |

Chapter 9 for a single bin fluctuation. Tests made with other starting values and cuts for the smaller peak yielded similar numbers of survivors.

To check our procedure, we also generated and analyzed 100 two-peak trial "experiments" from Equation (9.13), with the parameters of the smaller peak set to the values from the nine-parameter fit listed in Table 9.1. From these 100 trials, 61, or 61%, survived the cuts. When we repeated the analysis with the amplitude of the smaller peak reduced by a factor of 2 (i.e., $a_4/2$), the success rate dropped to 18%, and a further reduction by another factor of 2 ($a_4/4$) reduced the success rate to 5%. The results of analyses are summarized in Table 11.1.

These results offer strong support for the existence of the smaller peak, and indicate that in a 4000-event experiment we might detect with reasonable probability a peak with only one-fourth the amplitude of the current smaller peak. Clearly, a Monte Carlo simulation should play an important role in planning this type of experiment. A carefully planned Monte Carlo program may be much better (and easier) than a detailed theoretical analysis for finding an answer to the question "How much data will be needed to establish (or disprove) the existence of a specified feature in a distribution."

We offer a final word of caution on using the Monte Carlo technique to study the statistical significance of experimental results. For Examples 9.2 and 11.5, we used a very simple problem to illustrate this technique. Yet, there are many opportunities for errors, which can lead to erroneous conclusions about the significance of our Chapter 9 data. In a larger study, it would be very easy to make a simple mistake that might lie undetected in the program and have a subtle effect on the results. It is important to test the program under a variety of conditions, and to examine results at intermediate stages before drawing conclusions from the result. In particular, if the results of the program lead to conclusions that violate intuition about the experiment, we should check and recheck the calculation. The Monte Carlo method is very powerful, and can enable us to solve very difficult statistical problems in a straightforward manner, but like all powerful tools, it must be used with care.



## SUMMARY

*Variance of the fit*:

$$s^2 = \frac{1}{N-M}\frac{\Sigma[(1/\sigma_i^2)[y_i - y(x_i)]^2\}}{(1/N)\Sigma(1/\sigma_i^2)} = \frac{1}{N-m}\sum w_i[y_i - y(x_i)]^2$$

*Weighting factors*:

$$w_i = \frac{1/\sigma_i^2}{(1/N)\Sigma(1/\sigma_i^2)}$$

*Relationship between $s^2$ and $\chi^2$*:

$$\chi_\nu^2 = \frac{\chi^2}{\nu} = \frac{s^2}{(\sigma_i^2)}$$

where

$$(\sigma_i^2) = \left[\frac{1}{N}\sum\frac{1}{\sigma_i^2}\right]^{-1}$$

*Probability $P_\chi(\chi^2; \nu)$ that any random set of N data points will yield a value of chi-square as large as or larger than $\chi^2$*:

$$P_\chi(\chi^2; \nu) = \int_{\chi^2}^{\infty}\frac{z^{1/2(\nu-2)}e^{-z/2}}{2^{\nu/2}\Gamma(\nu/2)}\,dz$$

*Linear-correlation coefficient*:

$$r \equiv \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{[N\Sigma x_i^2 - (\Sigma x_i)^2]^{1/2}[N\Sigma y_i^2 - (\Sigma y_i)^2]^{1/2}}$$

*Probability $P_c(r, N)$ that any random sample of uncorrelated experimental data points would yield an experimental linear-correlation coefficient as large as or larger than $|r|$*:

$$P_c(r; \nu + 2) = \int_{|r|}^{1}\frac{1}{\sqrt{\pi}}\frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)}(1 - r^2)^{(\nu-2)/2}$$

*Sample covariance*:

$$s_{jk}^2 \equiv \frac{1/(N-1)\Sigma[(1/\sigma_i^2)(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)]}{(1/N)\Sigma(1/\sigma_i^2)} \quad \text{with} \quad \bar{x}_j = \frac{\Sigma(x_{ij}/\sigma_i^2)}{\Sigma(1/\sigma_i^2)}$$

*Sample variance*: $\sigma_j^2 = \sigma_{jj}^2$
*Sample linear-correlation coefficient*:

$$r_{jk} = \frac{s_{jk}^2}{s_j s_k}$$

*Multiple-correlation coefficient*:

$$R^2 \equiv \sum_{j=1}^{n}\left(b_j\frac{s_{jy}^2}{s_y^2}\right) = \sum_{j=1}^{n}\left(b_j\frac{s^j}{s_y}r_{jy}\right)$$

*F test:*

$$F = \frac{\chi^2_{\nu 1}}{\chi^2_{\nu 2}}$$

$$P_F(F; \nu_1, \nu_2) = \int_F^\infty p_f(f; \nu_1, \nu_2)\, df$$

*F test for multiple-correlation coefficient R (for $\nu = N - m$):*

$$F_R = \frac{R^2/(m-1)}{(1-R^2)/(N-m)} = \frac{R^2}{(1-R^2)} \times \frac{(N-m)}{(m-1)}$$

*F test for $\chi^2$ validity of adding $(m+1)$th term:*

$$F_\chi = \frac{\chi^2(m) - \chi^2(m+1)}{\chi^2(m+1)/(N-m-1)} = \frac{\Delta\chi^2}{\chi^2_\nu}$$

*Confidence limits:* $1\sigma \longrightarrow 68.3\%$; $2\sigma \longrightarrow 95.4\%$; $3\sigma \longrightarrow 99.7\%$

## EXERCISES

**11.1.** Discuss the meaning of $\chi^2$ and justify the relationship between it and the sample variance $s^2 = \chi^2_\nu$.

**11.2.** Compare the exact calculation of the gamma function $\Gamma(n)$ of Equation (11.7) with the approximate calculation of Equation (11.8) for $n = ½, 1, 1½, 4, 9½, 10$.

**11.3.** From Equation (11.6), show that the $\chi^2$-probability density for 1 degree of freedom can be written as

$$p(x^2) = \frac{e^{-x^2/2}}{\sqrt{2\pi x^2}}$$

Calculate to 1% the probability of obtaining a value of $\chi^2$ that is less than 2.00 by expanding the function in a Taylor series and integrating term by term.

**11.4.** For a typical number of degrees of freedom ($\nu \simeq 10$), find, by numerically integrating Equation (11.6), the range of probability $P_\chi(\chi^2, \nu)$ for finding $\chi^2$ as small as 0.5 or as large as 1.5. Use the approximation for the gamma function of Equation (11.8).

**11.5.** By numerically integrating Equation (11.6), find the probability of finding a value of $\chi^2_\nu = 1.5$ with $\nu = 100$ degrees of freedom. (Note that double-precision variables must be used.) Would you consider this to be a reasonably good fit?

**11.6.** Express the linear-correlation probability density of Equation (11.18) in terms of the approximation for the gamma function of Equation (11.8).

**11.7.** Work out the details of the calculation of the linear-correlation coefficients $r$ for Examples 6.1 and 6.2.

**11.8.** If a set of data yields a zero slope $b = 0$ when fitted with Equation (11.11), what can you say about the linear-correlation coefficient $r$? Justify this value in terms of the correlation between $x_i$ and $y_i$.

**11.9.** Find the linear-correlation coefficient $r_1$ between the independent variable $T_i$ and the dependent variable $V_i$ for the data of Example 7.1.

**11.10.** Find the correlation coefficient $r_2$ between $T_i^2$ and $V_i$ for the data of Example 7.1. Does the correlation justify the use of a quadratic term?



**11.11.** Express the multiple correlation $R$ in terms of $x_{ij}$, $y_i$, and their averages.

**11.12.** Evaluate the multiple-correlation coefficient $R$ for the data of Example 7.1.

**11.13.** Is a large value of $F$ good or bad? Explain.

**11.14.** If we wish to set as an arbitrary criterion a probability of 0.01 for the $F_\chi$ test, what would be the reasonable average value for $F$ test?

**11.15.** What different aspects of a fit do the $F_R$ and $F_\chi$ tests represent?

**11.16.** Apply the $F_\chi$ test for the quadratic term to the data of Example 7.1 and state your conclusions. (Refer to Table 7.4.)

**11.17.** Show the intermediate steps in the derivation of Equation (11.43).

**11.18.** Estimate from Figure 11.2 the 90% confidence limit for each of the two mean lifetimes ($a_4$ and $a_5$) of Example 8.1 when all variables are allowed to find their optimum values.

# APPENDIX A

# NUMERICAL METHODS

There are several reasons why we might want to fit a function to a data sample, and several different techniques that we might use. If we wish to estimate parameters that describe the parent population from which the data are drawn, then the maximum-likelihood or least-squares method is best. If we wish to interpolate between entries in data tables to find values at intermediate points or to find numerically derivatives or integrals of tabulated data, then an interpolation technique will be more useful. Additionally, if we wish to obtain intermediate values between calculated coordinate pairs in order to plot a smooth curve on a graph, then we may wish to use a spline fitting method. In this appendix we shall summarize some standard methods for treating the latter two types of problems, as well as some methods of finding the roots of nonlinear functions, a different sort of interpolation problem.

## A.1 POLYNOMIAL INTERPOLATION

With modern fast computers, the need for interpolating within tables to find intermediate values of tabulated functions has reduced markedly. Nevertheless, there are situations in which it may be convenient to represent a complicated function by a simple approximation over a limited range. For example, in a large Monte Carlo calculation, where computing time is a significant consideration, we may approximate a complex function by a simpler polynomial that can be calculated quickly. Alternatively, we may save time by creating a probability integral once at the beginning of the program. and interpolating to find values of $x$ corresponding to the randomly chosen values of $y$.

For many purposes a linear or quadratic interpolation is satisfactory; that is, we fit a straight line to two coordinate pairs, or a parabola to three, and use the equation of the fitted polynomial to find values of $y$ at nearby values of $x$. Higher orders may be necessary for functions that have strong variations, but in general, it





is better and more convenient to represent a function over a limited region by a series of low-order approximations.

### Lagrange's Interpolation Method

Here is a method that is easy to remember and can be used to expand a function to any order. We know it works because of the theorem that states that if you can find any $n$th-degree polynomial that passes exactly through $n + 1$ points, then you have found the one and only $n$th-degree polynomial that passes through those points. Think about it. It is obvious for $n = 1$ (2 points).

Let us start with an easy problem. Suppose we have two coordinate pairs $(x_0, y_0)$ and $(x_1, y_1)$, and we want to find the straight line that passes through both of them. We write a function of the form

$$P(x) = y_0 A_0(x) + y_1 A_1(x) \tag{A.1}$$

and search for a function $A_0(x)$ that is 1 when $x = x_0$ and 0 when $x = x_1$, and a function $A_1(x)$ that is 1 when $x = x_1$ and 0 when $x = x_0$. We can guess the form. If we write $A_0(x)$ as a fraction and set its numerator to $(x - x_1)$, then $A_0(x)$ will be 0 for $x = x_1$ and will be $(x_0 - x_1)$ for $x = x_0$. But we want $A_0(x) = 1$ for $x = x_0$, so the denominator of $A_0$ must be $(x_0 - x_1)$. We can make similar arguments for $A_2(x)$ and thus write as our interpolation equation

$$P(x) = y_0 \frac{(x - x_1)}{(x_0 - x_1)} + y_1 \frac{(x - x_0)}{(x_1 - x_0)} \tag{A.2}$$

Suppose we want a parabola that passes through three points. Then we simply write

$$P(x) = y_0 A_0(x) + y_1 A_1(x) + y_2 A_2(x) \tag{A.3}$$

and, following the previous arguments, write

$$P(x) = y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)}{(x_1 - x_0)} \frac{(x - x_2)}{(x_1 - x_2)} + y_2 \frac{(x - x_0)}{(x_2 - x_0)} \frac{(x - x_1)}{(x_2 - x_1)} \tag{A.4}$$

The expansion to higher orders should be obvious. The $k$th term in an $n$th order expansion is given by the following product in which the $j = k$ term must be omitted:

$$\prod_{j=0}^{n} \frac{(x - x_j)}{(x_k - x_j)} y_k \quad (\text{excluding } j = k) \tag{A.5}$$

Note that the intervals in $x$ need not be equally spaced. The interpolation for a well-behaved function $y = f(x)$ is completely general.

### Newton's Divided Differences

Although the Lagrange interpolation method is especially easy to derive and provides a convenient way of interpolating between points in a function or table, it is

not very convenient for repetitive calculations. It is not very convenient as an expansion either, because increasing the order of the expansion requires adding another factor to each term as well as adding another term. What we require is a more familiar form—a discrete analog of the Taylor expansion. For this we turn to Newton's method of divided differences.

There are several forms of the divided differences expansion, roughly characterized by the method we choose to define the differences, forward, backward, or about a central point. We shall restrict ourselves here to forward differences; that is, we calculate the variation of $y$ with respect to $x$ by taking increments in the positive $x$ direction.

Again, consider a set of data points, $(x_0, y_0)$, $(x_1, y_1)$, $(x_2, y_2)$, . . . . Let us assume that we wish to make a linear interpolation from $x_0$ to some point x with a first-degree polynomial. We define the zeroth divided difference as the function itself $f(x)$ evaluated at $x = x_0$:

$$f[x_0] \equiv f(x_0) = y_0 \tag{A.6}$$

The first divided difference is defined to be

$$f[x_0, x_1] \equiv \frac{f[x_1] - f[x_0]}{(x_1 - x_0)} \tag{A.7}$$

which is the slope of a linear function. Then, for a linear function,

$$f[x, x_0] = f[x_0, x_1] \tag{A.8}$$

or

$$\frac{f[x_0] - f[x]}{(x_0 - x)} = \frac{f[x_1] - f[x_0]}{(x_1 - x_0)} \tag{A.9}$$

which, on rearrangement of the terms, gives the first-order expansion

$$\begin{aligned} P_1(x) &= f[x_0] + (x - x_0)\frac{f[x_1] - f[x_0]}{(x_1 - x_0)} \\ &= f[x_0] + (x - x_0) f[x_0, x_1] \end{aligned} \tag{A.10}$$

where we have written $P_1(x)$ instead of $f(x)$ to indicate that the expansion is a polynomial approximation to the function $f(x)$.

To find the second-order expansion, we consider the second divided differences

$$f[x_0, x_1, x_2] \equiv \frac{f[x_2, x_1] - f[x_1, x_0]}{(x_2 - x_1)(x_1 - x_0)} \tag{A.11}$$

which corresponds to the slope of the slope, or the second derivative. This must be constant for a second-order function, so we have

$$f[x, x_0, x_1] = f[x_0, x_1, x_2] \tag{A.12}$$

which leads to the second-order expansion

$$P_2(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \tag{A.13}$$

The general form for the $n$th-order expansion should again be obvious.



## Remainders

The extrapolation formula for an $n$th-order expansion is only exact when the function itself is an $n$th-degree polynomial. Otherwise, the *remainder* at $x$ after $n$ terms $R_n(x)$, defined as the difference between the original function $f(x)$ and the expansion $P_n(x)$, is given by

$$\begin{aligned} R_n(x) &= f(x) - P_n(x) \\ &= (x - x_0)(x - x_1) \cdots (x - x_n) f[x, x_0, x_1, \ldots, x_n] \end{aligned} \tag{A.14}$$

Calculation of the remainder requires the value of the function $f(x)$ at $x$, which is generally not available. (If it were, we might not be doing this expansion.) However, it may be possible to make an estimate of $f_n(x)$, or to use a nearby value, and thus find an estimate of $R_n(x)$. An expression for the remainder can also be obtained in terms of the $(n + 1)$th derivative of the function.[1]

## Uniform Spacing

The divided difference expressions have a particular convenient form when the intervals in $x$ are uniform; that is, if $x_2 - x_1 = x_3 - x_2 = x_i - x_{i-1} = h$. The divided difference of the previous discussion can be written

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{(x_1 - x_0)} = \frac{\Delta f(x_0)}{h}$$

or

$$\Delta f(x_0) \equiv f(x_1) - f(x_0) \quad \text{and} \quad h = x_1 - x_2 \tag{A.15}$$

and higher-order differences become

$$\Delta^2 f(x_0) \equiv \Delta[\Delta f(x_0)] = \Delta f(x_1) - \Delta f(x_0), \text{ etc.} \tag{A.16}$$

If we define the relative distance along the interval by

$$\alpha = (x - x_0)/h \tag{A.17}$$

we can write for the $n$th-order expansion,

$$\begin{aligned} P_n(x) &= f(x_0) + \alpha \Delta f(x_0) + \alpha(\alpha - 1)\Delta^2 f(x_0)/2! + \cdots \\ &\quad + \alpha(\alpha - 1) \cdots (\alpha - n - 1)\Delta^n f(x_0)/n! \end{aligned} \tag{A.18}$$

Equation (A.18) is a finite difference analog of the familiar Taylor expansion with the important difference that the factors multiplying the coefficients $\Delta^i f(x_0)/n!$ are not successive powers of the relative distance from the starting point, but rather the product of relative distances from successive points used in the expansion, because $(\alpha - 1) = (x - x_0 - h)/h = (x - x_1)/h$, and so forth.

## Extrapolation

Equations (A.15) through (A.18) are perfectly general for fitting exactly $n$ sequential equally spaced data points with a polynomial of degree $n - 1$. In principle, the

[1] See Hildebrand (1956) for a derivation.

**TABLE A.1**
**Uniform differences for cos θ**

| θ (degrees) | $y$ | $\Delta_1$ | $\Delta_2$ | $\Delta_3$ | $\Delta_4$ | $\Delta_5$ |
|---|---|---|---|---|---|---|
| 0 | 1.0000 | −0.0489 | −0.0931 | 0.0139 | 0.0078 | −0.0021 |
| 18 | 0.9511 | −0.1420 | −0.0792 | 0.0217 | 0.0056 | |
| 36 | 0.8090 | −0.2212 | −0.0575 | 0.0273 | | |
| 54 | 0.5878 | −0.2788 | −0.0302 | | | |
| 72 | 0.3090 | −0.3090 | | | | |
| 90 | −0.0000 | | | | | |

**TABLE A.2**
**Extrapolation from 0 to 10° and from 0 to 75° in various orders**

| θ | | Order | | | | |
|---|---|---|---|---|---|---|
| (degrees) | cos θ | 1 | 2 | 3 | 4 | 5 |
| 10 | 0.9848 | 0.9728 | 0.9843 | 0.9851 | 0.9848 | 0.9848 |
| 75 | 0.2588 | 0.7961 | 0.1819 | 0.2481 | 0.2589 | 0.2588 |

position of the first data point $(x_0, y_0)$ can be anywhere, but for optimum interpolation, the values of $x_0$ and $x_n$ should straddle the interpolation point $x$ and be approximately equidistant from it.

The same formula can be used for extrapolating to values beyond the region of data, but the uncertainties in the validity of the approximation increase as $x$ gets farther from the average of $x_1$ and $x_n$. The approximation is limited by both the degree of the interpolating polynomial and by uncertainties in the coefficients of the polynomial resulting from fluctuations in the data.

**Example A.1.** Table A.1 shows a uniform divided difference table for the cosine function for a range of the argument θ between 0 and 90°. Table A.2 shows values of cos θ for θ = 10 and 75° calculated from the divided difference table in orders 1 through 5. The interpolation starts at 0° so that only the top row of Table A.1 is used and thus, θ > 18°, the calculation is an extrapolation. The true value of cos θ is also listed. As we should expect, the large extrapolation to 75° is very poor in low order. Usually, an approximation can be improved by increasing the number of terms in the expansion. However, the better method would be to drop to a different line of the table; that is, to ensure that the calculation is an interpolation rather than an extrapolation.

## A.2. BASIC CALCULUS: DIFFERENTIATION AND INTEGRATION

Let us review some basic principles of differential calculus before considering discrete methods that are applicable to computer calculations.



### Differentiation

Let $f(x)$ be a function of the variable $x$. If $x$ increases by an amount $\Delta x$, the function varies by an amount $\Delta f = f(x + \Delta x) - f(x)$. The ratio $\Delta f/\Delta x$ is a measure of the relative variation of $f(x)$ with $x$. In the limit, as $\Delta x$ becomes infinitesimally small, the ratio $\Delta f/\Delta x$ for a continuous function $f(x)$ approaches an asymptotic value, the *derivative* $df/dx$ of the function $f(x)$ with respect to $x$.

$$\frac{df(x)}{dx} \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{A.19}$$

The derivative of $f(x)$ at $x = x_0$ is written $\frac{df(x_0)}{dx}$ and corresponds to the slope of the function evaluated at $x_0$ or the tangent to the curve at that point.

**Example A.2** To find the derivative of $f(x) = x^n$, we can expand the function $f(x + \Delta x)$ to first order in a Taylor series.

Thus, with $n = 4$, we have $f(x) = x^4$ and $df/dx = 4x^3$.

$$\frac{d(x^n)}{dx} = \lim_{\Delta x \to 0} \frac{(x^n + nx^{n-1}\Delta x) - x^n}{\Delta x}$$
$$= \frac{nx^{n-1}\Delta x}{\Delta x} = nx^{n-1}$$

**Example A.3** For $f(x) = \sin x$, we can write

$$\sin(x + \Delta x) = (\sin x)(\cos \Delta x) + (\sin \Delta x)(\cos x)$$

and again expand $f(x)$ to obtain

$$\frac{d(\sin x)}{dx} = \lim_{\Delta x \to 0} \frac{\sin(x + \Delta x) - \sin x}{\Delta x}$$
$$= \lim_{\Delta x \to 0} \frac{(\sin x)(\cos \Delta x) + (\sin \Delta x)(\cos x) - \sin x}{\Delta x}$$
$$= \frac{\sin x + (\Delta x)(\cos x) - \sin x}{\Delta x} = \cos x$$

Similarly, for $f(x) = \cos x$, we find $df/dx = -\sin x$.

**SUMS AND PRODUCTS** The derivative of a sum of functions is equal to the sum of the derivatives of the individual functions. Consider the function

$$f(x) = g(x) + h(x)$$

The derivative of this function is the sum of the derivatives of the individual terms.

$$\frac{df(x)}{dx} = \frac{dg(x)}{dx} + \frac{dh(x)}{dx}$$

The derivative of a product of functions, however, is not equal to the product of the derivatives. Consider the function

$$f(x) = g(x) \times h(x)$$

We can rewrite Equation (A.19) as

$$\lim_{\Delta x \to 0} f(x + \Delta x) = \lim_{\Delta x \to 0} \left[ f(x) = \Delta x \frac{df(x)}{dx} \right] \tag{A.20}$$

and show that

$$\begin{aligned} \frac{d[g(x) \times h(x)]}{dx} &= \lim_{\Delta x \to 0} \frac{g(x + \Delta x)h(x + \Delta x) - g(x)h(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left\{ \left[ g(x) + \Delta x \frac{dg(x)}{dx} \right] \left[ h(x) + \Delta x \frac{dh(x)}{dx} \right] - g(x)h(x) \right\} \\ &= g(x) \frac{dh(x)}{dx} + h(x) \frac{dg(x)}{dx} \end{aligned}$$

**FUNCTIONS OF FUNCTIONS** If the function $f(x)$ can be expressed as a function of a function $g(x)$ of $x$,

$$f(x) = f[g(x)]$$

the derivative of $f(x)$ with respect to $x$ can be expressed in terms of the derivative of $g(x)$ with respect to $x$. If we expand the definition of Equation (A.19) for the derivative, we can make use of the relationship of Equation (A.20) to expand still further.

$$\begin{aligned} \frac{df(x)}{dx} &= \lim_{\Delta x \to 0} \frac{f\left[ g(x) + \Delta x \dfrac{dg(x)}{dx} \right] - f[g(x)]}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{f[g(x)] + \Delta x \dfrac{dg(x)}{dx} \dfrac{df(x)}{dg(x)} - f[g(x)]}{\Delta x} \\ &= \frac{df(x)}{dg(x)} \frac{dg(x)}{dx} \end{aligned} \tag{A.21}$$

**Example A.4** If $f(x) = (a - bx^3)^2$, define $g(x) = a + bx^3$ so that $f(x) = [g(x)]^2$. The first factor in Equation (A.21) is the derivative of a square, and the second factor is the derivative of a cubic polynomial.

$$\frac{df(x)}{dg(x)} = 2g(x) = 2(a + bx^3) \qquad \frac{dg(x)}{dx} = 3bx^2$$

$$\frac{df(x)}{dx} = 2(a + bx^3)3bx^2 = 6bx^2(a + bx^3)$$

**HIGHER-ORDER DERIVATIVES** Higher-order derivatives are defined as derivatives of derivatives. For example, the second derivative of a function $f(x)$ is just the derivative of the first derivative.



$$\frac{d^2f(x)}{dx^2} \equiv \frac{d}{dx}\left[\frac{df(x)}{dx}\right]$$

For the $n$th-order derivative $d^nf(x)/dx^n$, we simply take the derivative $n$ times in succession. For example, if $f(x) = x^4$ as in Example A.2, the second derivative is $12x^2$. Similarly, the fourth derivative of either $\sin x$ or $\cos x$ is equal to itself.

**PARTIAL DERIVATIVES** If the function $f(x, y)$ is dependent on two variables $x$ and $y$, we must define derivatives of the function with respect to each of the independent variables. To determine the *partial derivative* of $f$ with respect to $x$, $\partial f/\partial x$, we consider that $y$ is a constant and proceed as we would for an ordinary derivative. Similarly, to determine the partial derivative $\partial f/\partial y$ we consider that $x$ is constant.

$$\frac{\partial f(x, y)}{\partial x} \equiv \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} = \frac{df(x)}{dx}$$

$$\frac{\partial f(x, y)}{\partial y} \equiv \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} = \frac{df(y)}{dy}$$

Higher-order partial derivatives include not only higher-order derivatives with respect to one variable, but also cross-partial derivatives with respect to two or more variables simultaneously.

$$\frac{\partial^2 f(x, y)}{\partial x^2} \equiv \frac{\partial}{\partial x}\left[\frac{\partial f(x, y)}{\partial x}\right]$$

$$\frac{\partial^2 f(x, y)}{\partial x\, \partial y} \equiv \frac{\partial}{\partial x}\left[\frac{\partial f(x, y)}{\partial y}\right] = \frac{\partial}{\partial y}\left[\frac{\partial f(x, y)}{\partial x}\right] = \frac{\partial^2 f(x, y)}{\partial y\, \partial x}$$

**MINIMA AND MAXIMA** A function $f(x)$ is said to have a *local minimum* at $x = x_{min}$ if the values of $f(x_{min} \pm \Delta x)$ are larger than the value of $f(x_{min})$ for infinitesimal changes $\Delta x$ about $x_{min}$. Similarly, the function has a *local maximum* if the values of $f(x_{max} \pm \Delta x)$ are smaller than $f(x_{max})$. At either a minimum or a maximum of a function, the derivative of the function is zero,

$$\frac{df(x_m)}{dx} = 0$$

corresponding to a tangent that is parallel to the $x$-axis.

The question of whether the function is a minimum or a maximum at $x_m$ can be resolved by examining the second derivative. If the second derivative is positive, the curvature of the function is upward and $f(x_m)$ is a minimum. If the second derivative is negative, the $f(x_m)$ is a maximum.

**FUNCTIONS OF MORE THAN ONE VARIABLE** With functions of more than one variable, for example $f(x, y)$, we can still consider the function to have a minimum in parameter space, but we must be careful to assure that the function has a minimum simultaneously with respect to all parameters.

### Integration

Integration is the inverse of differentiation. To find the integral $F(x)$ of the function $f(x)$,

$$F(x) = \int f(x)dx$$

we must find a function $F(x)$ such that $\frac{dF(x)}{dx} = f(x)$.

However, this definition is not unique. An undetermined constant must be added to the solution to allow for the fact that the derivative of a constant is zero.

**Example A.5** Consider the integral of the function $f(x) = x^3$. We observe that $F(x) = x^4/4$ is a solution:

$$\frac{dF(x)}{dx} = \frac{d(x^4/4)}{dx} = x^3 = f(x)$$

However, $F(x) = x^4/4 + C$ is also a solution, where $C$ is any quantity that is not a function of $x$. Thus, the solution to an *indefinite integral* must include an added constant.

A *definite integral* is the integral of a function between two specific values of the independent variable, and is written

$$I = \int_a^b f(x)dx$$

To find the definite integral of a function, we integrate it, calculate the value of the integral at $x = b$ and at $x = a$, and find the difference between the two values. This is equivalent to calculating the area under the function $f(x)$ between the two limits $a$ and $b$.

**Example A.6** Consider the integral of the function $f(x) = x^3$ between the limits $x = 1.0$ and $x = 2.0$.

$$I = \int_{1.0}^{2.0} f(x)dx = \int_{1.0}^{2.0} x^3 dx = x^4/4\Big|_{1.0}^{2.0} = [2^4 - 1^4]/4 = 15/4$$

Note that a definite integral is not a function of variable of integration $x$.

From the results of Example A.3,

$$\int \sin x\,dx = -\cos x + C \qquad \text{and} \qquad \int \cos x\,dx = \sin x + C$$

## A.3. NUMERICAL DIFFERENTIATION AND INTEGRATION

With the interpolation expressions discussed in Section A.1, it is relatively straightforward to obtain expressions for derivatives and integrals in terms of expansions to order $n$.



### Differentiation

We can differentiate Equation (A.18) to find approximations for the derivatives of the function $f(x)$. We obtain

$$\frac{dP_n(x)}{dx} = \frac{1}{h}\frac{dP_n(x)}{d\alpha} = [\Delta f(x_0) + (2\alpha - 1)\Delta^2 f(x_0)/2! + (3\alpha^2 - 6\alpha + 2)\Delta^3 f(x_0)/3! + \cdots]/h \tag{A.22}$$

and

$$\frac{d^2P_n(x)}{dx^2} = \frac{1}{h^2}\frac{d}{d\alpha}\left[\frac{dP_n(x)}{d\alpha}\right] = [\Delta^2 f(x_0) + (\alpha - 1)\Delta^3 f(x_0) + \cdots]/h^2 \tag{A.23}$$

We should note that the use of forward differences introduces an asymmetry in the calculation. For a general solution, we could replace the forward differences by central differences, which are taken symmetrically about a central starting point. For a particular problem, we can usually arrange the expansion to provide reasonable symmetry of the differences about the point of interest. Thus, we can replace Equations (A.22) and (A.23) by

$$\frac{dP_n(x)}{dx} = \Delta f(x_0)/h = \frac{f(x + h/2) - f(x - h/2)}{h} \tag{A.24}$$

and

$$\frac{d^2P_n(x)}{dx^2} = \Delta^2 f(x_0)/h^2 = \frac{f(x + h) - 2f(x) + f(x - h)}{h^2} \tag{A.25}$$

### Integration

Integrating Equation (A.18) leads to expressions for calculating the numerical integral in various orders, depending on the number of terms in the polynomial approximation. There are various forms for each order, depending on how we choose the limits of integration. We quote three of the most useful forms with the remainder estimates.

First-order, endpoint trapezoidal

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f^{(2)}(\xi)$$

(first-order closed-end trapezoidal)

$$\int_{x_0}^{x_2} f(x)\,dx = 2hf(x_1) + \frac{h^3}{3}f^{(2)}(\xi) \quad \text{(first-order open end)}$$

$$\int_{x_0}^{x_2} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90}f^{(4)}(\xi)$$

(second-order closed-end Simpson's rule)

The factors $f^{(n)}(\xi)$ in the remainder estimates represent the $n$th derivative of the function evaluated at some (unknown) value of x in the range of the integral.

Note the large reduction on the error estimate in going from either of the first-order approximations to the second-order approximation.

For an integral over an extended range of $x$, it is usually advisable to employ a series of first- or second-order integrals over sections of the function, rather than to attempt to fit a large region with a higher-order function. In fact, it can be shown that the gain in accuracy in going from a second- to a third-order numerical integral is relatively small, and, for the same number of calculations of the ordinate $y_i$, the second-order Simpson rule may be more accurate than the third-order form. This relation applies in general to even and odd orders, so that, to make a significant improvement in the numerical integration of a function, one should advance to the next higher *even* order.

Thus, to find the integral by Simpson's rule of $f(x)$ over an extended range between $x = x_0$ and $x = x_n$, we divide the region into $n$ equal intervals in $x$, with $nh = (x_n - x_0)$, to obtain

$$\int_{x_0}^{x_n} f(x)\,dx = \frac{h}{3}[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 4f(x_{n-1}) + f(x_n)] - \frac{nh^5}{180} f^{(4)}(\xi) \qquad \text{(A.26)}$$

where $\xi$ is the value of $x$ somewhere in the range of integration.

**Program A.1** SIMPSON (Appendix E) calculates an extended integral by the second-order approximation of Equation (A.26). See Programs 11.1 and 11.2 for examples of the use of this routine.

The user supplies four arguments:

1. FUNCT: the name of the function to be integrated. The function must have one real argument. If other arguments are required, they must be made accessible to the function as global variables.
2. NINT: the number of *double* intervals. The interval is calculated as DX = (HILIM-LOLIM)\(2*NINT);
3. LOLIM and
4. HILIM: the integration limits.

## A.4 CUBIC SPLINES

If we attempt to represent by an $n$th-degree polynomial a function that is tabulated at $n + 1$ points, we are apt to obtain disappointing results if $n$ is large. The polynomial will necessarily coincide with the data points, but may exhibit large oscillations between points. In addition, if there are many data points, the calculations can become rather cumbersome. It is often better to make several low-order polynomial fits to separate regions of the function, and this procedure is usually satisfactory for simple interpolation in tables. However, if we want a smooth function, which passes through the data points, the results may not be satisfactory.

Suppose we have calculated a function at $n + 1$ points, and want to represent the function as a smooth curve on a graph. The $n$th-order polynomial is out—too wiggly. Breaking the curve up into small sections produces disjointed segments on the plot. It is unlikely that they will combine to form a smooth curve. What do we do now? Reach for our pencil and trusty drafting spline? No, we call up our spline fitting subroutine and let it join up the separate fits for us.



Spline fitting procedures have other uses besides plotting pretty curves on graphs, but the plotting function is of interest to us and is easily illustrated. Suppose we choose to make a series of cubic fits to successive groups of data points. What conditions do we need to produce a smooth curve that passes through the data points? We want the first and second derivatives, as well as the function itself, to be continuous at the data points. Suppose we consider a separate cubic polynomial for each interval on the graph, or a total of $n$ polynomials for the $n + 1$ points. Then we write the polynomial equation, take derivatives, and, at each data point, equate the first and second derivatives of the left-side polynomial to those of the right-side polynomial.

Following the method discussed in Thompson (1984), we begin by writing the Taylor series for the cubic polynomial for interval $i$, expanded about the point $x_i$

$$y(x) = y(x_i) + (x - x_i)\frac{dy(x_i)}{dx} + (x - x_i)^2 \frac{d^2y(x_i)}{dx^3}\Big/2! + (x - x_i)^3 \frac{d^3y(x_i)}{dx^3}\Big/3! \qquad \text{(A.27)}$$

where the function and derivatives are evaluated at $x_i$. This can be written in a more concise form as

$$y(x) = y_i + (x - x_i)y_i' + (x - x_i)^2 y_i''/2 + (x - x_i)^3(y_{i+1}'' - y_i'')/6h \qquad \text{(A.28)}$$

where $y_i'$ and $y_i''$ stand for the first and second derivatives evaluated at $x = x_i$ and the third derivative has been replaced by its divided difference form, which is exact for a cubic function. At $x = x_i$, we have $y = y_i$, as required. We can also set $x = x_{i+1} = x_i + h$ and solve the equation

$$y(x_{i+1}) = y_i + (x_{i+1} - x_i)y_i' + (x_{i+1} - x_i)^2 y_i''/2 + (x_{i+1} - x_i)^3(y_{i+1}'' + y_i'')/6h \qquad \text{(A.29)}$$

to obtain

$$y_{i+1} - y_i = hy_i' + h^2[2y_i'' + y_{i+1}'']/6 \qquad \text{(A.30)}$$

We repeat the calculation, using the equation for $y(x)$ in interval $i - 1$ [i.e., we replace $i$ by $i - 1$ in Equation (A.29)],

$$y(x) = y_{i-1} + (x - x_{i-1})y_{i-1}' + (x - x_{i-1})^2 y_{i-1}''/2 + (x - x_{i-1})^3(y_i'' - y_{i-1}')/6h \qquad \text{(A.31)}$$

and again require that $y(x) = y(x_i)$ at the $i$th data point and obtain

$$y_i - y_{i-1} = hy_{i-1}' + h^2[2y_{i-1}'' + y_i'']/6 \qquad \text{(A.32)}$$

To establish the continuity conditions at the data points, we need the first derivative in the interval $i$,

$$y'(x) = y'_i + (x - x_i)y''_i + (x - x_i)^2(y''_{i+1} - y''_i)/2h \tag{A.33}$$

which we equate to the first derivative in the interval $i - 1$ at the position $x = x_i$, to obtain

$$y'_i - y'_{i-1} = h[y''_i + y''_{i-1}]/2 \tag{A.34}$$

Similarly, equating the derivatives at the boundary $x = x_{i+1}$ gives

$$y'_{i+1} - y'_i = h[y''_{i+1}y''_i]/2 \tag{A.35}$$

(Repeating the procedure with the second derivative leads to an identity, because our use of the divided difference form for the third derivative assures continuity of the second derivative across the boundaries.) Eliminating the first derivatives from Equations (A.30), (A.32), (A.34), and (A.35) gives us the spline equation

$$y''_{i-1} + 4y''_i + y''_{i+1} = D_i \tag{A.36}$$

with

$$D_i = y[y_{i+1} - 2y_i + y_{i-1}]/h^2 \tag{A.37}$$

Note that the $D_i$ are proportional to the second differences of the tabulated data and are all known. We can write Equation (A.36) as a set of linear equations relating the unknown variables $y''$, beginning with $i = 2$ and ending with $i = n - 1$:

$$y''_1 + 4y''_2 + y''_3 = D_2 \tag{A.38a}$$

$$y''_2 + 4y''_3 + y''_4 = D_3 \tag{A.38b}$$

$$\vdots$$

$$y''_{n-3} + 4y''_{n-2} + y''_{n-1} = D_{n-2} \tag{A.38c}$$

$$y''_{n-2} + 4y''_{n-1} + y''_n = D_{n-1} \tag{A.38d}$$

These equations can be solved for the second derivatives $y''_i$, as long as we know the values of $y''_1$ and $y''_n$. One possibility is to set the second derivatives to 0 to obtain *natural splines*. Alternatively, we may use the true second derivatives, if they are known, or a numerical approximation.

For example, suppose we have only four points to consider. Then, if we know $y''_1$ and $y''_4$, we can solve the simultaneous Equations (A.38a) and (A.38b) for $y_2$ and $y_3$. Similarly, if we have a full set of $n$ equations, we can rewrite Equation (A.38a) to express $y''_2 = (D_2 - y''_1 - y''_3)/4$, and substitute this expression into Equation (A.38b) to eliminate $y''_2$. Then, we repeat the procedure to eliminate $y''_3$ from the next equation. We continue this procedure until we reach the last equation, which will contain only terms in $y''_1$, $y''_{n-1}$, and $y''_n$. Because $y''_1$ and $y''_n$ are known, we can solve this equation for $y''_{n-1}$, and then work back down the chain determining successively $y''_{n-2}$, $y''_{n-3}$, and so forth, until we reach Equation (A.38a) from which we determine the last unknown $y''_2$. Once we have found the values of the $y''_i$, we can find the $y'_i$ from Equation (A.30) or (A.32), and use Equation (A.28) to interpolate in each interval.

The solution of Equation (A.38) is discussed in several textbooks. Essentially, one sets up recursion relations to build a table of the second derivatives $y''$. The method is illustrated by the computer routines SPLINEMAKE listed in Appendix E.



An interesting alternative method of solving the set of simultaneous equations, Equations (A.38), is to set them up in a spreadsheet program. Then, when the boundary values $y''_1$ and $y''_n$ are supplied, the program will readjust the variables until they stabilize at the solutions to Equations (A.38). Although this method is not very practical for graphical applications where we want to build the solution into our plotting program, it does provide a quick way of finding the second derivatives and an interesting illustration of the solution.

As with all techniques, a certain amount of care must be exercised in using spline routines. The choice of a second derivative at the boundary may have an important effect on the interpolation at the ends of the function, and a wrong choice, for example, can produce undesirable shapes at the edges of a plot. Then too, although the spline routine assures a smooth variation between the data points, with continuity of the function and first and second derivative across the points, it cannot guarantee that there will be no peculiar oscillation between the points.

**Program A.2 SPLINE INTERPOLATION (Appendix E)**
SPLINEMAKE numerically calculates a table of second derivatives for a spline interpolation by the method discussed in the previous paragraphs.
SPLINEINT performs the interpolation. For simplicity, we have chosen to store only the second derivatives and to calculate the first and third derivatives as needed in functions D1YDX1 and D3YDX3. If speed is important, the derivatives could be computed and stored in arrays.

A spline interpolation routine is especially useful for plotting curves on graphs. The routine has been used to produce many of the graphs in this book.

## A.5 ROOTS OF NONLINEAR EQUATIONS

Finding roots of nonlinear equations is essentially the reverse of an interpolation problem. When we interpolate a function, our object is to find a value of the dependent variable $y$ at a specific value of the independent variable $x$. When we are searching for the root of a function, we are trying to find the value of $x$ at a particular value, usually 0, of $y$. However, interchanging the variables completely changes the nature of the problem. Interpolation involves straightforward application of well-defined equations that are independent of the form of the original function: Finding roots of nonlinear equations may require different equations for different problems and almost always requires some sort of a search and iteration procedure.

The diffraction of light by a single slit provides an interesting example of a nonlinear equation. It is well known that the position of the interference maxima and minima from double slits and diffraction gratings can be determined analytically from consideration of the phase difference between the rays that pass through each slit, but only the minima of the diffraction pattern of a single slit can be found in this way. To find the position of a maximum, with the exception of the central one, we must differentiate the expression for the intensity with respect to the phase $\alpha$:

$$I = I_0\left(\frac{\sin\alpha}{\alpha}\right)^2 \quad \text{with} \quad \alpha = \frac{\pi a}{\lambda}\sin\theta \tag{A.39}$$

In Equation (A.39), $I_0$ is the intensity of the light at the central maximum ($\theta = 0$), $I$ is the intensity at angle $\theta$, $\lambda$ is the wavelength of the light, and $a$ is the slit width. The position of the maximum is given by solving

$$\frac{dI}{d\alpha} = 2I_0\left(\frac{\sin\alpha}{\alpha^3}\right)(\alpha\cos\alpha - \sin\alpha) = 0 \tag{A.40}$$

to obtain the value $\alpha_r$ at the root of the equation

$$f(\alpha) = \alpha_r - \tan\alpha_r = 0 \tag{A.41}$$

The first root is at $\alpha_r = 0$. The other roots cannot be calculated analytically and must be found by an iterative method. An approximate solution can be obtained by rewriting Equation (A.41) as

$$\alpha_r = \tan\alpha_r \tag{A.42}$$

and plotting separately the left and right sides to find the intersection of the straight line and the tangent curves. There are several mathematical ways to solve the problem, but making a plot of the function is always a good starting procedure.

### Trial-and-Error: The Half-Interval Method

With a personal computer, trial-and-error may be a suitable method for solving the occasional root finding problem. An orderly approach is advisable and the half-interval method is convenient. The procedure is to write a little program that requests a trial value of the root and calculates the function and displays its value. The initial trial value might be obtained from a graph, or perhaps by mapping the function for various values of the independent variable $x$, until a reasonable estimate of the root has been obtained. Then, a second trial $x$ is submitted, which produces a value of $y$ on the other side of the root. The half-interval method begins at this point. The procedure is to select a third trial value that is midway between the two that bracket the root. For the fourth trial value, we use the mean of the most recent value, and whichever of the two previous trials was on the other side of the root. The process continues until the root is found to the desired accuracy.

This rather primitive method of root finding could be improved with a little programming to let the program decide which root to choose, to calculate the mean, and perform the next trial. The program could proceed in a loop until the root had been found to a predefined degree of accuracy, or the calculation could be stopped manually. However, if we are willing to program that little bit of logic, slightly more effort will produce a much faster root-finding program.

### Secant Methods

The gain in speed comes from using the slope of the function in the calculation. We begin with two trial estimates of the root, $x_k$ and $x_{k+1}$, preferably, but not necessarily, on either side of the root. Then we write an expression for a linear interpolation between the two points. Equation (A.10) gives



$$f(x) = y_k + (x - x_k)\frac{(y_{k+1} - y_k)}{(x_{k+1} - x_k)} \tag{A.43}$$

where we have written $y_k = f(x_k)$ and so forth. Setting $f(x) = 0$ and solving for $x$ gives us an approximation to the value of $x$ at the root:

$$x = x_k - y_k\frac{x_{k+1} - x_k}{y_{k+1} - y_k} = \frac{x_k y_{k+1} - x_{k+1}y_k}{y_{k+1} - y_k} \tag{A.44}$$

For the next trial, we replace $x_{k+1}$ or $x_{k+2}$ by the value $x$ found in Equation (A.44) and repeat the calculation. The process can be repeated until the root is approximated as closely as desired. This is the first-order secant method.

There are various ways of choosing which of the previous values of $x$ ($x_k$ or $x_{k+1}$) to keep for the next iteration. The simplest is to keep the most recent value and discard the older value. Another way is to choose whichever is closer to the root [i.e., gives a smaller value of $f(x)$]. A third is to start the process with two values that straddle the root (i.e., give opposite signs for $y_1$ and $y_2$) and to continue to choose values that straddle the root after each iteration. This is the Regulo-Falsi method.

Clearly any method will find the root most quickly if the starting values are close to the root, but, in principle, the secant methods will almost always find a root of the function, eventually. With some functions, such as those that are antisymmetric about the root, there is the possibility that the search by the Regulo-Falsi method, for example, will jump back and forth across the root and never approach it. Additionally, for functions with several roots, we may not always find the one we want. Problems may also arise if two roots are very close together.

### Newton-Raphson Method

Instead of calculating the slope by finite differences, as in the secant method, we could use the tangent, or derivative of the function, if it can be calculated. Then, we can replace Equation (A.43) by

$$f(x) = y_k + (x - x_k)\frac{df(x_k)}{dx} \tag{A.45}$$

where $x_k$ and $y_k$ are the values of $x$ and $f(x)$ after the $k$th iteration. We find the next estimate $x_{k+1}$ for the root, as before, by setting $f(x)$ in Equation (A.45) to zero to obtain

$$x_{k+1} = x_k - y_k \div \frac{df(x_k)}{dx} \tag{A.46}$$

**Example A.7** Table A.3 shows steps in an iterative calculation of the second and third roots of Equation (A.41) by the secant and Newton-Raphson methods. Starting values were chosen by examining a plot of $\tan x$ versus $x$.

### Simultaneous Nonlinear Equations

In the examples of alternate fitting methods in Section 6.6, we obtained two pairs of coupled, nonlinear equations, Equations (6.24) and (6.27), which we wished to