# STATISTICS

To reduce the size of this section's PostScript file, we have divided it into three PostScript files. We present the following index:

PART 1



PART 2



PART 3

# 28. STATISTICS

Revised May 1996.

## 28.1. General [1–6]

A probability density function with known parameters enables us to predict the frequency with which a random variable will take on a particular value (if discrete) or lie in a given range (if continuous). In *parametric* statistics we have the opposite problem of estimating the parameters of the p.d.f. from a set of actual observations.

We refer to the true p.d.f. as the *population*; the data form a *sample* from this population. A *statistic* is any function of the data, plus known constants, which does not depend upon any of the unknown parameters. A statistic is a random variable if the data have random errors. An *estimator* is any statistic whose value is intended as a meaningful guess for the value of an unknown parameter; we denote estimators with hats, *e.g.*, $\widehat{\alpha}$.

Often it is possible to construct more than one reasonable estimator. Let $\alpha$ represent the true value of a parameter to be estimated; $\alpha$ is a vector $\boldsymbol{\alpha}$ if there is more than one parameter. Then if $\widehat{\alpha}$ is an estimator for $\alpha$, desirable properties for $\widehat{\alpha}$ are: (a) *Unbiased*; bias $b = E(\,\widehat{\alpha}\,) - \alpha$, where the expectation value is taken over a hypothetical set of similar experiments in which $\widehat{\alpha}$ is constructed the same way. The bias may be due to statistical properties of the estimator or to *systematic* errors in the experiment. If we can estimate the average bias $b$ we usually subtract it from $\widehat{\alpha}$ to obtain a new $\widehat{\alpha}' \equiv \widehat{\alpha} - b$. However, $b$ may depend upon $\alpha$ or other unknowns, in which case we usually try to choose an estimator which minimizes its average size. (b) *Minimum variance*; the minimum possible value of $\mathrm{Var}(\widehat{\alpha})$ is given by the Rao-Cramér-Frechet bound:

$$\mathrm{Var}_{\mathrm{min}} = \left[1 + \partial b/\partial\alpha\right]^2 / I(\alpha)\ ; \tag{28.1}$$

$$I(\alpha) = E\left\{\left[\frac{\partial}{\partial\alpha}\sum_{i=1}^{n} \ln f(x_i;\,\alpha)\right]^2\right\}\ .$$

(Compare with Eq. (28.6) below.) The sum is over all data and $b$ is the bias, if any; the $x_i$ are assumed independent and distributed as $f(x_i;\alpha)$, and the allowed range of $x$ must not depend upon $\alpha$. The ratio $\epsilon = \mathrm{Var}_{\mathrm{min}}/\mathrm{Var}(\,\widehat{\alpha}\,)$ is the *efficiency*. An *efficient* estimator (with $\epsilon = 1$) exists only for certain cases. The square root of the variance expresses the expected spread of $\widehat{\alpha}$ about its average value, as would be observed in a large number of repeats of the same measurement. (c) *Minimum mean-squared error* (mse); mse $= E[(\,\widehat{\alpha} - \alpha\,)^2] = V(\,\widehat{\alpha}\,) + b^2$. The mse combines the error due to any bias quadratically with the variance, which expresses only the spread about $E(\,\widehat{\alpha}\,)$, as distinct from $\alpha$, the true value. (d) *Robust*; a robust estimator is not sensitive to errors in our assumptions, *e.g.*, to departures from the assumed p.d.f. due to such factors as noise.

These criteria (and others) allow us to evaluate any procedure for obtaining $\widehat{\alpha}$. In many cases these criteria conflict. The bias, variance, and mse may depend on the unknown $\alpha$. In this case the optimum prescription for $\widehat{\alpha}$ may depend on the range in which we assume $\alpha$ to lie.

Following are techniques in common use for obtaining estimators and their standard errors $\sigma(\widehat{\alpha}) = \sqrt{\mathrm{Var}(\widehat{\alpha})}$. When the conditions of the central limit theorem are satisfied, the interval $\widehat{\alpha} \pm \sigma(\widehat{\alpha})$ forms a 68.3% *confidence interval.* This is a random interval in that its endpoints depend upon the randomly sampled data; its meaning here will be taken to be that in 68.3% of all similar experiments the interval will include the true value $\alpha$. One should be aware that in most practical cases the central limit theorem is only approximately satisfied and accordingly confidence intervals which depend on that are only approximate. Confidence intervals are discussed in Section 28.6 below.

## 28.2. Data with a common mean

Suppose we have a set of $N$ independent measurements $y_i$ assumed to be unbiased measurements of the same unknown quantity $\mu$ with a common, but unknown, variance $\sigma^2$ resulting from measurement error. Then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i = E(y) \tag{28.2}$$

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \widehat{\mu})^2 = \frac{N}{N-1}\left(E(y^2) - \widehat{\mu}^2\right) \tag{28.3}$$

are unbiased estimators of $\mu$ and $\sigma^2$. The variance of $\widehat{\mu}$ is $\sigma^2/N$. If the common p.d.f. of the $y_i$ is Gaussian, these statistics are independent. Then, for large $N$, the standard deviation of $\widehat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. If the $y_i$ are Gaussian or $N$ is large enough that the central limit theorem applies, then $\widehat{\mu}$ is an efficient estimator for $\mu$. Otherwise $\widehat{\mu}$ is sometimes subject to large fluctuations, *e.g.*, if the p.d.f. for $y_i$ has long tails. In this case the median of the $y_i$ may be a more robust estimator for $\mu$, provided the median and mean are expected to lie at the same point in the p.d.f. for $y$. For Gaussian $y$, the median has asymptotic (large-$N$) efficiency $2/\pi \approx 0.64$. Student's $t$-distribution provides an example in which there are large tails. In this case, for large $N$ the efficiency of the sample median relative to the sample mean is ($\infty$, $\infty$, 1.62, 1.12, 0.96, 0.80, 0.64) for (1, 2, 3, 4, 5, 8, $\infty$) degrees of freedom.

If $\sigma^2$ is known, $\widehat{\mu}$ as given in Eq. (28.2) is still the best estimator for $\mu$; if $\mu$ is known, substitute it for $\widehat{\mu}$ in Eq. (28.3) and replace $N-1$ by $N$, to obtain a somewhat better estimator $\widehat{\sigma}^2$.

If the $y_i$ have different, known, variances $\sigma_i^2$, then

$$\widehat{\mu} = \frac{1}{w}\sum^{N} w_i\, y_i \ , \tag{28.4}$$

is an unbiased estimator for $\mu$ with smaller variance than Eq. (28.2), where $w_i = 1/\sigma_i^2$ and $w = \sum w_i$. The standard deviation of $\widehat{\mu}$ is $1/\sqrt{w}$.

## 28.3. The method of maximum likelihood

### 28.3.1. *General*:

"From a theoretical point of view, the most important general method of estimation so far known is the *method of maximum likelihood*" [1]. We suppose that a set of independently measured quantities $x_i$ came from a p.d.f. $f(x;\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is an unknown set of parameters. The method of maximum likelihood consist of finding the set of values, $\widehat{\boldsymbol{\alpha}}$, which maximizes the joint probability density for all the data, given by

$$\mathscr{L}(\boldsymbol{\alpha}) = \prod_i f(x_i;\boldsymbol{\alpha}) \ , \tag{28.5}$$

where $\mathscr{L}$ is called the likelihood. It is usually easier to work with $\ln \mathscr{L}$, and since both are maximized for the same set of $\boldsymbol{\alpha}$, it is sufficient to solve the *likelihood equation*

$$\frac{\partial \ln \mathscr{L}}{\partial \alpha_n} = 0 \ . \tag{28.6}$$

The solution is called the *maximum likelihood estimate* of $\boldsymbol{\alpha}$. The importance of the approach is shown by the following proposition, proved in Ref. 1:

*If an efficient estimate* $\widehat{\boldsymbol{\alpha}}$ *of* $\boldsymbol{\alpha}$ *exists, the likelihood equation will have a unique solution equal to* $\widehat{\boldsymbol{\alpha}}$.

In evaluating $\mathscr{L}$, it is important that any normalization factors in the $f$'s which involve $\boldsymbol{\alpha}$ be included. However, we will only be interested in the maximum of $\mathscr{L}$ and in ratios of $\mathscr{L}$ at different $\boldsymbol{\alpha}$'s; hence any multiplicative factors which do not involve the parameters we want to estimate may be dropped; this includes factors which depend on the data but not on $\boldsymbol{\alpha}$.

If the solution to Eq. (28.6) is at a maximum, $\partial \ln \mathscr{L}/\partial\alpha_n$ will have negative slope in its vicinity. In many practical problems, one often uses nonlinear algorithms for finding the maximum, and must be alert to various possibilities for error: (a) Eq. (28.6) may yield a minimum, therefore one must check the second derivative; (b) there may be more than one maximum—one must try to find the global maximum; (c) the global maximum may lie at a boundary of the physical region, in which case Eq. (28.6) will not find it.

If an unbiased, efficient estimator exists, this method will find it. If $\partial \ln \mathscr{L}/\partial\alpha_n$ is linear in the vicinity of the root, an efficient estimator is guaranteed; other efficient cases are discussed in the literature. For large data samples, the central limit theorem will usually assure this condition in some significant neighborhood of zero; hence the estimator is usually efficient in that case, provided certain conditions are met (*e.g.*, that the solution does not lie on a boundary). In this case, in the neighborhood of the maximum $\ln \mathscr{L}$ is a downward-curving paraboloid and $\mathscr{L}$ is proportional to a multivariate Gaussian.

The results of two or more experiments may be combined by forming the product of the $\mathscr{L}$'s, or the sum of the $\ln \mathscr{L}$'s.

Under a one-to-one change of parameters from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\alpha})$, the maximum likelihood estimate is simply $\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})$, given the solution $\widehat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$. That is, the maximum likelihood solution for $\boldsymbol{\beta}$ is found by simple substitution of $\widehat{\boldsymbol{\alpha}}$ into the transformation equation. It is possible that the new solution $\widehat{\boldsymbol{\beta}}$ will be a biased solution for the true value of $\boldsymbol{\beta}$ even if $\widehat{\boldsymbol{\alpha}}$ is not biased, and vice-versa. In the asymptotic limit (of large amounts of data) both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ will (usually) converge to unbiased solutions, but at different rates.

Except in special cases like the least-squares method, the value of the likelihood function at the solution does not necessarily tell us whether the final fit was a sensible description of the data or not. In special cases such as the one discussed in Sec. 28.3.3, one can define a quantity which approaches the $\chi^2$-distribution in the limit of a large number of counts in the experiment, but in general some other strategy must be used. For example, data generated by Monte Carlo simulations of the experiment can be analyzed by the same method. If the experimental likelihood is lower than that of some agreed-upon fraction of these results, one should question the appropriateness of the p.d.f. At the same time one can check for bias in the solution.

### 28.3.2. *Error estimates*:

The covariance matrix $V$ may be estimated from

$$V_{nm} = \left( E\left[ -\frac{\partial^2 \ln \mathscr{L}}{\partial\alpha_n\,\partial\alpha_m}\bigg|_{\widehat{\alpha}} \right] \right)^{-1} . \tag{28.7}$$

If $\partial \ln \mathscr{L}/\partial\alpha_n$ is linear, the "expectation" operation in Eq. (28.7) has no effect because the second derivative of $\ln \mathscr{L}$ is constant. Otherwise, it may be approximated by taking the average of the quantity in square brackets over a range of $\alpha_n$ and $\alpha_m$ near the solution. For complex cases it may be more practical to evaluate $s$-standard-deviation errors from the contour

$$\ln \mathscr{L}(\boldsymbol{\alpha}) = \ln \mathscr{L}_{\max} - s^2/2 \ , \tag{28.8}$$

where $\ln \mathscr{L}_{\max}$ is the value of $\ln \mathscr{L}$ at the solution point (compare with Eq. (28.32), below). The extreme limits of this contour parallel to the $\alpha_n$ axis give an approximate $s$-standard-deviation confidence interval in $\alpha_n$. These intervals may not be symmetric and they may even consist of two or more disjoint intervals. This procedure gives one-standard-deviation errors in $\alpha_n$ equal to $\sqrt{V_{nn}}$ (not summed) of Eq. (28.7) if the estimator is efficient. If it is not efficient, the level of confidence implied by the value of $s$ is only approximate.

### 28.3.3. *Application to Poisson-distributed data*:

In the case of Poisson-distributed data in a counting experiment, the unbinned maximum likelihood method (where the index $i$ in Eq. (28.5) labels events) is preferred if the total number of events is very small. If there are enough events to justify binning them in a histogram, then one may alternatively maximize the likelihood function for the contents of the bins (so $i$ labels bins). This is equivalent to minimizing [7]

$$\chi^2 = \sum_i \left[ 2(N_i^{\rm th} - N_i^{\rm obs}) + 2N_i^{\rm obs} \ln(N_i^{\rm obs}/N_i^{\rm th}) \right] . \tag{28.9}$$

where $N_i^{\rm obs}$ and $N_i^{\rm th}$ are the observed and theoretical (from $f$) contents of the $i$th bin. In bins where $N_i^{\rm obs} = 0$, the second term is zero. This function asymptotically behaves like a classical $\chi^2$ for purposes of point estimation, interval estimation, *and goodness-of-fit.* It also guarantees that the area under the fitted function $f$ is equal to the sum of the histogram contents (as long as the overall normalization of $f$ is effectively left unconstrained during the fit), which is not the case for $\chi^2$ statistics based on a least-squares procedure with traditional weights.

## 28.4. Propagation of errors

Suppose that $F(x;\boldsymbol{\alpha})$ is some function of variable(s) $x$ and the fitted parameters $\boldsymbol{\alpha}$, with a value $\widehat{F}$ at $\widehat{\boldsymbol{\alpha}}$. The variance matrix of the parameters is $V_{mn}$. To first order in $\alpha_m - \widehat{\alpha}_m$, $F$ is given by

$$F = \widehat{F} + \sum_m \frac{\partial F}{\partial \alpha_m}(\alpha_m - \widehat{\alpha}_m) \ , \tag{28.10}$$

and the variance of $F$ about its estimator is given by

$$(\Delta F)^2 = E[(F - \widehat{F})^2] = \sum_{mn} \frac{\partial F}{\partial \alpha_m}\frac{\partial F}{\partial \alpha_n} V_{mn} \ , \tag{28.11}$$

evaluated at the $x$ of interest. For different functions $F_j$ and $F_k$, the covariance is

$$E[(F_j - \widehat{F}_j)(F_k - \widehat{F}_k)] = \sum_{mn} \frac{\partial F_j}{\partial \alpha_m}\frac{\partial F_k}{\partial \alpha_n} V_{mn} \ . \tag{28.12}$$

If the first-order approximation is in serious error, the above results may be very approximate. $\widehat{F}$ may be a biased estimator of $F$ even if the $\widehat{\boldsymbol{\alpha}}$ are unbiased estimators of $\boldsymbol{\alpha}$. Inclusion of higher-order terms or direct evaluation of $F$ in the vicinity of $\widehat{\boldsymbol{\alpha}}$ will help to reduce the bias.

## 28.5. Method of least squares

The *method of least squares* can be derived from the maximum likelihood theorem. We suppose a set of $N$ measurements at points $x_i$. The $i$th measurement $y_i$ is assumed to be chosen from a Gaussian distribution with mean $F(x_i;\boldsymbol{\alpha})$ and variance $\sigma_i^2$. Then

$$\chi^2 = -2\ln\mathscr{L} + \text{constant} = \sum_1^N \frac{[y_i - F(x_i;\boldsymbol{\alpha})]^2}{\sigma_i^2} \ . \tag{28.13}$$

Finding the set of parameters $\boldsymbol{\alpha}$ which maximizes $\mathscr{L}$ is the same as finding the set which minimizes $\chi^2$.

In many practical cases one further restricts the problem to the situation in which $F(x_i;\boldsymbol{\alpha})$ is a linear function of the $\alpha_m$'s,

$$F(x_i;\boldsymbol{\alpha}) = \sum_n \alpha_n\, f_n(x)\ , \tag{28.14}$$

where the $f_n$ are $k$ linearly independent functions (*e.g.*, 1, $x$, $x^2$, ..., or Legendre polynomials) which are single-valued over the allowed range of $x$. We require $k \leq N$, and at least $k$ of the $x_i$ must be distinct. We wish to estimate the linear coefficients $\alpha_n$. Later we will discuss the nonlinear case.

If the point errors $\epsilon_i = y_i - F(x_i;\boldsymbol{\alpha})$ are Gaussian, then the minimum $\chi^2$ will be distributed as a $\chi^2$ random variable with $n = N - k$ degrees of freedom. We can then evaluate the goodness-of-fit (confidence level) from Figs. 27.1 or 27.3, as per the earlier discussion. The confidence level expresses the probability that a *worse* fit would be obtained in a large number of similar experiments under the assumptions that: (a) the model $y = \sum \alpha_n\, f_n$ is correct and (b) the errors $\epsilon_i$ are Gaussian and unbiased with variance $\sigma_i^2$. If this probability is larger than an agreed-upon value (0.001, 0.01, or 0.05 are common choices), the data are *consistent* with the assumptions; otherwise we may want to find improved assumptions. As for the converse, most people do not regard a model as being truly *inconsistent* unless the probability is as low as that corresponding to four or five standard deviations for a Gaussian ($6\times 10^{-3}$ or $6\times 10^{-5}$; see Sec. 28.6.1). If the $\epsilon_i$ are not Gaussian, the method of least squares still gives an answer, but the goodness-of-fit test would have to be done using the correct distribution of the random variable which is still called "$\chi^2$."

Finding the minimum of $\chi^2$ in the linear case is straightforward:

$$\begin{aligned} -\frac{1}{2}\frac{\partial\chi^2}{\partial\alpha_m} &= \sum_i f_m(x_i)\left(\frac{y_i - \sum_n \alpha_n\, f_n(x_i)}{\sigma_i^2}\right) \\ &= \sum_i \frac{y_i\, f_m(x_i)}{\sigma_i^2} - \sum_n \alpha_n \sum_i \frac{f_n(x_i)\, f_m(x_i)}{\sigma_i^2}\ . \end{aligned} \tag{28.15}$$

With the definitions

$$g_m = \sum_i y_i\, f_m(x_i)/\sigma_i^2 \tag{28.16}$$

and

$$V^{-1}_{mn} = \sum_i f_n(x_i)\, f_m(x_i)/\sigma_i^2\ , \tag{28.17}$$

the $k$-element column vector of solutions $\widehat{\boldsymbol{\alpha}}$, for which $\partial\chi^2/\partial\alpha_m = 0$ for all $m$, is given by

$$\widehat{\boldsymbol{\alpha}} = V\,\boldsymbol{g}\ . \tag{28.18}$$

With this notation, $\chi^2$ for the special case of a linear fitting function (Eq. (28.14)) can be rewritten in the compact form

$$\chi^2 = \chi^2_{\min} + (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^T V^{-1} (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \ . \tag{28.19}$$

**Nonindependent $y_i$'s**

Eq. (28.13) is based on the assumption that the likelihood function is the product of independent Gaussian distributions. More generally, the measured $y_i$'s are not independent, and we must consider them as coming from a multivariate distribution with nondiagonal covariance matrix $S$, as described in Sec. 27.3.3. The generalization of Eq. (28.13) is

$$\chi^2 = \sum_{jk} [y_j - F(x_j;\boldsymbol{\alpha})] S^{-1}_{jk} [y_k - F(x_k;\boldsymbol{\alpha})] \ . \tag{28.20}$$

In the case of a fitting function that is linear in the parameters, one may differentiate $\chi^2$ to find the generalization of Eq. (28.15), and with the extended definitions

$$\begin{aligned} g_m &= \sum_{jk} y_j \ f_m(x_k) S^{-1}_{jk} \\ V^{-1}_{mn} &= \sum_{jk} f_n(x_j) \ f_m(x_k) S^{-1}_{jk} \end{aligned} \tag{28.21}$$

solve Eq. (28.18) for the estimators $\widehat{\boldsymbol{\alpha}}$.

The problem of constructing the covariance matrix $S$ is simplified by the fact that contributions to $S$ (not to its inverse) are additive. For example, suppose that we have three variables, all of which have independent statistical errors. The first two also have a common error resulting in a positive correlation, perhaps because a common baseline with its own statistical error (variance $s^2$) was subtracted from each. In addition, the second two have a common error (variance $a^2$), but this time the values are anticorrelated. This might happen, for example, if the sum of the two variables is a constant. Then

$$\begin{aligned} S = & \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} \\ & + \begin{pmatrix} s^2 & s^2 & 0 \\ s^2 & s^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & a^2 & -a^2 \\ 0 & -a^2 & a^2 \end{pmatrix} \ . \end{aligned} \tag{28.22}$$

If unequal amounts of the common baseline were subtracted from variables 1, 2, and 3—*e.g.*, fractions $f_1$, $f_2$, and $f_3$, then we would have

$$S = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$$

$$+\begin{pmatrix} f_1^2 s^2 & f_1 f_2 s^2 & f_1 f_3 s^2 \\ f_1 f_2 s^2 & f_2^2 s^2 & f_2 f_3 s^2 \\ f_1 f_3 s^2 & f_2 f_3 s^2 & f_3^2 s^2 \end{pmatrix} . \qquad (28.23)$$

While in general this "two-vector" representation is not possible, it underscores the procedure: Add zero-determinant correlation matrices to the matrix expressing the independent variation.

Care must be taken when fitting to correlated data, since off-diagonal contributions to $\chi^2$ are not necessarily positive. It is even possible for all of the residuals to have the same sign.

**Example: straight-line fit**

For the case of a straight-line fit, $y(x) = \alpha_1 + \alpha_2\, x$, one obtains, for independent measurements $y_i$, the following estimates of $\alpha_1$ and $\alpha_2$,

$$\widehat{\alpha}_1 = (g_1\, V_{22}^{-1} - g_2\, V_{12}^{-1})/D \ , \qquad (28.24)$$

$$\widehat{\alpha}_2 = (g_2\, V_{11}^{-1} - g_1\, V_{12}^{-1})/D \ , \qquad (28.25)$$

where

$$(V_{11}^{-1},\, V_{12}^{-1},\, V_{22}^{-1}) = \sum(1,\, x_i,\, x_i^2)/\sigma_i^2 \ , \qquad (28.26a)$$

$$(g_1,\, g_2) = \sum(1,\, x_i) y_i/\sigma_i^2 \ . \qquad (28.26b)$$

respectively, and

$$D = V_{11}^{-1}\, V_{22}^{-1} - (V_{12}^{-1})^{\,2} \ . \qquad (28.27)$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} V_{22}^{-1} & -V_{12}^{-1} \\ -V_{12}^{-1} & V_{11}^{-1} \end{pmatrix} \ . \qquad (28.28)$$

The estimated variance of an interpolated or extrapolated value of $y$ at point $x$ is:

$$(\,\widehat{y} - y_{\rm true})^2\Big|_{\rm est} = \frac{1}{V_{11}^{-1}} + \frac{V_{11}^{-1}}{D}\left(x - \frac{V_{12}^{-1}}{V_{11}^{-1}}\right)^2 \ . \qquad (28.29)$$

**28.5.1.** ***General comments***:

If $y$ is not linear in the fitting parameters $\boldsymbol{\alpha}$, the solution vector may have to be found by iteration. If we have a first guess $\boldsymbol{\alpha}_0$, then we may expand to obtain

$$\left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha} = \left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha_0} + V_{\alpha_0}^{-1}\cdot(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)+\ldots\ , \tag{28.30}$$

where $\partial\chi^2/\partial\alpha$ is a vector whose $m$th component is $\partial\chi^2/\partial\alpha_m$, and $(V_{mn}^{-1}) = \frac{1}{2}\partial^2\chi^2/\partial\alpha_m\partial\alpha_n$. (See Eqns. 28.7 and 28.17. When evaluated at $\widehat{\boldsymbol{\alpha}}$, $V^{-1}$ is the inverse of the covariance matrix.) The next iteration toward $\widehat{\boldsymbol{\alpha}}$ can be obtained by setting $\partial\chi^2/\partial\alpha_m|_{\alpha} = 0$ and neglecting higher-order terms:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 - V_{\alpha_0}\cdot\partial\chi^2/\partial\alpha|_{\alpha_0} \tag{28.31}$$

If $V$ is constant in the vicinity of the minimum, as it is when the model function is linear in the parameters, then $\chi^2$ is parabolic as a function of $\boldsymbol{\alpha}$ and Eq. (28.31) gives the solution immediately. Otherwise, further iteration is necessary. If the problem is highly nonlinear, considerable difficulty may be encountered. There may be secondary minima, and $\chi^2$ may be decreasing at physical boundaries. Numerical methods have been devised to find such solutions without divergence [8,11]. In particular, the CERN program MINUIT [11] offers several iteration schemes for solving such problems.

Note that minimizing any function proportional to $\chi^2$ (or maximizing any function proportional $\ln\mathscr{L}$) will result in the same parameter set $\widehat{\boldsymbol{\alpha}}$. Hence, for example, if the variances $\sigma_j^2$ are unknown but assumed equal and independent, one can still solve for $\widehat{\boldsymbol{\alpha}}$. One cannot, however, evaluate goodness-of-fit, and the covariance matrix is known only to within a constant multiplier. The scale can be estimated at least roughly from the size of $\chi^2$ compared to its expected size.

Additional information can be extracted from the behavior of the (normalized) residuals, $r_j = (y_j - F(x_j;\boldsymbol{\alpha})/\sigma_j$, which should themselves distribute normally with a mean of 0.

If the data covariance matrix $S$ has been correctly evaluated (or, equivalently, the $\sigma_j$'s, if the data are independent), then the $s$-standard deviation limits on the parameters are given by a set $\boldsymbol{\alpha}'$ such that

$$\chi^2(\boldsymbol{\alpha}) = \chi^2_{\min} + s^2\ . \tag{28.32}$$

This equation, a special case of 28.8, defines a contour in $\boldsymbol{\alpha}$-space; compare with the linear case in Eq. (28.19). It is often convenient for estimating errors in applications to nonlinear cases, where the matrix $V^{-1}|_{\alpha}$ may be a rapidly varying function of $\boldsymbol{\alpha}$. If the problem is highly nonlinear, such contours only approximately define the desired confidence regions which would have some given probability of covering the true value of $\boldsymbol{\alpha}$.

The method of least squares is sometimes used in cases where the distribution is not Gaussian or not known to be Gaussian. In such cases it can still be used, but it is then not a special case of the maximum-likelihood method, and the theorems having to do with that approach no longer apply. However, if (a) the distribution of $y_i - \sum \alpha_k f_k(x_i)$ has an expectation value of zero (unbiased) and (b) has a finite, known, fixed variance $\sigma_i^2$ (does not depend on $\boldsymbol{\alpha}$), then estimates of $\boldsymbol{\alpha}$ obtained by minimizing $\chi^2$ will be unbiased and have the smallest possible variance of all linear unbiased estimates (Gauss-Markov theorem). This statement is more general than the least-squares method as a special case of the maximum likelihood method in that the distributions do not have to be Gaussian, but more restrictive in that it applies only when the fitting function is linear in the $\alpha_k$'s.