STATISTICS

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

PART 1

PART 2

29. STATISTICS

Revised April 1998 by F. James (CERN).

29.1. Parameter estimation [1–4]

A probability density function $f(x;\alpha)$ (p.d.f.) with known parameters α enables us to predict the frequency with which random data x 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 α from a set of actual observations.

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 (the *estimate* $\widehat{\alpha}$) is intended as a meaningful guess for the value of the parameter α, or the vector $\boldsymbol{\alpha}$ if there is more than one parameter.

Since we are free to choose any function of the data as an estimator of the parameter α, we will try to choose that estimator which has the best properties. The most important properties are (a) *consistency*, (b) *bias*, (c) *efficiency*, and (d) *robustness*.

(a) An estimator is said to be *consistent* if the estimate $\widehat{\alpha}$ converges to the true value α as the amount of data increases. This property is so important that it is possessed by all commonly used estimators.

(b) The *bias*, $b = E(\widehat{\alpha}) - \alpha$, is the difference between the true value and the expectation of the estimates, where the expectation value is taken over a hypothetical set of similar experiments in which $\widehat{\alpha}$ is constructed the same way. When $b = 0$ the estimator is said to be unbiased. The bias may be due to statistical properties of the estimator or to *systematic* errors in the experiment. If we can estimate the b we can subtract it from $\widehat{\alpha}$ to obtain a new $\widehat{\alpha}' \equiv \widehat{\alpha} - b$. However, b may depend upon α or other unknowns, in which case we usually try to choose an estimator which minimizes its average size.

(c) *Efficiency* is the inverse of the ratio between the *variance of the estimates* $\mathrm{Var}(\widehat{\alpha})$ and the minimum possible value of the variance. Under rather general conditions, the minimum variance is given by the Rao-Cramér-Frechet bound:

$$\mathrm{Var}_{\mathrm{min}} = \left[1 + \partial b/\partial\alpha\right]^2 / I(\alpha) \; ; \tag{29.1}$$

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

(Compare with Eq. (29.6) 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 α. *Mean-squared error*, $\mathrm{mse} = E[(\widehat{\alpha} - \alpha)^2] = V(\widehat{\alpha}) + b^2$ is a convenient quantity which combines in the appropriate way the errors due to bias and efficiency.

(d) *Robustness*; is the property of being insensitive to departures from assumptions in the p.d.f. due to such factors as noise.

For some common estimators the above properties are known exactly. More generally, it is always possible to evaluate them by Monte Carlo simulation. Note that they will often depend on the unknown α.

29.2. 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 μ with a common, but unknown, variance σ^2 resulting from measurement error. Then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i = E(y) \tag{29.2}$$

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \widehat{\mu})^2 = \frac{N}{N-1}\left(E(y^2) - \widehat{\mu}^2\right) \tag{29.3}$$

are unbiased estimators of μ and σ^2. The variance of $\widehat{\mu}$ is σ^2/N. If the common p.d.f. of the y_i is Gaussian, these estimates are uncorrelated. Then, for large N, the standard deviation of $\widehat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. Again if the y_i are Gaussian, $\widehat{\mu}$ is an efficient estimator for μ. Otherwise the mean is in general not the most efficient estimator. For example, if the y follow a double-exponential distribution, the most efficient estimator of the mean is the sample median (the value for which half the y_i lie above and half below). This is discussed in more detail in Ref. 2, section 8.7.

If σ^2 is known, it does not improve the estimate $\widehat{\mu}$, as can be seen from Eq. (29.2); however, if μ is known, substitute it for $\widehat{\mu}$ in Eq. (29.3) and replace $N-1$ by N, to obtain a somewhat better estimator of σ^2.

If the y_i have different, known, variances σ_i^2, then the weighted average

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{N} w_i\, y_i \ , \tag{29.4}$$

is an unbiased estimator for μ with smaller variance than Eq. (29.2), where $w_i = 1/\sigma_i^2$ and $w = \sum w_i$. The standard deviation of $\widehat{\mu}$ is $1/\sqrt{w}$.

29.3. The method of maximum likelihood

29.3.1. *Parameter estimation by maximum likelihood*:

"From a theoretical point of view, the most important general method of estimation so far known is the *method of maximum likelihood*" [3]. 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 consists of finding the set of values, $\widehat{\boldsymbol{\alpha}}$, which maximizes the joint probability density for all the data, given by

$$\mathscr{L}(\boldsymbol{\alpha}) = \prod_i f(x_i;\boldsymbol{\alpha}) \ , \tag{29.5}$$

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*

$$\frac{\partial \ln \mathscr{L}}{\partial \alpha_n} = 0 \ . \tag{29.6}$$

When the solution to Eq. (29.6) is a maximum, it 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 $\boldsymbol{\alpha}$. The results of two or more independent experiments may be combined by forming the product of the $\mathscr{L}$'s, or the sum of the $\ln \mathscr{L}$'s.

Most commonly the solution to Eq. (29.6) will be found using a general numerical minimization program such as the CERN program MINUIT [8] which contains considerable code to take account of the many special cases and problems which can arise.

Under a one-to-one change of parameters from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\alpha})$, the maximum likelihood estimate $\widehat{\boldsymbol{\alpha}}$ transforms to $\boldsymbol{\beta}(\widehat{\boldsymbol{\alpha}})$. That is, the maximum likelihood solution is invariant under change of parameter. However, many properties of $\widehat{\boldsymbol{\alpha}}$, in particular the bias, are not invariant under change of parameter.

29.3.2. *Confidence intervals from the likelihood function*:

The covariance matrix V may be estimated from

$$V_{nm} = \left(E \left[-\frac{\partial^2 \ln \mathscr{L}}{\partial \alpha_n \, \partial \alpha_m} \bigg|_{\widehat{\alpha}} \right] \right)^{-1} \ . \tag{29.7}$$

In the asymptotic case (or a linear model with Gaussian errors), $\mathscr{L}$ is Gaussian, $\ln \mathscr{L}$ is a (multidimensional) parabola, and the second derivative in Eq. (29.7) is constant, so the "expectation" operation has no effect. This leads to the usual approximation of calculating the error matrix of the parameters by inverting the second derivative matrix of $\ln \mathscr{L}$. In this asymptotic case, it can be seen that a numerically equivalent way of determining s-standard-deviation errors is from the contour given by the $\boldsymbol{\alpha}'$ such that

$$\ln \mathscr{L}(\boldsymbol{\alpha}') = \ln \mathscr{L}_{\max} - s^2/2 \ , \tag{29.8}$$

where $\ln \mathscr{L}_{\max}$ is the value of $\ln \mathscr{L}$ at the solution point (compare with Eq. (29.32), below). The extreme limits of this contour parallel to the α_n axis give an approximate s-standard-deviation confidence interval in α_n. These intervals may not be symmetric and in pathological cases they may even consist of two or more disjoint intervals.

Although asymptotically Eq. (29.7) is equivalent to Eq. (29.8) with $s = 1$, the latter is a better approximation when the model deviates from linearity. This is because Eq. (29.8) is invariant with respect to even a non-linear transformation of parameters $\boldsymbol{\alpha}$, whereas Eq. (29.7) is not. Still, when the model is non-linear or errors are not Gaussian, confidence intervals obtained with both these formulas are only approximate. The true coverage of these confidence intervals can always be determined by a Monte Carlo simulation, or exact confidence intervals can be determined as in Sec. 29.6.3.

29.3.3. *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. (29.5) 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 [5]

$$\chi^2 = \sum_i \left[2(N_i^{\rm th} - N_i^{\rm obs}) + 2N_i^{\rm obs} \ln(N_i^{\rm obs}/N_i^{\rm th}) \right] . \tag{29.9}$$

where $N_i^{\rm obs}$ and $N_i^{\rm th}$ are the observed and theoretical (from f) contents of the ith bin. In bins where $N_i^{\rm obs} = 0$, the second term is zero. This function asymptotically behaves like a classical χ^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 χ^2 statistics based on a least-squares procedure with traditional weights.

29.4. 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

$$F = \widehat{F} + \sum_m \frac{\partial F}{\partial \alpha_m} (\alpha_m - \widehat{\alpha}_m) \ , \tag{29.10}$$

and the variance of F about its estimator is given by

$$(\Delta F)^2 = E[(F - \widehat{F})^2] = \sum_{mn} \frac{\partial F}{\partial \alpha_m} \frac{\partial F}{\partial \alpha_n} V_{mn} \ , \tag{29.11}$$

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

$$E[(F_j - \widehat{F}_j)(F_k - \widehat{F}_k)] = \sum_{mn} \frac{\partial F_j}{\partial \alpha_m} \frac{\partial F_k}{\partial \alpha_n} V_{mn} \ . \tag{29.12}$$

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.

29.5. 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 ith measurement y_i is assumed to be chosen from a Gaussian distribution with mean $F(x_i;\boldsymbol{\alpha})$ and variance σ_i^2. Then

$$\chi^2 = -2\ln\mathscr{L} + \text{constant} = \sum_1^N \frac{[y_i - F(x_i;\boldsymbol{\alpha})]^2}{\sigma_i^2} \ . \tag{29.13}$$

Finding the set of parameters $\boldsymbol{\alpha}$ which maximizes $\mathscr{L}$ is the same as finding the set which minimizes χ^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 α_m's,

$$F(x_i;\boldsymbol{\alpha}) = \sum_n \alpha_n \, f_n(x) \ , \tag{29.14}$$

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 α_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 χ^2 will be distributed as a χ^2 random variable with $n = N - k$ degrees of freedom. We can then evaluate the goodness-of-fit (confidence level) from Figs. 28.1 or 28.3, 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 ϵ_i are Gaussian and unbiased with variance σ_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 (6×10^{-3} or 6×10^{-5}; see Sec. 29.6.4). If the ϵ_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 "χ^2."

Minimizing χ^2 in the linear case is straightforward:

$$-\frac{1}{2}\frac{\partial\chi^2}{\partial\alpha_m} = \sum_i f_m(x_i)\left(\frac{y_i - \sum_n \alpha_n \ f_n(x_i)}{\sigma_i^2}\right)$$

$$= \sum_i \frac{y_i \ f_m(x_i)}{\sigma_i^2} - \sum_n \alpha_n \ \sum_i \frac{f_n(x_i) \ f_m(x_i)}{\sigma_i^2} \ . \tag{29.15}$$

With the definitions

$$g_m = \sum_i y_i \; f_m(x_i)/\sigma_i^2 \tag{29.16}$$

and

$$V_{mn}^{-1} = \sum_i f_n(x_i) \; f_m(x_i)/\sigma_i^2 \; , \tag{29.17}$$

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

$$\widehat{\boldsymbol{\alpha}} = V \; \boldsymbol{g} \; . \tag{29.18}$$

With this notation, χ^2 for the special case of a linear fitting function (Eq. (29.14)) can be rewritten in the compact form

$$\chi^2 = \chi^2_{\rm min} + (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^T V^{-1} (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) \; . \tag{29.19}$$

Nonindependent y_i's

Eq. (29.13) 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. 28.3.3. The generalization of Eq. (29.13) is

$$\chi^2 = \sum_{jk} [y_j - F(x_j;\boldsymbol{\alpha})] S^{-1}_{jk} [y_k - F(x_k;\boldsymbol{\alpha})] \; . \tag{29.20}$$

In the case of a fitting function that is linear in the parameters, one may differentiate χ^2 to find the generalization of Eq. (29.15), and with the extended definitions

$$\begin{aligned} g_m &= \sum_{jk} y_j \; f_m(x_k) S^{-1}_{jk} \\ V^{-1}_{mn} &= \sum_{jk} f_n(x_j) \; f_m(x_k) S^{-1}_{jk} \end{aligned} \tag{29.21}$$

solve Eq. (29.18) 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

$$S = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} + \begin{pmatrix} s^2 & s^2 & 0 \\ s^2 & s^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & a^2 & -a^2 \\ 0 & -a^2 & a^2 \end{pmatrix} . \tag{29.22}$$

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

$$S = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} + \begin{pmatrix} f_1^2 s^2 & f_1 f_2 s^2 & f_1 f_3 s^2 \\ f_1 f_2 s^2 & f_2^2 s^2 & f_2 f_3 s^2 \\ f_1 f_3 s^2 & f_2 f_3 s^2 & f_3^2 s^2 \end{pmatrix} . \tag{29.23}$$

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 χ^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 α_1 and α_2,

$$\widehat{\alpha}_1 = (g_1\, \Lambda_{22} - g_2\, \Lambda_{12})/D \ , \tag{29.24}$$

$$\widehat{\alpha}_2 = (g_2\, \Lambda_{11} - g_1\, \Lambda_{12})/D \ , \tag{29.25}$$

where

$$(\Lambda_{11},\, \Lambda_{12},\, \Lambda_{22}) = \sum (1,\, x_i,\, x_i^2)/\sigma_i^2 \ , \tag{29.26a}$$

$$(g_1,\, g_2) = \sum (1,\, x_i) y_i/\sigma_i^2 \ . \tag{29.26b}$$

respectively, and

$$D = \Lambda_{11}\, \Lambda_{22} - (\Lambda_{12})^{\,2} \ . \tag{29.27}$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} \Lambda_{22} & -\Lambda_{12} \\ -\Lambda_{12} & \Lambda_{11} \end{pmatrix} . \tag{29.28}$$

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

$$(\widehat{y} - y_{\rm true})^2 \Big|_{\rm est} = \frac{1}{\Lambda_{11}} + \frac{\Lambda_{11}}{D} \left(x - \frac{\Lambda_{12}}{\Lambda_{11}} \right)^2 \ . \tag{29.29}$$

29.5.1. ***Confidence intervals from the chisquare function***:

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

$$\left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha} = \left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha_0} + V_{\alpha_0}^{-1}\cdot(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0) + \dots \ , \tag{29.30}$$

where $\partial\chi^2/\partial\alpha$ is a vector whose mth 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. 29.7 and 29.17. 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:

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 - V_{\alpha_0}\cdot\partial\chi^2/\partial\alpha|_{\alpha_0} \ . \tag{29.31}$$

If V is constant in the vicinity of the minimum, as it is when the model function is linear in the parameters, then χ^2 is parabolic as a function of $\boldsymbol{\alpha}$ and Eq. (29.31) 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 χ^2 may be decreasing at physical boundaries. Numerical methods have been devised to find such solutions without divergence [7,8]. In particular, the CERN program MINUIT [8] offers several iteration schemes for solving such problems.

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

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 σ_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

$$\chi^2(\boldsymbol{\alpha}') = \chi^2_{\rm min} + s^2 \ . \tag{29.32}$$

This equation gives confidence intervals in the same sense as 29.8, and all the discussion of Sec. 29.3.2 applies as well here, substituting $-\chi^2/2$ for $\ln\mathscr{L}$.