20. STATISTICS

Revised June 1994 with the help of R. Cousins, F. James, G. Lynch, B. P. Roe, and M. Roos.

20.1. 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 α represent the true value of a parameter to be estimated; α is a vector $\boldsymbol{\alpha}$ if there is more than one parameter. Then if $\widehat{\alpha}$ is an estimator for α, 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 α 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:

$$\mathrm{Var}_{\mathrm{min}} = \left[1 + \partial b/\partial\alpha\right]^2 / I(\alpha)\ ; \tag{20.1}$$

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

(Compare with Eq. (20.6) 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. 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); $\mathrm{mse} = E[(\,\widehat{\alpha} - \alpha\,)^2] = V(\,\widehat{\alpha}\,) + b^2$. 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 α, 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 α. In this case the optimum prescription for $\widehat{\alpha}$ may depend on the range in which we assume α 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 α. 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 20.5 below.

20.2. 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

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i = E(y) \tag{20.2}$$

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \widehat{\mu})^2 = \frac{N}{N-1}\left(E(y^2) - \widehat{\mu}^2\right) \tag{20.3}$$

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 statistics are independent. Then, for large N, the variance of $\widehat{\sigma}^2$ is $2\sigma^4/N$. 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 μ. 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 μ, 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 (∞, ∞, 1.62, 1.12, 0.96, 0.80, 0.64) for (1, 2, 3, 4, 5, 8, ∞) degrees of freedom.

If σ^2 is known, $\widehat{\mu}$ as given in Eq. (20.2) is still the best estimator for μ; if μ is known, substitute it for $\widehat{\mu}$ in Eq. (20.3) and replace $N-1$ by N, to obtain a somewhat better estimator $\widehat{\sigma}^2$.

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

$$\widehat{\mu} = \frac{1}{w}\sum^{N} w_i\ y_i\ , \tag{20.4}$$

is an unbiased estimator for μ with smaller variance than Eq. (20.2), where $w_i = 1/\sigma_i^2$ and $w = \sum w_i$. The variance of $\widehat{\mu}$ is $1/w$.

20.3. The method of maximum likelihood

20.3.1. *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

$$\mathscr{L}(\boldsymbol{\alpha}) = \prod_i f(x_i;\boldsymbol{\alpha})\ , \tag{20.5}$$

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*

$$\frac{\partial \ln\mathscr{L}}{\partial\alpha_n} = 0\ . \tag{20.6}$$

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. (20.6) 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. (20.6) 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. (20.6) 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. 20.3.3, one can define a quantity which approaches the χ^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.

20.3.2. *Error estimates*:

The covariance matrix V may be estimated from

$$V_{nm} = \left(E\left[-\frac{\partial^2 \ln\mathscr{L}}{\partial\alpha_n\,\partial\alpha_m}\bigg|_{\widehat{\alpha}} \right] \right)^{-1} . \tag{20.7}$$

If $\partial\ln\mathscr{L}/\partial\alpha_n$ is linear, the "expectation" operation in Eq. (20.7) 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 α_n and α_m near the solution. For complex cases it may be more practical to evaluate s-standard-deviation errors from the contour

$$\ln\mathscr{L}(\boldsymbol{\alpha}) = \ln\mathscr{L}_{\rm max} - s^2/2 \ , \tag{20.8}$$

where $\ln\mathscr{L}_{\rm max}$ is the value of $\ln\mathscr{L}$ at the solution point (compare with Eq. (20.27), 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 they may even consist of two or more disjoint intervals. This procedure gives one-standard-deviation errors in α_n equal to $\sqrt{V_{nn}}$ (not summed) of Eq. (20.7) if the estimator is efficient. If it is not efficient, the level of confidence implied by the value of s is only approximate.

20.3.3. ***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. (20.5) 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]

$$\chi^2 = \sum_i \left[2(N_i^{\rm th} - N_i^{\rm obs}) + 2N_i^{\rm obs} \ln(N_i^{\rm obs}/N_i^{\rm th}) \right] . \tag{20.9}$$

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.

20.4. 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

$$\chi^2 = -2\ln\mathscr{L} + \text{constant} = \sum_1^N \frac{[y_i - F(x_i;\boldsymbol{\alpha})]^2}{\sigma_i^2} . \tag{20.10}$$

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,

$$F(x_i;\boldsymbol{\alpha}) = \sum_n \alpha_n\, f_n(x) , \tag{20.11}$$

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 $k \le N$, 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. 19.1 or 19.3, 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. 20.5.1). 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."

Finding the minimum of χ^2 in the linear case is straightforward:

$$\begin{aligned} -\frac{1}{2}\frac{\partial\chi^2}{\partial\alpha_m} &= \sum_i f_m(x_i)\left(\frac{y_i - \sum_n \alpha_n f_n(x_i)}{\sigma_i^2}\right) \\ &= \sum_i \frac{y_i\, f_m(x_i)}{\sigma_i^2} - \sum_n \alpha_n \sum_i \frac{f_n(x_i)\, f_m(x_i)}{\sigma_i^2} \ . \end{aligned} \tag{20.12}$$

With the definitions

$$g_m = \sum_i y_i\, f_m(x_i)/\sigma_i^2 \tag{20.13}$$

and

$$\left(V_{\widehat{\alpha}}^{-1}\right)_{mn} = \sum_i f_n(x_i)\, f_m(x_i)/\sigma_i^2 \ , \tag{20.14}$$

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

$$\widehat{\boldsymbol{\alpha}} = V_{\widehat{\alpha}}\, \boldsymbol{g} \ . \tag{20.15}$$

Non-independent y_i's

More generally, the measured y_i's are not independent. Then the set of σ_i^2's must be replaced by the $N\times N$ covariance matrix V_y. Then, if H is the $N\times k$ matrix with element $H_{in} = f_n(x_i)$, the solution $\widehat{\boldsymbol{\alpha}}$ is given by the solution to the *normal equation*

$$(H^T\, V_y^{-1}\, H)\, \widehat{\boldsymbol{\alpha}} = H^T\, V_y^{-1}\, \boldsymbol{y} \ , \tag{20.16a}$$

or, formally,

$$\widehat{\boldsymbol{\alpha}} = (H^T\, V_y^{-1}\, H)^{-1}\, H^T\, V_y^{-1}\, \boldsymbol{y} \equiv D\boldsymbol{y} \ , \tag{20.16b}$$

where $\boldsymbol{y}$ is the N-element vector of measured y_i's. The normal equations may be solved by numerical methods much more computationally efficient than brute application of Eq. (20.16b). In particular, $H^T\ V_y^{-1}\ H$ is sometimes singular or nearly singular. In such cases there is at least one f_n which may be expressed as a linear combination of others (or nearly so) when evaluated at the data points. The best procedure is usually to drop such functions from the expansion (or set $\widehat{\alpha}_n = 0$). See Press [8], Maindonald [9], or Basilevsky [10] for discussions.

In terms of the $k \times N$ matrix D, the standard covariance matrix for the $\widehat{\boldsymbol{\alpha}}$ is estimated by

$$V_{\widehat{\alpha}} = D\ V_y\ D^T\ . \tag{20.17}$$

If the measured y_i's are independent, V_y is diagonal with ii^{th} element σ_i^2 and $V_{\widehat{\alpha}}$ is obtained from Eq. (20.14) above.

The expected covariance [see Eq. (19.9)] of $\widehat{\alpha}_n$ and $\widehat{\alpha}_m$ is estimated by

$$E\big[(\alpha_n - \widehat{\alpha}_n)(\alpha_m - \widehat{\alpha}_m)\big] = (V_{\widehat{\alpha}})_{nm}\ . \tag{20.18}$$

Even when the y_i's are independent (diagonal V_y), $\widehat{\alpha}_n$ and $\widehat{\alpha}_m$ may not be (nondiagonal $V_{\widehat{\alpha}}$).

For the model function $y = \sum \alpha_n\ f_n(x)$, the estimated variance of an interpolated or extrapolated value of y at a point x is

$$\begin{aligned} E\left[(y - \widehat{y}\,)^2\right] &= \sigma^2(y) \\ &= \sum_{n,m} (V_{\widehat{\alpha}})_{nm}\ f_n(x)\ f_m(x)\ . \end{aligned} \tag{20.19}$$

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,

$$\widehat{\alpha}_1 = (g_1\ S_{22} - g_2\ S_{12})/D\ , \tag{20.20}$$

$$\widehat{\alpha}_2 = (g_2\ S_{11} - g_1\ S_{12})/D\ , \tag{20.21}$$

where

$$(S_{11},\, S_{12},\, S_{22}) = \sum (1,\, x_i,\, x_i^2)/\sigma_i^2\ , \tag{20.22a}$$

$$(g_1,\, g_2) = \sum (1,\, x_i) y_i/\sigma_i^2\ . \tag{20.22b}$$

respectively, and

$$D = S_{11}\, S_{22} - S_{12}^2\ . \tag{20.23}$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} S_{22} & -S_{12} \\ -S_{12} & S_{11} \end{pmatrix}\ . \tag{20.24}$$

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

$$(\,\widehat{y} - y_{\text{true}})^2\Big|_{\text{est}} = \frac{1}{S_{11}} + \frac{S_{11}}{D}\left(x - \frac{S_{12}}{S_{11}}\right)^2\ . \tag{20.25}$$

20.4.1. ***General comments***:

If y is not linear in the fitting parameters α_n, or if the errors σ_i depend upon y and therefore on α_n, the solution vector may have to be found by iteration of Eqs. (20.13)–(20.15) or Eq. (20.16b). The same results may be obtained by numerical techniques from the sum of squares, χ^2, directly, if we have a reasonable first guess $\boldsymbol{\alpha}_0$ for the solution vector:

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0 - \left(\frac{\partial^2\chi^2}{\partial\alpha^2}\right)^{-1}_{\alpha_0} \cdot \left.\frac{\partial\chi^2}{\partial\alpha}\right|_{\alpha_0} \tag{20.26a}$$

and

$$V_{\widehat{\alpha}} = 2\left(\frac{\partial^2\chi^2}{\partial\alpha^2}\right)^{-1}_{\widehat{\alpha}} \ , \tag{20.26b}$$

where $\partial\chi^2/\partial\alpha$ is a k-element vector whose n^{th} element is $\partial\chi^2/\partial\alpha_n$, $\partial^2\chi^2/\partial\alpha^2$ is a $k \times k$ matrix with mn^{th} element $\partial^2\chi^2/(\partial\alpha_m \cdot \partial\alpha_n)$, and all derivatives are to be evaluated at the points indicated. If "χ^2" is a true χ^2, the second-derivative matrix is independent of $\boldsymbol{\alpha}$; therefore the shape of the χ^2 as a function of $\boldsymbol{\alpha}$ is a paraboloid and Eq. (20.26a) will give the solution immediately. Otherwise one may need to iterate Eq. (20.26a) to arrive at a solution (Newton-Raphson method). The CERN program MINUIT [11] offers several iteration schemes for solving such problems.

Note that in Eq. (20.16b), one needs only a matrix proportional to V_y to find $\widehat{\boldsymbol{\alpha}}$. Hence, for example, if the variances σ_i^2 of the errors are unknown but assumed equal and independent, and $E(\epsilon_i) = 0$, one can still solve for $\widehat{\boldsymbol{\alpha}}$. One cannot, however, solve for $V_{\widehat{a}}$ or evaluate goodness-of-fit. These can be estimated from the *residuals*, $r_i = \widehat{y}(x_i) - y_i$, where $\widehat{y}(x_i)$ is the fitted curve at x_i, because study of the r_i enables one to estimate V_y. In addition, the residuals can be used to look for evidence of bias such as trends in the data not incorporated in the model [4].

The errors on the solution $\widehat{\boldsymbol{\alpha}}$ are independent of the value of χ^2 at minimum—they depend only upon the shape about the minimum. Eq. (20.26b) implies that s-standard-deviation limits on the elements of $\widehat{\boldsymbol{\alpha}}$ are given by the set of $\boldsymbol{\alpha}'$ such that

$$\chi^2(\boldsymbol{\alpha}') = \chi^2_{\rm min} + s^2 \ ; \tag{20.27}$$

(This is a special case of Eq. (20.8).) This equation, which defines a contour in $\boldsymbol{\alpha}$-space, is often convenient for estimating errors in applications of least-squares techniques to nonlinear cases, where the second derivative [Eq. (20.26b)] may be a rapidly varying function of $\boldsymbol{\alpha}$. If the problem is highly nonlinear, all such contours are, at best, only approximations to desired exact confidence regions which would have some given probability of covering the true value of $\boldsymbol{\alpha}$. It may be that Eq. (20.27) will define a set of disjoint regions. In addition, iteration of Eq. (20.26a) may require sophisticated techniques [8,11] to reach convergence in a practical amount of computation. For example, in cases involving many variables in $\boldsymbol{\alpha}$, especially if the correlations are large, simplex or other techniques which do not involve explicit calculation of derivatives are often to be preferred. Such techniques are designed to find their way through complicated nonlinear problems without diverging to infinite $\boldsymbol{\alpha}$ (unless the minimum is actually at infinity).

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 σ_i^2 (does not depend on $\boldsymbol{\alpha}$), then estimates of $\boldsymbol{\alpha}$ obtained by minimizing χ^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 α_k's.

For example, in counting experiments one often groups the data in bins in order to associate a Poisson error with each bin. In this case y_i is the bin height and the error depends on the expectation value of the theory in each bin, as estimated by the best fit of the model. Since the variances are functions of the fitting parameters, neither the conditions of the Gauss-Markov theorem nor the assumption of Gaussian distributions (with or without fixed variances) in the context of the maximum likelihood approach are valid without a large-number approximation, and so an unbiased or efficient least-squares fit is not guaranteed. In such cases it seems more sensible to use one of the approaches discussed in Sec. 20.3.3.

20.5. Errors and confidence intervals

We measure a mass, lifetime, or other physical quantity under the assumption that a "true answer" α exists. The conditions of the measurement introduce a random element, and our measurement (or combination of measurements) $\widehat{\alpha}_{\rm exp}$ samples a distribution with p.d.f. $f(\widehat{\alpha};\alpha)$. The unknown constant α appears as a parameter. We suppose that for every value of α we can find two values $\gamma_1(\alpha,\varepsilon)$ and $\gamma_2(\alpha,\varepsilon)$ such that repeated experiments would produce results in the interval $\gamma_1 < \widehat{\alpha} < \gamma_2$ a fraction $1-\varepsilon$ of the time, where

$$1 - \varepsilon = \int_{\gamma_1}^{\gamma_2} f(\widehat{\alpha};\alpha)\, d\widehat{\alpha} \ . \tag{20.28}$$

This situation is shown in Fig. 20.1 (ignore the "unphysical region" part of the graph for now), where the region between the curves

$\gamma_1(\alpha,\varepsilon)$ and $\gamma_2(\alpha,\varepsilon)$ is indicated by the domain $D(\varepsilon)$. It can be argued that since the point $(\alpha_{\mathrm{actual}}, \widehat{\alpha}_{\mathrm{exp}})$ belongs to D, then our statement that repeated experiments would produce values of $\widehat{\alpha}$ in the interval $\gamma_1 < \widehat{\alpha} < \gamma_2$ is equivalent to the statement that the *confidence interval* $c_1 < \alpha < c_2$ includes α_{actual} with probability $1-\varepsilon$ [1,6]. (We will call ε the *confidence coefficient*.) In this "confidence interval" or frequentist approach, α is a parameter, not a statistical variable. Instead, c_1 and c_2 vary from experiment to experiment and *are* statistical variables. It is very different to say that a lifetime τ is to be found in the interval $\tau_0 \pm \sigma_\tau$ with 68% probability than to say that the interval $\tau_0 \pm \sigma_\tau$ (which can vary from experiment to experiment) includes the actual, fixed, value of the lifetime with 68% probability.

The actual choice of γ_1 and γ_2, such that $\int_{\gamma_1}^{\gamma_2} f(\widehat{\alpha};\alpha)\,d\widehat{\alpha} = 1-\varepsilon$, can be made in an infinite number of ways, but in practical situations there are usually additional criteria. For a Gaussian distribution, for example, choosing the limits symmetric about the mean minimizes the length of the interval. The area of the excluded tail on either side is then $\varepsilon/2$. For a Poisson distribution negative values cannot occur, so $\gamma(\widehat{\alpha},\alpha)$ (with $\widehat{\alpha}$ an integer and α the Poisson mean) might be taken as the curve below which ε of the area under the distribution lies. (In this case the curve really consists of discrete points, since $\widehat{\alpha}$ can have only discrete values.) For $\varepsilon = 0.05$ the curve starts at $(\alpha,\widehat{\alpha}) = (3.0, 0)$. If in a given experiment no decays to a certain final state are seen, we might then conclude that $\alpha < 3.0$ excludes the actual value of α with 95% probability. This statement can be converted to a similar statement about the branching fraction.

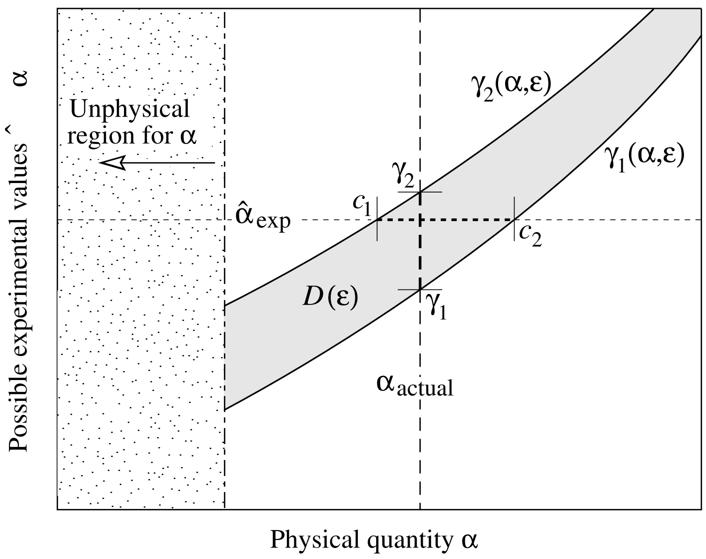

Figure 20.1: Confidence intervals for a single unknown parameter α. One might think of the p.d.f. $f(\widehat{\alpha};\alpha)$ as being plotted out of the paper as a function of $\widehat{\alpha}$ along each vertical line of constant α. The domain $D(\varepsilon)$ contains a fraction $1-\varepsilon$ of the area under each of these functions.

In Sec. 19 we discussed such confidence limits for a χ^2 distribution (where ε was called CL). Here we discuss confidence intervals for the Gaussian and Student's t-distribution, and confidence limits for the Poisson case. We then discuss the much more contentious situation in which the horizontal line at ordinate $\widehat{\alpha}$ in Fig. 20.1 enters $D(\varepsilon)$ at a boundary for unphysical values of α, so that at least c_1 is undefined—for example, if we find $\widehat{m}^2 = -30 \pm 50$ eV^2.

Extensive tables and graphs were once used to find confidence intervals and limits, but by now their main function is to confirm that software is working. FORTRAN mathematical libraries (IMSL, NAG, CERNLIB) are readily available, and a wide range of distributions are available in personal computer spreadsheet applications such as Microsoft® Excel [12]. Its built-in functions CHIDIST, NORMDIST, and TDIST (Student's t-distribution), along with "Solver," were used to produce or check the numbers given in this section.

20.5.1. ***Gaussian errors***:

If the data are such that the distribution of the estimator(s) satisfies the central limit theorem discussed in Sec. 19.3.3, the Gaussian distribution is the basis of the error analysis. If there is more than one parameter being estimated, the multivariate Gaussian is used. For the univariate case with known σ,

$$1 - \varepsilon = \int_{\widehat{\mu}-\delta}^{\widehat{\mu}+\delta} f(x;\, \widehat{\mu},\, \sigma^2)\, dx = \text{erf}\left(\frac{x - \widehat{\mu}}{\sqrt{2}\,\sigma}\right) \tag{20.29}$$

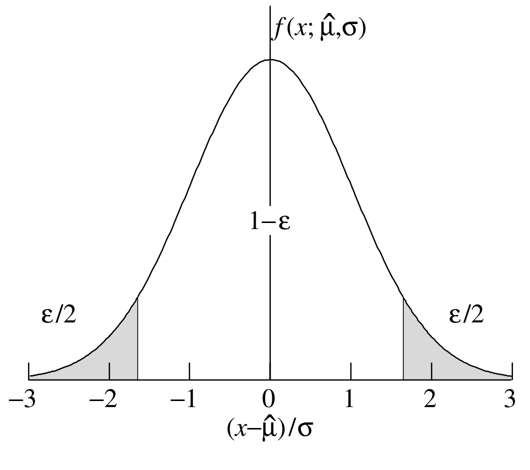

Figure 20.2: Illustration of a symetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by ε, are as shown.

is the probability that the true value of μ will fall within $\pm\delta$ ($\delta > 0$) of the measured $\widehat{\mu}$. This interval will cover μ in a fraction $1 - \varepsilon$ of all similar measurements. Fig. 20.2 shows a $\delta = 1.64\sigma$

Table 20.1: Area of the tails ε outside $\pm\delta$ from the mean of a Gaussian distribution.

ε (%)	δ	ε (%)	δ
31.73	1σ	20	1.28σ
4.55	2σ	10	1.64σ
0.27	3σ	5	1.96σ
6.3×10^{-3}	4σ	1	2.58σ
5.7×10^{-5}	5σ	0.1	3.29σ
2.0×10^{-7}	6σ	0.01	3.89σ

confidence interval unshaded. The choice $\delta = \sqrt{\mathrm{Var}(\,\widehat{\mu}\,)} \equiv \sigma$ gives an interval called the *standard error* which has $1-\varepsilon = 68.27\%$ if σ is known. Confidence coefficients ε for other frequently used choices of δ are given in Table 20.1. For other δ, find ε as the ordinate of Fig. 19.1 on the $n = 1$ curve at $\chi^2 = (\delta/\sigma)^2$. We can set a one-sided (upper or lower) limit by excluding above $\widehat{\mu} + \delta$ (or below $\widehat{\mu} - \delta$); ε's for such limits are 1/2 the values in the table above.

We have increased confidence that the interval covers the true value as $1-\varepsilon$ increases, or χ^2 increases. We must be careful to distinguish this case from the other major use of Fig. 19.1, evaluation of goodness-of-fit (Sec. 20.4.0). In that case we have increased confidence in the fit as χ^2 decreases. In an attempt to reduce possible confusion in this discussion, we will use the ε notation (which corresponds to notation used in hypothesis testing [4]) when discussing confidence intervals and CL notation when discussing goodness-of-fit. Elsewhere in this *Review*, where the confusion between fit confidence level and interval (usually an upper or lower limit) confidence level does not arise, we follow the common practice of using "CL" to refer to the confidence level of the interval. This CL is understood to represent $1-\varepsilon$.

If the variance σ^2 of the estimator is not known, but must be estimated from the data, then we need to incorporate the error in $\widehat{\sigma}$ into our confidence interval using Student's t distribution. If we have N data points with which we estimate k parameters, the Gaussian approximation is adequate for $N - k \gg 1$. Otherwise replace δ by a factor $T\widehat{\sigma}$, T being defined by

$$1-\varepsilon = \int_{-T}^{T} f(t; N-k)\; dt \;, \tag{20.30}$$

where f for the Student's t-distribution is defined in Table 19.1. T is tabulated in Ref. 13 and in Table 20.2.

For multivariate $\boldsymbol{\alpha}$ we must consider pairwise correlations. Assuming a multivariate Gaussian, Eq. (19.20), and subsequent discussion the standard error ellipse for the pair $(\,\widehat{\alpha}_m, \widehat{\alpha}_n)$ may be drawn as in Fig. 20.3.

The minimum χ^2 or maximum likelihood solution is at $(\,\widehat{\alpha}_m, \widehat{\alpha}_n)$. The standard errors σ_m and σ_n are defined as shown,

Table 20.2: t limits containing $1-\varepsilon$ of the area of Student's t-distribution $f(t;N-k)$.

	ε (%)					
$N-k$	31.73	10.00	5.00	4.55	1.00	0.27
1	1.84	6.31	12.71	13.97	63.66	235.8
2	1.32	2.92	4.30	4.53	9.92	19.21
3	1.20	2.35	3.18	3.31	5.84	9.22
4	1.14	2.13	2.78	2.87	4.60	6.62
5	1.11	2.01	2.57	2.65	4.03	5.51
10	1.05	1.81	2.23	2.28	3.17	3.96
20	1.03	1.72	2.09	2.13	2.85	3.42
∞	1.00	1.64	1.96	2.00	2.58	3.00

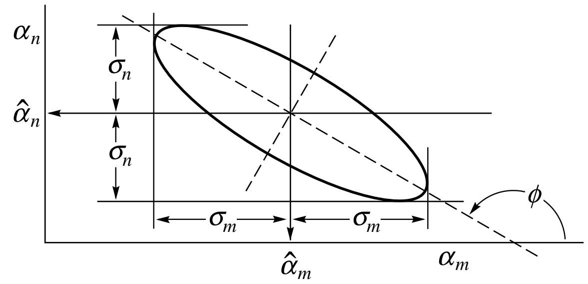

Figure 20.3: Standard error ellipse for the estimators $\widehat{\alpha}_m$ and $\widehat{\alpha}_n$. In this case the correlation is negative.

where the ellipse is at a constant value of $\chi^2 = \chi^2_{\text{min}} + 1$ or $\ln \mathscr{L} = \ln \mathscr{L}_{\text{max}} - 1/2$. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{mn}\,\sigma_m\,\sigma_n}{\sigma_m^2 - \sigma_n^2}\ . \tag{20.31}$$

For non-Gaussian or nonlinear cases, one may construct an analogous contour from the same χ^2 or $\ln\mathscr{L}$ relations. Any other parameters $\widehat{\alpha}_\ell, \ell \neq m, n$ must be allowed freely to find their optimum values for every trial point.

For any unbiased procedure (*e.g.*, least squares or maximum likelihood) being used to estimate k parameters α_i, $i = 1,\ldots,k$, the probability $1-\varepsilon$ that the true values of all k lie within the s-standard deviation ellipsoid may be found from Fig. 19.1. Read the ordinate as ε; the correct value of ε occurs on the $n = k$ curve at $\chi^2 = s^2$. For example, for $k = 2$, the probability that the true values of α_1 and α_2 simultaneously lie within the one-standard-deviation error ellipse ($s = 1$), centered on $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$, is 39%. This probability only assumes Gaussian errors, unbiased estimators, and that the model describing the data in terms of the α_i is correct.

20.5.2. ***Poisson processes—upper limits*:**

Because the outcome of a Poisson process is an integral number of events, n_0, it is usually not possible to set confidence intervals for the true Poisson parameter μ at a certain exact ε. For large n_0 an approximate interval can be set using the Gaussian approximation, in our section on Probability, Sec. 19.3.2, and the techniques of Sec. 20.5.1.

For small n_0 we can define an upper limit N for μ as being that value of μ such that it would be at least $1 - \varepsilon$ (*e.g.*, 90% or 95%) probable that a random observation of n would then lie above the observed n_0. Thus

$$1 - \varepsilon = \sum_{n=n_0+1}^{\infty} f(n;N) \ ; \qquad \varepsilon = \sum_{n=0}^{n_0} f(n;N) \ . \tag{20.32}$$

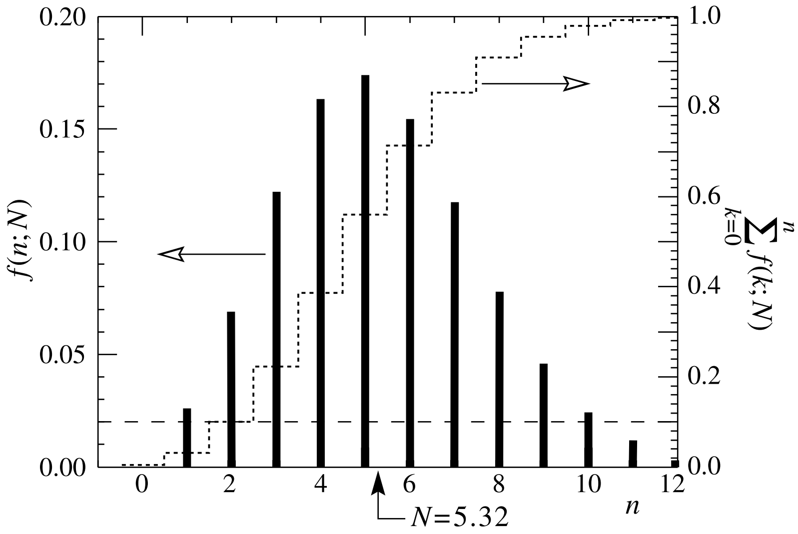

Figure 20.4: Illustration of Eq. (20.32) Poisson probabilities for an assumed mean of N. With an observed count $n_0 = 2$, $N = 5.32$ as shown gives summed probability $\varepsilon = 10\%$. The dotted summed probability curve (scale on right) has been displaced by -0.5 for clarity.

Fig. 20.4 illustrates the case with $n_0 = 2$ and $1 - \varepsilon = 90\%$, for which it may be shown that $N = 5.32$. For any given n_0 and desired ε we can obtain N from the χ^2 Confidence Level figure because of a relation between the Poisson and the χ^2 distributions: read the ordinate as ε, find χ^2 on the curve for $n = 2(n_0 + 1)$; then $N = \chi^2/2$. Some useful values are given in Table 20.3.

The meaning of these upper limits is that, for a given true μ, the probability is *at least* $1 - \varepsilon$ that one will observe n_0 which will result in N which is $\geq \mu$. The probability for that to occur may be higher than $1-\varepsilon$; for example, if $\mu \leq 2.30$ a "90%" upper limit will actually exceed μ 100% of the time. Note from Eq. (20.32) that for $n_0 = 0$, $N = -\ln\varepsilon$.

Table 20.3: Poisson upper limits N for n_0 observed events.

n_0	$\varepsilon =$ 10%	$\varepsilon =$ 5%	n_0	$\varepsilon =$ 10%	$\varepsilon =$ 5%
0	2.30	3.00	6	10.53	11.84
1	3.89	4.74	7	11.77	13.15
2	5.32	6.30	8	13.00	14.44
3	6.68	7.75	9	14.21	15.71
4	7.99	9.15	10	15.41	16.96
5	9.27	10.51	11	16.60	18.21

20.5.3. *Bounded physical region**:

The measurement of a physical constant α results in an estimator $\widehat{\alpha}$, together with some knowledge of experimental error and therefore knowledge of $f(\widehat{\alpha};\alpha)$, the parameterized p.d.f. that allows us to state the probability with which repeated experiments would produce results in a given range. It does *not* permit us to comment about α itself, which in this language is a constant, not a statistical variable. At the beginning of this section we introduced the confidence interval, or frequentist, approach to the problem, and were able to say that with a given probability the unknown parameter could be found between (statistical) limits c_1 and c_2. But what if a physical boundary exists? Although polarization should be less than one and mass or its square should be greater than zero, experimental results do not always fall inside such a physical boundary because of statistical fluctuations.

However one might set a limit, there is little question about how to report and combine data [14]. A given experiment finds an unbiased estimator $\widehat{\alpha} = -5 \pm 10$ for a physical constant (*e.g.* the square of the mass of a neutrino, in eV^2). This value should be reported as the primary result. In case the true value is zero, for example, this "unphysical" result would not be unlikely. It can be combined with the results of other such experiments by forming the appropriately weighted average of unbiased results, including negative ones, to find an unbiased estimator which expresses our best knowledge of the parameter.

What if we wish to extend our concept of confidence limit to such a situation? The question of how to calculate an upper limit in the vicinity of a physical boundary is one of the most divisive in high-energy physics. We present two main approaches: The confidence interval, or frequentist, method, and the Bayesian method. "Classical method" is applied to one or the other by various writers, so we avoid the term.

1. The method of confidence intervals [1,15]. This is the approach described in the introduction, and requires little further explanation.

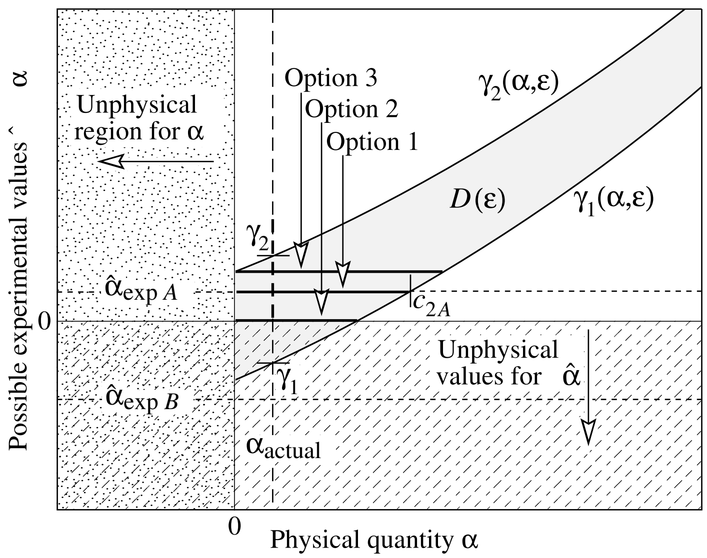

Figure 20.5: The situation near a physical boundary. In Fig. 20.1 the horizontal line for a given $\widehat{\alpha}_{\rm exp}$ crossed the domain $D(\varepsilon)$, bounded by $\gamma_1(\alpha,\epsilon)$ and $\gamma_2(\alpha,\epsilon)$ entirely in the physical region, entering at c_1 and leaving at c_2. The limits γ_1 and γ_2 cannot be defined in a region where α is not defined, so the functions cannot be continued into the unphysical region. As a result c_1 (for experiment A) or c_1 and c_2 (for experiment B) cannot be defined. Options 1, 2, and 3 label the ways one might define confidence intervals, as described in the text.

It is presently the method in favor [1,6,14]. For a Gaussian distribution it gives the same result as the Bayesian approach with a flat prior distribution (see below) if the region containing α with the stated probability is far from an unphysical region, as in Fig. 20.1. Two cases in which this is untrue are shown in Fig. 20.5, where as a matter of convenience we assume that α must be positive. As before, we can define limits γ_1 and γ_2 for each value of the unknown parameter α, such that we can expect that a fraction $1-\varepsilon$ of repeated experiments to produce results between these limits. Since this can be done for each value of α, the limits are described by the functions $\gamma_1(\alpha,\varepsilon)$ and $\gamma_2(\alpha,\varepsilon)$. However, these *cannot* be extended into a region in which α makes no sense. Experimental result $\widehat{\alpha}_{{\rm exp}\,A}$, indicated in Fig. 20.5, is positive, but if the true value is $\alpha_{\rm actual}$ a significant fraction of repetitions of the experiment would produce negative $\widehat{\alpha}$. In these cases there is no horizontal intercept c_1, so without further assumptions we cannot make a statement about the region which would cover $\alpha_{\rm actual}$ in a given fraction of experiments. Experimental result $\widehat{\alpha}_{{\rm exp}\,B}$ presents a more serious problem, since it is so negative that there is no physical α for which the point $(\widehat{\alpha},\alpha)$ lies in the domain $D(\varepsilon)$. The reason why the frequentist method gives no confidence interval is clear: This measured value of $\widehat{\alpha}$ would be unlikely *no matter what the true value of α was.*

There are several *ad hoc* ways to set confidence limits in such cases, although many frequentists would prefer to stop with the weighted average of unbiased results—if the outcome is exceedingly unlikely, one should look to the experiment, not to the statistics. The methods we list below all involve placing c_1 on the physical boundary, which in our example is at $\alpha = 0$.

1. If $\widehat{\alpha}_{\rm exp} > \gamma_1(0,\varepsilon)$, as in Experiment A, c_2 is defined. Use it for the upper limit, whether or not $\widehat{\alpha}_{\rm exp} > 0$.
2. If $\widehat{\alpha}_{\rm exp} < 0$, as in Experiment B, use the c_2 corresponding to $\widehat{\alpha}_{\rm exp} = 0$.
3. If c_1 is not defined, "lift up" $\widehat{\alpha}$ to $\gamma_2(0,\varepsilon)$, where $c_1 = 0$. Use the corresponding c_2 as the upper limit.

These three options are shown in in Fig. 20.5; note that there are regions where more than one of them can be used, with different results. The third option is certainly the most conservative. For Gaussian $f(\widehat{\alpha};\alpha)$ the upper limit c_2 is a one-sided Gaussian confidence limit; read the tables for a 90% two-sided limit to obtain 95% one-sided limit. Alternatively, read the intercepts of the dotted lines in Fig. 20.7. (The horizontal axis is incorrectly labeled for this application.)

2. The Bayesian approach [3]. This is the approach favored in the older literature, and has (unfortunately and incorrectly) been referred to as the "PDG method" in certain papers. To begin with, it is argued that while α is not a statistical variable, our *knowledge* of α is less than complete, and it is fair to describe our uncertainty by treating α as a statistical variable. The parameterized p.d.f. $f(\widehat{\alpha};\alpha)$ is replaced by the conditional p.d.f. $f(\widehat{\alpha}|\alpha)$. The confidence limit question can then be rephrased: Our measurements provide $f(\widehat{\alpha}|\alpha)$, that is, information about $\widehat{\alpha}$ for a fixed and unknown value of α, while we really want to know $g(\alpha|\widehat{\alpha})$, which tells us that, given our measurement $\widehat{\alpha}$, the "true answer" α lies between α and $\alpha + d\alpha$ with probability $g(\alpha|\widehat{\alpha})\, d\alpha$. The connection is provided by Bayes' theorem (Eq. (19.7):

$$g(\alpha|\widehat{\alpha}) = \frac{f(\widehat{\alpha}|\alpha)\,\pi(\alpha)}{\int f(\widehat{\alpha}|\alpha)\,\pi(\alpha)\,d\alpha}\ . \tag{20.33}$$

Here $\pi(\alpha)$ represents our "advance knowledge" of the value of α. In the usual case we claim no prior knowledge, so that before the experiment all physically reasonable values of α are equally probable: $\pi(\alpha)$ is a constant over the region of interest and zero in the unphysical region. This assumption leads to the conclusion that

$$g(\alpha|\widehat{\alpha}) = \begin{cases} f(\widehat{\alpha}|\alpha)/\int f(\widehat{\alpha}|\alpha)\,d\alpha & \text{if } \alpha \text{ is in the physical region;} \\ 0 & \text{otherwise;} \end{cases} \tag{20.34}$$

where this time the integral is over the physical region. In Fig. 20.6 we assume that an ensemble of experiments would produce values for $\widehat{\alpha}$ which distribute as shown, with a significant probability of obtaining results with unphysical values. With our assumed step function $\pi(\alpha)$, the effect of Eq. (20.33) or (20.34) is to replace this distribution with the function shown by the shaded region, except that it is renormalized to unit area. By stating our confidence at the 90% level that α lies below the beginning of the dark shaded region, we mean that 90% of the area in the physical region is in the light shaded region.

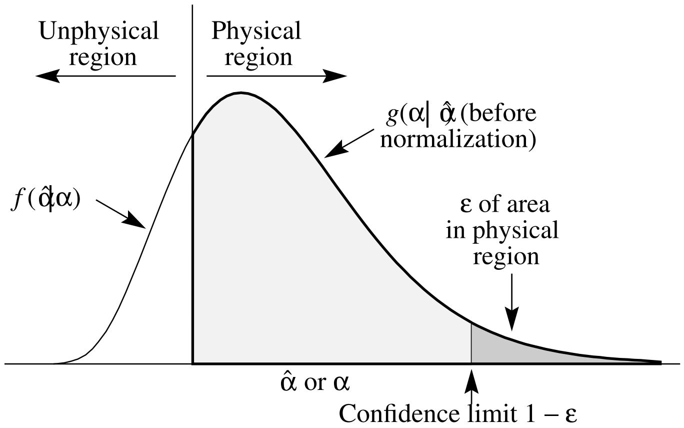

Figure 20.6: An example of a bounded physical region, in which a measurement $\widehat{\alpha}$ can fall in an unphysical region with significant probability. If we assume that α, the quantity we are trying to measure, cannot lie in the unphysical region (0 probability) but can lie anywhere in the physical region ("no prior knowledge"), then Bayes' theorem says that our new knowledge of the distribution of α, given our measurement $\widehat{\alpha}$, is given by the shaded function after appropriate renormalization.

In most cases of interest in this Review, $\widehat{\alpha}$ is assumed to be a random value from a Gaussian distribution. Application of the procedure sketched in Fig. 20.6 then leads to the family of curves shown in Fig. 20.7. The confidence limit set by this method is always greater than the [one-sided] confidence interval set without the restriction of an unphysical region, and approaches it from above as the tail in the unphysical region becomes unimportant. It is also greater than any of the limits shown in Fig. 20.5. With a small modification (exclusion of that portion of the negative tail inside the physical region in the confidence interval definition), it smoothly approaches the usual two-sided confidence interval for Gaussian distributions.

Even so, it is not a valid confidence limit. If it were, the interval would include the true value of α with *exactly* $1-\varepsilon$ probability no matter what the true value was. If the true answer is zero, our procedure, by guaranteeing a limit greater than zero for any experiment, also guarantees that the confidence interval for any ε includes α_{actual} with 100% probability. Only as α increases does the probability decrease toward the α-independent Gaussian result.

The error function corresponding to the right axis of Fig. 20.7 shows the probability that $\widehat{\alpha}/\sigma$ at or below the given value should occur. If the experimental value is exceeding improbable, then the

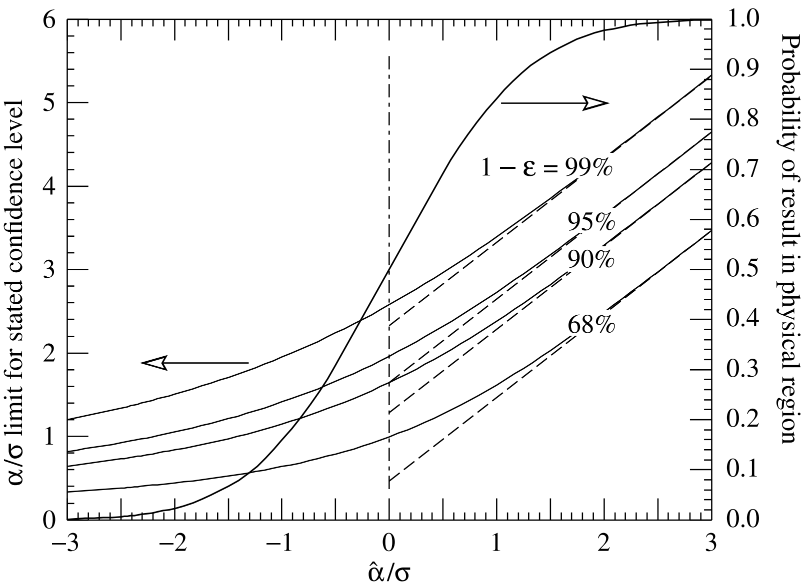

Figure 20.7: Application of the Bayesian scheme shown in Fig. 20.6 to the case of Gaussian $f(\widehat{\alpha}|\alpha)$. For example, if our measurement $\widehat{\alpha}$ is 1.0 standard deviations negative, then we conclude that $\alpha < 1.15\sigma$ with 90% probability—however, there is only a 31% probability that an experimental result as low as this would occur. Note that these are upper limits, so that the asymptote for large $\widehat{\alpha}/\sigma$ corresponds to a one-sided confidence interval, *e.g.*, the asymptote for a 95% confidence level is $\alpha < \widehat{\alpha} + 1.64\sigma$, corresponding to a 90% confidence interval for a two-sided distribution. The dashed lines show the frequentist limit; if Option 3 is used, these are extended horizontally to the right for negative $\widehat{\alpha}/\sigma$.

formal confidence limit obtained by this or any other method means very little.

What about the arbitrariness of $\pi(\alpha)$? If the square of the neutrino mass is measured ($\alpha = m_\nu^2$), then should we not take the prior knowledge distribution as proportional to $\sqrt{\alpha}$?† There are other attractive options. Jeffreys points out that if $\pi(\alpha)\,d\alpha = d\alpha/\alpha$, then the distributions for α and α^n are proportional [16], but there are practical difficulties with this approach. Lynch has investigated prior distributions that are constant in α, α^2, and $\sqrt{\alpha}$ in the context of Gaussian $f(\widehat{\alpha}|\alpha)$, and has observed that assuming a prior distribution that is flat in α gives results that are much more satisfactory than one gets from the others [17]: All three methods have the property that the probability that the calculated limit contains the correct answer is 100% when $\alpha = 0$ and approaches the proper value when $\alpha >> \sigma$, but the approach to the proper value as α increases is much faster when the prior distribution is taken to be flat in α. In this case the approach is also monotonic, giving it the "conservative" property that for no value of α will the method produce a limit that has a probability of being correct that is less than the stated confidence limit. Although there is nothing unique about the limit calculated with a constant $\pi(\alpha)$, it has desirable features and no obvious replacement.

Summary: If there is a significant probability of obtaining an estimator corresponding to an unphysical value for a parameter, there is no universally accepted way way to make a statement of the sort "α is less that c_2 with probability $1-\varepsilon$." A variety of upper limits can be defined, but no method is entirely satisfactory. The Bayesian method with a flat prior distribution gives a reasonable upper limit which combines everything we know about the unknown quantity α into a physically reasonable value, but it does not give a complete summary of the information contained in the experiment.

20.5.4. *Poisson processes with background* [18] :

If we observe n_0 events in a Poisson process which has two components, signal and background, estimating a limit on the signal is more complicated. Let μ_S be the unknown mean (the Poisson parameter) for the signal and μ_B be the mean for the sum of all backgrounds. Assume μ_B is known with negligible error; however we don't know n_B, the actual number of events resulting from the background. We do know that $n_B \leq n_0$. If $\mu_B + \mu_S$ is large, the Gaussian approximation to the Poisson distribution (see Sec. 19.3.2) is usually adequate, and one can define confidence intervals or limits as above, assuming $\widehat{n}_B \approx \mu_B$ and therefore $\widehat{\mu}_S = n_0 - \mu_B$ with variance equal to n_0 (larger than $\widehat{\mu}_S$ to allow for the error in $\widehat{n}_B$).

Otherwise an upper limit can be defined by extension of the argument of the preceding section. Let N be the desired upper limit on μ_S with confidence coefficient ε. Set N to be that value of μ_S such that any random repeat of the current experiment with $\mu_S = N$ and the same μ_B would observe *more* than n_0 events in total *and* would have $n_B \leq n_0$, all with probability $1-\varepsilon$. For any assumed N and μ_B we can calculate this probability:

$$1 - \varepsilon = 1 - \frac{e^{-(\mu_B+N)} \sum_{n=0}^{n_0} \frac{(\mu_B+N)^n}{n!}}{e^{-\mu_B} \sum_{n=0}^{n_0} \frac{\mu_B^n}{n!}} \ . \tag{20.35}$$

We adjust N to obtain a desired ε. For $\mu_B = 0$ this converges to Eq. (20.32). As in that case (see the last paragraph of Section 20.5.2) this gives a *conservative* upper limit in that for any given true μ_S we get a true probability $\geq 1-\varepsilon$ that $N \geq \mu_S$, averaged over a large set of identically performed experiments. For $\varepsilon = 0.10$, Fig. 20.8 shows N as a function of n_0 and μ_B.

Averaging of experiments and other comparisons require that n_0 and μ_B be quoted and the technique used for upper limit extraction be given.

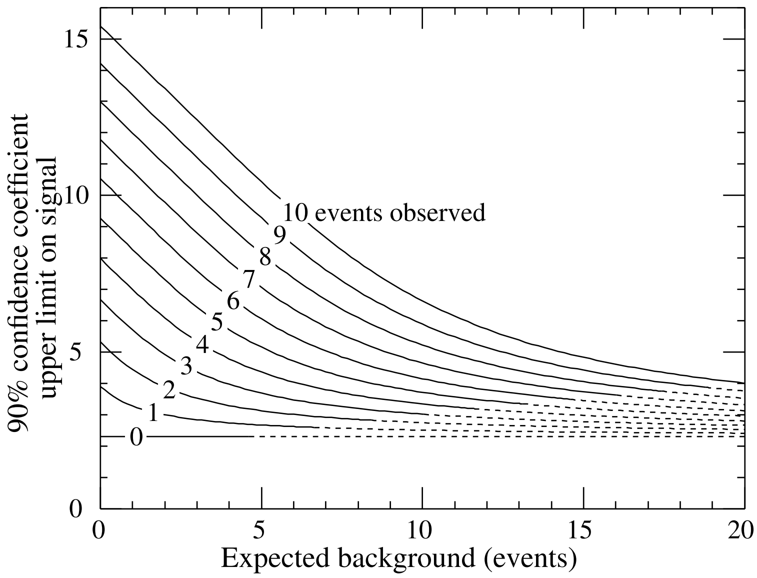

Figure 20.8: 90% confidence coefficient upper limit on the number of signal events as a function of the expected number of background events. For example, if the expected background is 8 events and 5 events are observed, then the signal is 4.0 (approximately) or less with 90% confidence. Dashed portions indicate regions where it is to be expected that the number observed would exceed the number actually observed $\geq$ 99% of the time, even in the complete absence of signal.

If $\mu_B \gg n_0$ the experimenter should question the probability of observing n_B as that n_0. If this is very small the background, μ_B, may not have been calculated properly and the upper limit for μ_S obtained under those assumptions may be too low. For example, in Fig. 20.8, the dashed portions of the curves lie in the region where n_0 is expected to exceed the observed value 99% of the time (or more), even in the complete absence of signal. In these regions one should be cautious about accepting the results of the measurement.

As in the Gaussian case (Sec. 20.5.3), whenever $n_0 < \mu_B$ some experimenters may prefer to use N calculated as if $n_0 \approx \mu_B$ rather than the smaller value obtained from the observed n_0.

20.6. Propagation of errors

Suppose we have a set of N random variables y_i which may be direct measurements or derived estimators $\widehat{\alpha}$, and we have a covariance matrix $V(y)$ for these. We can make a transformation to a different set of variables $f_n \equiv f_n(y)$, $j = 1, \ldots, M$ $(M \leq N)$ and obtain best estimates for the f_n from

$$\widehat{f}_n \approx f_n(\,\widehat{y}\,) + \frac{1}{2}\sum_{k,n}^{N} V_{kn}(\,\widehat{y}\,)\left[\frac{\partial^2 f_n}{\partial y_k\, \partial y_n}\right]_{\widehat{y}} \tag{20.36}$$

with covariance matrix

$$V_{ij}(\widehat{f}\,) \approx \sum_{n,m} \frac{\partial f_i}{\partial y_n}\bigg|_{\widehat{y}} \frac{\partial f_j}{\partial y_m}\bigg|_{\widehat{y}} V_{nm}(\widehat{y}\,) \ . \tag{20.37}$$

For a single-valued function f of a single measurement y with variance σ^2 (*i.e.*, $M = 1, N = 1$), this becomes

$$\begin{aligned} \widehat{f} &\approx \ f(\,\widehat{y}\,) + \frac{1}{2}\sigma^2 f''(\,\widehat{y}\,) \\ V(\,\widehat{f}\,) &\approx \ \sigma^2[f'(\,\widehat{y}\,)]^2 \ , \end{aligned} \tag{20.38}$$

where the primes denote differentiation with respect to y, evaluated at $\widehat{y}$.

These approximations are based on a Taylor expansion of f about the true value of y. If f is approximately linear in y over a range of roughly $\widehat{y}_i \pm \sigma(y_i)$, the approximation is good and the second-order terms in (20.36) and (20.38) can be neglected. This is what is usually done. However, if linearity is badly violated (*e.g.*, $f \propto 1/y$ and $\widehat{y}$ is no more than a few σ from zero), it should be recognized that propagation of errors will give very approximate results. In such cases $\widehat{f} \approx f(\,\widehat{y}\,)$ may be a biased estimator for f even if $\widehat{y}$ is unbiased for y, and the second-order terms in (20.36) and (20.38) will help to reduce that bias.

*In addition to the references cited, communications with R.D. Cousins, F. James, G. Lynch, and B. Roe have been invaluable in formulating this section.

†There is an additional problem: Even if we set a confidence limit on m_ν^2 by some particular recipe, it translates into a different confidence limit on $\sqrt{m_\nu^2}$ except when a Bayesian procedure with Jeffreys' $\pi(\alpha) \propto 1/\alpha$ prior distribution is used.

References:

1. H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press, New Jersey (1958).
2. J. Orear, "Notes on Statistics for Physicists," Univ. of Calif. Radiation Lab. Report UCRL-8417 (1958), unpublished; "Notes on Statistics for Physicists, Revised," Cornell Univ. Report CLNS 82/511 (1982), unpublished.
3. F.T. Solmitz, Annu. Rev. Nucl. Sci. **14**, 375–402 (1964).
4. W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam and London, 1971).
5. L. Lyons, *Statistics for Nuclear and Particle Physicists,* (Cambridge University Press, New York (1986)).
6. B.P. Roe, *Probability and Statistics in Experimental Physics,* (Springer-Verlag, New York, 208 pp., 1992).

7. S. Baker and R. Cousins, Nucl. Instr. and Meth. **221**, 437 (1984).
8. W.H. Press *et al.*, *Numerical Recipes* (Cambridge University Press, New York, 1986).
9. W.H. Maindonald *et al.*, *Statistical Computation* (John Wiley and Sons, Inc., New York, 1984).
10. A. Basilevsky *et al.*, *Applied Matrix Algebra in the Statistical Sciences* (North Holland, New York, 1983).
11. F. James and M. Roos, "MINUIT, Function Minimization and Error Analysis," CERN D506 (Long Writeup). Available from the CERN Program Library Office, CERN-DD Division, CERN, CH-1211, Geneva 21, Switzerland.
12. Microsoft® is a registered trademark of Microsoft Corporation.
13. M. Abramowitz and I. Stegun, eds., *Handbook of Mathematical Functions* (Dover, New York, 1972).
14. F. James and M. Roos, Phys. Rev. **D44**, 299 (1991).
15. J. Neyman, *Lectures and Conferences on Mathematical Statistics,"* Washington (1938).
16. H. Jeffreys, *Theory of Probability,* 3rd ed. (Oxford Univ. Press, Oxford, England, 447 pp., 1961).
17. G. Lynch, "Calculations of Upper Limits," Particle Data Group Note PDG-93-03 (Sept. 1993) (unpublished).
18. O. Helene, Nucl. Instr. and Meth. **212**, 319 (1983).