28. STATISTICS

Revised April 1998 by F. James (CERN). Minor updates October 1999 by R. Cousins (UCLA). (More revisions anticipated following workshop at CERN in Jan. 2000.)

28.1. Parameter estimation [1–5]

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 depends on the chosen metric, i.e., if $\widehat{\alpha}$ is an unbiased estimator of α, then $(\widehat{\alpha})^2$ is generally not an unbiased estimator of α^2. 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:

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

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

(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 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.

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

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

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

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 [$\sim \exp(-\sqrt{2}|y-\mu|/\sigma)$], 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, Sec. 8.7.

If σ^2 is known, it does not improve the estimate $\widehat{\mu}$, as can be seen from Eq. (28.2); however, if μ is known, substitute it for $\widehat{\mu}$ in Eq. (28.3) 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

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

is an unbiased estimator for μ with smaller variance than an unweighted average; here $w_i = 1/\sigma_i^2$ 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. *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

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

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*

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

When the solution to Eq. (28.6) 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. (28.6) will be found using a general numerical minimization program such as the CERN program MINUIT [9], 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.

28.3.2. ***The likelihood is not a p.d.f. for the parameters***:

Recall the definition of a probability *density* function: a function $p(\alpha)$ is a p.d.f. for α if $p(\alpha)d\alpha$ is the *probability* for α to be within α and $\alpha + d\alpha$. The likelihood function $\mathscr{L}(\alpha)$ is *not* a p.d.f. for α, so in general it is nonsensical to integrate the likelihood function with respect to its parameter(s).

Consider, for example, the Poisson probability for obtaining n when sampling from a distribution with mean α: $f(n;\alpha) = \alpha^n \exp(-\alpha)/n!$. If one obtains $n = 3$ in a particular experiment, then $\mathscr{L}(\alpha) = \alpha^3 \exp(-\alpha)/6$. Nothing in the construction of $\mathscr{L}$ makes it a probability *density*, *i.e.*, a function which one can multiply by $d\alpha$ in order to obtain a probability.

In Bayesian theory (see Sec. 28.6), one constructs the posterior p.d.f. for α by multiplying the prior p.d.f. for α by $\mathscr{L}$. If the prior p.d.f. is uniform, integrating the posterior p.d.f. may give the appearance of integrating $\mathscr{L}$. But note that the prior p.d.f. crucially provides the *density* which makes it sensible to multiply by $d\alpha$ to obtain a probability. In non-Bayesian applications, such as those considered in this section, only likelihood *ratios* are used (or equivalently, differences in $\ln \mathscr{L}$).

28.3.3. ***Confidence intervals from the likelihood function***:

The covariance matrix V may be estimated from

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

(Here and below, the superscript -1 indicates matrix inversion, followed by application of the subscripts.)

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. (28.7) 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

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

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 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. (28.7) is equivalent to Eq. (28.8) with $s = 1$, the latter is a better approximation when the model deviates from linearity. This is because Eq. (28.8) is invariant with respect to even a non-linear transformation of parameters $\boldsymbol{\alpha}$, whereas Eq. (28.7) 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. 28.6.3.

28.3.4. ***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. (Sometimes it is "extended" to include the total number of events as a Poisson-distributed observable.) 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 [6]

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

where $N_i^{\rm obs}$ and $N_i^{\rm th}$ are the observed and theoretical (from f) contents of the 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.

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 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

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

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

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

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 $k \le N$, 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 (significance level) from Figs. 27.1 or 27.3, as per the earlier discussion. The significance 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. 28.6.4). 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:

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

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, χ^2 for the special case of a linear fitting function (Eq. (28.14)) can be rewritten in the compact form

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

Nonindependent $\boldsymbol{y_i}$'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 χ^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

$$S = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} + \begin{pmatrix} f_1^2 s^2 & f_1 f_2 s^2 & f_1 f_3 s^2 \\ f_1 f_2 s^2 & f_2^2 s^2 & f_2 f_3 s^2 \\ f_1 f_3 s^2 & f_2 f_3 s^2 & f_3^2 s^2 \end{pmatrix} . \tag{28.23}$$

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 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,

$$\widehat{\alpha}_1 = (g_1\,\Lambda_{22} - g_2\,\Lambda_{12})/D , \tag{28.24}$$

$$\widehat{\alpha}_2 = (g_2\,\Lambda_{11} - g_1\,\Lambda_{12})/D , \tag{28.25}$$

where

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

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

respectively, and

$$D = \Lambda_{11}\,\Lambda_{22} - (\Lambda_{12})^{\,2} . \tag{28.27}$$

The covariance matrix of the fitted parameters is:

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} \Lambda_{22} & -\Lambda_{12} \\ -\Lambda_{12} & \Lambda_{11} \end{pmatrix} . \tag{28.28}$$

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

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

28.5.1. ***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

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

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. 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 χ^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 χ^2 may be decreasing at physical boundaries. Numerical methods have been devised to find such solutions without divergence [8,9]. In particular, the CERN program MINUIT [9] 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 (known as "pulls"), $r_j = (y_j - F(x_j;\boldsymbol{\alpha})/\sigma_j$, which should themselves distribute normally with mean 0 and rms deviation 1.

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 each of the parameters are given by a set $\boldsymbol{\alpha}'$ such that

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

This equation gives confidence intervals in the same sense as 28.8, and all the discussion of Sec. 28.3.3 applies as well here, substituting $-\chi^2/2$ for $\ln\mathscr{L}$.

28.6. Exact confidence intervals

28.6.1. *Two methodologies*:

There are two different approaches to statistical inference, which we may call Frequentist and Bayesian. For the cases considered up to now, both approaches give the same numerical answers, even though they are based on fundamentally different assumptions. However, for exact results for small samples and for measurements near a physical boundary, the different approaches may yield very different confidence limits, so we are forced to make a choice. There is an enormous amount of literature devoted to the question of Bayesian vs non-Bayesian methods, most of it written by people who are fervent advocates of one or the other methodology, which often leads to exaggerated conclusions. For a reasonably balanced discussion, we recommend the following articles: by a statistician [10], and by a physicist [7].

28.6.2. ***Bayesian***: The Bayesian concept of probability is not based on limiting frequencies, but is more general and includes *degrees of belief.* It can therefore be used for experiments which cannot be repeated, where a frequency definition of probability would not be applicable (for example, one can consider the probability that it will rain tomorrow). Bayesian methods also allow for a natural way to input additional information such as physical boundaries and subjective information; in fact they *require* as input the *prior distribution* for any parameter to be estimated.

The Bayesian methodology, while well adapted to decision-making situations, is not in general appropriate for the objective presentation of experimental data. This can be seen from the following example.

An experiment sets out to measure the value of a parameter whose true value cannot be negative (such as the neutrino mass squared), but let us assume that the true value is in fact zero. We should then expect that about half of the time, an unbiased experimental measurement should yield a negative (unphysical) result. Now if our experiment produces a negative result, the question arises what value to report. If we wish to make a decision concerning the most likely value of this parameter, we would use a Bayesian approach which would assure that the reported value is positive, since it would be nonsense to assert that the most likely value is one which cannot be true. On the other hand, if we wish to report an unbiased result which can be combined with other measurements, it is better to report the unphysical result. Everyone understands what it means to quote a result of, for example, $m^2 = -1.2 \pm 2.0\ \mathrm{eV}^2$. This result could then be averaged with other results, half of which would be positive, and the average would eventually converge toward zero, the true value. If Bayesian estimates are averaged, they do not converge to the true value, since they have all been forced to be positive. (In Bayesian theory, the proper way to combine experiments invokes only one use of the prior. Thus it is problematic to combine properly several experiments which report only the estimate and the uncertainty; each has independently invoked a prior.)

28.6.3. *Frequentist, or classical confidence intervals*: As the name implies, the Frequentist concept of probability is based entirely on the limiting frequency, so it only makes sense in situations where experiments are repeatable, at least in principle. This is clearly the case for the kind of data we are concerned with, and the methods we present here are based on the Frequentist point of view.

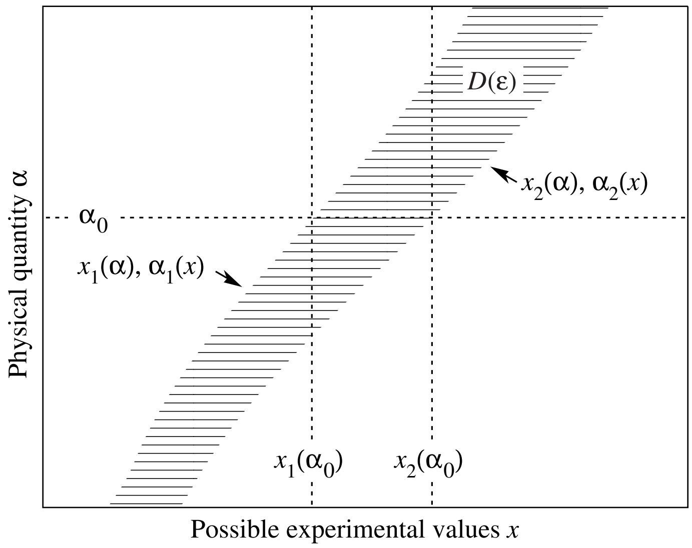

Figure 28.1: Confidence intervals for a single unknown parameter α. One might think of the p.d.f. $f(x;\alpha)$ as being plotted out of the paper as a function of x along each horizontal line of constant α. The domain $D(\varepsilon)$ contains a fraction $1-\varepsilon$ of the area under each of these functions.

The classical construction of exact confidence intervals which we describe here was first proposed by Neyman [11]. We consider the parameter α whose true value is fixed but unknown. The properties of our experimental apparatus are expressed in the function $f(x;\alpha)$ which gives the probability of observing data x if the true value of the parameter is α. This function must be known in order to interpret the results of an experiment. For a large complex experiment, f is usually determined numerically using Monte Carlo simulation.

Given $f(x;\alpha)$, we can find for every value of α, two values $x_1(\alpha,\varepsilon)$ and $x_2(\alpha,\varepsilon)$ such that

$$P(x_1 < x < x_2;\alpha) \;=\; 1-\varepsilon \;=\; \int_{x_1}^{x_2} f(x;\alpha)\,dx\ . \tag{28.33}$$

This is shown graphically in Fig. 28.1: a horizontal line segment $[x_1(\alpha,\varepsilon), x_2(\alpha,\varepsilon)]$ is drawn for representative values of α. The union of all intervals $[x_1(\alpha,\varepsilon), x_2(\alpha,\varepsilon)]$, designated in the figure as the domain $D(\varepsilon)$, is known as the *confidence belt.* Typically the curves $x_1(\alpha,\varepsilon)$ and $x_2(\alpha,\varepsilon)$ are monotonic functions of α, which we assume for this discussion.

Upon performing an experiment to measure x and obtaining the value x_0, one draws a vertical line through x_0 on the horizontal axis. The confidence interval for α is the union of all values of α for which the corresponding line segment $[x_1(\alpha, \varepsilon), x_2(\alpha, \varepsilon)]$ is intercepted by this vertical line. The confidence interval is an interval $[\alpha_1(x_0), \alpha_2(x_0)]$, where $\alpha_1(x_0)$ and $\alpha_2(x_0)$ are on the boundary of $D(\varepsilon)$. Thus, the boundaries of $D(\varepsilon)$ can be considered to be functions $x(\alpha)$ when constructing D, and then to be functions $\alpha(x)$ when reading off confidence intervals.

Such confidence intervals are said to have Confidence Level (CL) equal to $1-\varepsilon$.

Now suppose that some unknown particular value of α, say α_0 (indicated in the figure), is the true value of α. We see from the figure that α_0 lies between $\alpha_1(x)$ and $\alpha_2(x)$ if and only if x lies between $x_1(\alpha_0)$ and $x_2(\alpha_0)$. Thus we can write:

$$P\big[x_1(\alpha_0) < x < x_2(\alpha_0)\big] \;=\; 1-\varepsilon \;=\; P\big[\alpha_2(x) < \alpha_0 < \alpha_1(x)\big] \; . \tag{28.34}$$

And since, by construction, this is true for any value α_0, we can drop the subscript 0 and obtain the relationship we wanted to establish for the probability that the confidence limits will contain the true value of α:

$$P[\alpha_2(x) < \alpha < \alpha_1(x)] \;=\; 1-\varepsilon \; . \tag{28.35}$$

In this probability statement, α_1 and α_2 are the random variables (not α), and we can verify that the statement is true, as a limiting ratio of frequencies in random experiments, for any assumed value of α. In a particular real experiment, the numerical values α_1 and α_2 are determined by applying the algorithm to the real data, and the probability statement is (all too frequently) misinterpreted to be a statement about the true value α since this is the only unknown remaining in the equation. It should however be interpreted as the probability of obtaining values α_1 and α_2 which include the true value of α, in an ensemble of identical experiments. Any method which gives confidence intervals that contain the true value with probability $1-\varepsilon$ (no matter what the true value of α is) is said to have the correct *coverage*. The frequentist intervals as constructed above have the correct *coverage* by construction. Coverage is considered to be a critical property of confidence intervals [7]. (Power to exclude false values of α, related to the length of the intervals in a relevant measure, is also important.)

The condition of coverage Eq. (28.33) does not determine x_1 and x_2 uniquely, since any range which gives the desired value of the integral would give the same coverage. Additional criteria are thus needed. The most common criterion is to choose *central intervals* such that the area of the excluded tail on either side is $\varepsilon/2$. This criterion is sufficient in most cases, but there is a more general *ordering principle* which reduces to centrality in the usual cases and produces confidence intervals with better properties when in the neighborhood of a physical limit. This ordering principle, which consists of taking the interval which includes the largest values of a likelihood ratio, is described by Feldman and Cousins [12].

For the problem of a counting rate experiment in the presence of background, Roe and Woodroofe [13] have proposed a modification to Ref. [12] incorporating *conditioning*, i.e., conditional probabilities computed using constraints on the number of background events actually observed.

28.6.4. *Gaussian errors*:

If the data are such that the distribution of the estimator(s) satisfies the central limit theorem discussed in Sec. 27.3.3, the function $f(x;\alpha)$ is the Gaussian distribution. If there is more than one parameter being estimated, the multivariate Gaussian is used. For the univariate case with known σ,

$$1-\varepsilon = \int_{\mu-\delta}^{\mu+\delta} e^{\dfrac{-(x-\mu)^2}{2\sigma^2}}\, dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{28.36}$$

is the probability that the measured value x will fall within $\pm\delta$ of the true value μ. From the symmetry of the Gaussian with respect to x and μ, this is also the probability that the true value will be within $\pm\delta$ of the measured value. Fig. 28.2 shows a $\delta = 1.64\sigma$ confidence interval unshaded. The choice $\delta = \sqrt{\mathrm{Var}(\,\mu\,)} \equiv \sigma$ 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 28.1.

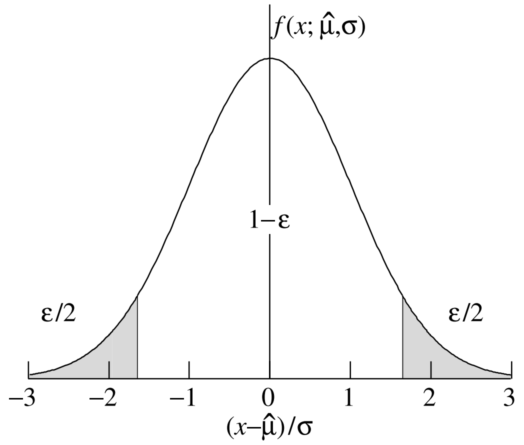

Figure 28.2: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by ε, are as shown.

For other δ, find ε as the ordinate of Fig. 27.1 on the $n=1$ curve at $\chi^2 = (\delta/\sigma)^2$. We can set a one-sided (upper or lower) limit by excluding above $\mu+\delta$ (or below $\mu-\delta$); ε's for such limits are 1/2 the values in Table 28.1.

For multivariate $\boldsymbol{\alpha}$ the scalar $\mathrm{Var}(\mu)$ becomes a full variance-covariance matrix. Assuming a multivariate Gaussian, Eq. (27.22), and subsequent discussion the standard error ellipse for the pair $(\,\widehat{\alpha}_m, \widehat{\alpha}_n)$ may be drawn as in Fig. 28.3.

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, where the ellipse is at a constant value of $\chi^2 = \chi^2_{\mathrm{min}} + 1$

Table 28.1: 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

or $\ln\mathscr{L} = \ln\mathscr{L}_{\rm max} - 1/2$. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{mn}\ \sigma_m\ \sigma_n}{\sigma_m^2 - \sigma_n^2}\ . \tag{28.37}$$

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.

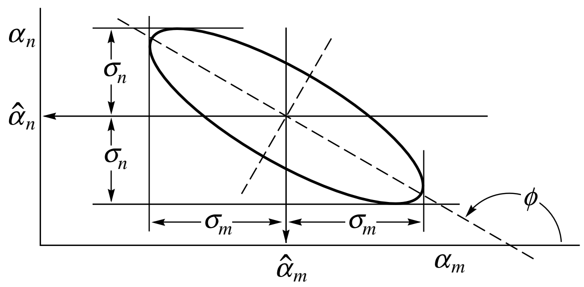

Figure 28.3: Standard error ellipse for the estimators $\widehat{\alpha}_m$ and $\widehat{\alpha}_n$. In this case the correlation is negative.

For any unbiased procedure (*e.g.*, least squares or maximum likelihood) used to estimate k parameters α_i, $i = 1,\ldots,k$, the probability $1-\varepsilon$ that the true values of all k parameters lie within an ellipsoid bounded by a fixed value of $\Delta\chi^2 = \chi^2 - \chi^2_{\rm min}$ may be found from Fig. 27.1. This is because the *difference*, $\Delta\chi^2 = \chi^2 - \chi^2_{\rm min}$, obeys the "$\chi^2$" p.d.f. given in Table 27.1, if the parameter n in the formula is taken to be k (rather than degrees-of-freedom in the fit). In Fig. 27.1, read the ordinate as ε and the abscissa as $\Delta\chi^2$. The correct values of ε are on the $n = k$ curve. For $k > 1$, the values of ε for given $\Delta\chi^2$ are much greater than for $k = 1$. Hence, using $\Delta\chi^2 = s^2$, which gives s-standard-deviation errors on a single parameter (*irrespective of the other parameters*), is not appropriate for a multi-dimensional ellipsoid. For example, for $k = 2$, the probability $(1-\varepsilon)$ 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 only 39%.

Table 28.2: $\Delta\chi^2$ corresponding to $(1-\varepsilon)$, for joint estimation of k parameters.

$(1-\varepsilon)$ (%)	$k=1$	$k=2$	$k=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

Values of $\Delta\chi^2$ corresponding to commonly used values of ε and k are given in Table 28.2. These probabilities assume Gaussian errors, unbiased estimators, and that the model describing the data in terms of the α_i is correct. When these assumptions are not satisfied, a Monte Carlo simulation is typically performed to determine the relation between $\Delta\chi^2$ and ε.

28.6.5. ***Upper limits and two-sided intervals***:

When a measured value is close to a physical boundary, it is natural to report a one-sided confidence interval (often an upper limit). It is straightforward to force the procedure of Sec. 28.6.3 to produce only an upper limit, by setting $x_2 = \infty$ in Eq. (28.33). Then x_1 is uniquely determined. Clearly this procedure will have the desired coverage, but *only if we always choose to set an upper limit.* In practice one might decide after seeing the data whether to set an upper limit or a two-sided limit. In this case the upper limits calculated by Eq. (28.33) will not give exact coverage, as has been noted in Ref. 12.

In order to correct this problem and assure coverage in all circumstances, it is necessary to adopt a *unified procedure*, that is, a single ordering principle which will provide coverage globally. Then it is the *ordering principle* which decides whether a one-sided or two-sided interval will be reported for any given set of data. The appropriate unified procedure and ordering principle are given in Ref. 12. We reproduce below the main results.

28.6.6. ***Gaussian data close to a boundary***:

One of the most controversial statistical questions in physics is how to report a measurement which is close to the edge or even outside of the allowed physical region. This is because there are several admissible possibilities depending on how the result is to be used or interpreted. Normally one or more of the following should be reported:

(a) The actual measurement should be reported, even if it is outside the physical region. As with any other measurement, it is best to report the value of a quantity which is nearly Gaussian distributed if possible. Thus one may choose to report mass squared rather than mass, or $\cos\theta$ rather than θ. For a complex quantity z close to zero, report $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ rather than amplitude and phase of z. Data carefully reported in this way can be unbiased, objective, easily interpreted and combined (averaged) with other data in a straightforward way, even if they lie partly or wholly outside the physical region. The reported error is a direct measure of the intrinsic accuracy of the result, which cannot always be inferred from the upper limits proposed below.

(b) If the data are to be used to make a decision, for example to determine the dimensions of a new experimental apparatus for an improved measurement, it may be appropriate to report a Bayesian upper limit, which must necessarily contain subjective feelings about the possible values of the parameter, as well as containing information about the physical boundary. Its interpretation requires knowledge of the prior distribution which was necessarily used to obtain it.

(c) If it is desired to report an upper limit in an objective way such that it has a well-defined statistical meaning in terms of a limiting frequency, then report the Frequentist confidence bound(s) as given by the unified Feldman-Cousins approach. This algorithm always gives a non-null interval (that is, the confidence limits are always inside the physical region, even for a measurement well outside the physical region), and still has correct global coverage. These confidence limits for a Gaussian measurement close to a non-physical boundary are summarized in Fig. 28.4. Additional tables are given in Ref. 12. It will also be useful to accumulate experience with the modification proposed by Roe and Woodroofe [13].

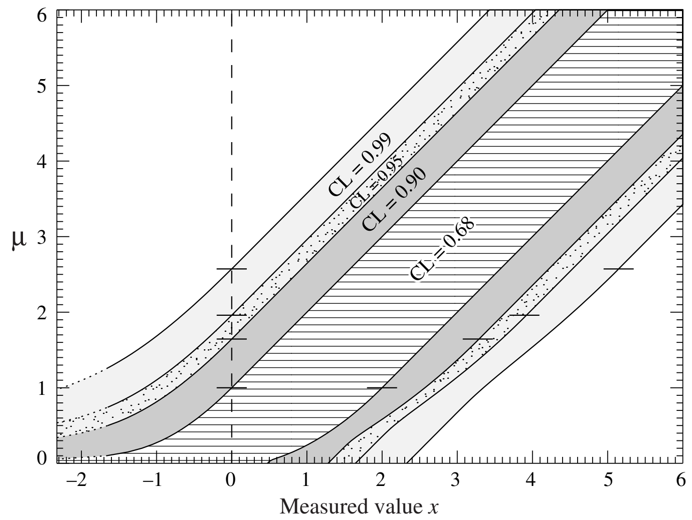

Figure 28.4: Plot of 99%, 95%, 90%, and 68.27% ("one σ") confidence intervals for a physical quantity μ based on a Gaussian measurement x (in units of standard deviations), for the case where the true value of μ cannot be negative. The curves become straight lines above the horizontal tick marks. The probability of obtaining an experimental value at least as negative as the left edge of the graph ($x = -2.33$) is less than 1%. Values of x more negative than -1.64 (dotted segments) are less than 5% probable, no matter what the true value of μ.

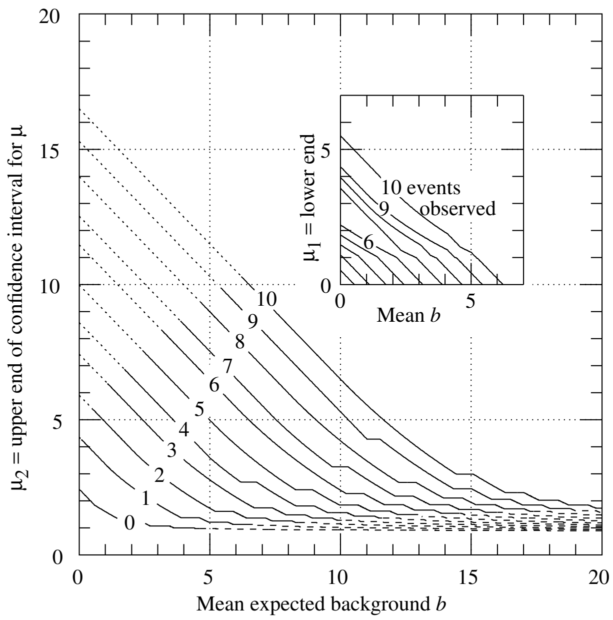

Figure 28.5: 90% confidence intervals $[\mu_1, \mu_2]$on the number of signal events as a function of the expected number of background events b. For example, if the expected background is 8 events and 5 events are observed, then the signal is 2.60 or less with 90% confidence. Dotted portions of the μ_2 curves on the upper left indicate regions where μ_1 is non-zero (as shown by the inset). Dashed portions in the lower right indicate regions where the probability of obtaining the number of events observed or fewer is less than 1%, even if $\mu = 0$. Horizontal curve sections occur because of discrete number statistics. Tables showing these data as well as the CL = 68.27%, 95%, and 99% results are given in Ref. 12.

28.6.7. *Poisson data for small samples*:

When the observable is restricted to integer values (as in the case of Poisson and binomial distributions), it is not generally possible to construct confidence intervals with exact coverage for all values of α. In these cases the integral in Eq. (28.33) becomes a sum of finite contributions and it is no longer possible (in general) to find consecutive terms which add up exactly to the required confidence level $1 - \varepsilon$ for all values of α. Thus one constructs intervals which happen to have exact coverage for a few values of α, and unavoidable over-coverage for all other values.

In addition to the problem posed by the discreteness of the data, we usually have to contend with possible background whose expectation must be evaluated separately and may not be known precisely. For these reasons, the reporting of this kind of data is even more controversial than the Gaussian data near a boundary as discussed above. This is especially true when the number of observed counts is greater than the expected background. As for the Gaussian case, there are at least three possibilities for reporting such results depending on how the result is to be used:

(a) The actual measurements should be reported, which means (1) the number of recorded counts, (2) the expected background, possibly with its error, and (3) normalization factor which turns the number of counts into a cross section, decay rate, *etc.* As with Gaussian data, these data can be combined with that of other experiments, to make improved upper limits for example.

(b) A Bayesian upper limit may be reported. This has the advantages and disadvantages of any Bayesian result as discussed above. It is especially difficult to find an acceptable prior probability distribution for this case.

(c) An upper limit (or confidence region) with optimal coverage can be reported using the unified approach of Ref. 12. At the moment these confidence limits have been calculated only for the case of exactly known background expectation. The main results can be read from Fig. 28.5 or from Table 28.3; more extensive tables can be found in Ref. 12.

None of the above gives a single number which quantifies the quality or sensitivity of the experiment. This is a serious shortcoming of most upper limits including those of method (c), since it is impossible to distinguish, from the upper limit alone, between a clean experiment with no background and a lucky experiment with fewer observed counts than expected background. For this reason, we suggest that in addition to (a) and (c) above, a measure of the sensitivity should be reported whenever expected background is larger or comparable to the number of observed counts. The best such measure we know of is that proposed and tabulated in Ref. 12, defined as the average upper limit that would be attained by an ensemble of experiments with the expected background and no true signal.

References:

1. A. Stuart and A. K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2 *Classical Inference and Relationship* 5th Ed., (Oxford Univ. Press, 1991), and earlier editions by Kendall and Stuart.
2. W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics* (North Holland, Amsterdam and London, 1971).
3. H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press, New Jersey (1958).
4. B.P. Roe, *Probability and Statistics in Experimental Physics,* (Springer-Verlag, New York, 208 pp., 1992).
5. G. Cowan, *Statistical Data Analysis* (Oxford University Press, Oxford, 1998).
6. For a review, see S. Baker and R. Cousins, Nucl. Instrum. Methods **221**, 437 (1984).
7. R.D. Cousins, Am. J. Phys. **63**, 398 (1995).
8. W.H. Press *et al.*, *Numerical Recipes* (Cambridge University Press, New York, 1986).
9. F. James and M. Roos, "MINUIT, Function Minimization and Error Analysis," CERN D506 (Long Writeup). Available from the CERN Program Library Office, CERN-IT Division, CERN, CH-1211, Geneva 21, Switzerland.
10. B. Efron, Am. Stat. **40**, 11 (1986).

Table 28.3: Poisson limits $[\mu_1, \mu_2]$ for n_0 observed events in the absence of background.

	CI = 90%		CI = 95%	
n_0	μ_1	μ_2	μ_1	μ_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

11. J. Neyman, Phil. Trans. Royal Soc. London, Series A, **236**, 333 (1937), reprinted in *A Selection of Early Statistical Papers on J. Neyman* (University of California Press, Berkeley, 1967).
12. G.J. Feldman and R.D. Cousins, Phys. Rev. **D57**, 3873 (1998).
13. B.P. Roe and M.B. Woodroofe, Phys. Rev. **D60**, 053009 (1999).
14. F. James and M. Roos, Phys. Rev. **D44**, 299 (1991).