31. STATISTICS

Revised April 1998 by F. James (CERN); February 2000 by R. Cousins (UCLA); October 2001 by G. Cowan (RHUL).

A probability density function (p.d.f.) $f(x;\theta)$ with known parameter θ (or parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$) 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 statistics we are concerned with the inverse problem, that of making inferences about the parameters from observed data.

There are two main approaches to statistical inference, which we may call frequentist and Bayesian. In frequentist statistics, probability is interpreted as the frequency of the outcome of a repeatable experiment. Estimators are used to measure values for unknown parameters, and confidence intervals can be constructed which contain the unknown true value of a parameter with a specified probability. Statistical tests can be constructed which, depending on the outcome of the experiment, accept or reject hypotheses. One does not, however, define a probability for a parameter θ, which is treated as a constant whose value may be unknown.

In Bayesian statistics, the interpretation of probability is more general and includes *degree of belief*. One can then speak of a p.d.f. for a parameter θ, which expresses one's state of knowledge about where its true value lies. Bayesian methods allow for a natural way to input additional information such as physical boundaries and subjective information; in fact they *require* as input the *prior* p.d.f. for the parameters, *i.e.*, the degree of belief about the parameters' values before carrying out the measurement. Using Bayes' theorem Eq. (30.4), the prior degree of belief is updated by the data from the experiment.

For many inference problems, the frequentist and Bayesian approaches give the same numerical answers, even though they are based on fundamentally different interpretations of probability. For small data samples, however, and for measurements of a parameter near a physical boundary, the different approaches may yield different results, so we are forced to make a choice. For a discussion of Bayesian vs. non-Bayesian methods, see References written by a statistician[1], by a physicist[2], or the more detailed comparison in Ref. [3].

Frequentist statistics provides the usual tools for reporting objectively the outcome of an experiment without needing to incorporate prior beliefs concerning the parameter being measured or the theory being tested. Bayesian techniques, on the other hand, are often used to treat systematic uncertainties, where the author's subjective beliefs about, say, the accuracy of the measuring device may enter. Bayesian statistics also provides a useful framework for discussing the validity of different theoretical interpretations of the data. This aspect of a measurement, however, will usually be treated separately from the reporting of the result.

31.1. Parameter estimation

Here we review *point estimation* of parameters. An *estimator* $\widehat{\theta}$ (written with a hat) is a function of the data whose value, the *estimate*, is intended as a meaningful guess for the value of the parameter θ.

There is no fundamental rule dictating how an estimator must be constructed. One tries therefore to choose that estimator which has the best properties. The most important of these are (a) *consistency*, (b) *bias*, (c) *efficiency*, and (d) *robustness*.

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

(b) The *bias*, $b = E[\widehat{\theta}] - \theta$, is the difference between the expectation value of the estimator and the true value of the parameter. The expectation value is taken over a hypothetical set of similar experiments in which $\widehat{\theta}$ is constructed in the same way. When $b = 0$ the estimator is said to be unbiased. The bias depends on the chosen metric, *i.e.*, if $\widehat{\theta}$ is an unbiased estimator of θ, then $\widehat{\theta}^2$ is not in general an unbiased estimator for θ^2. If we have an estimate $\widehat{b}$ for the bias we can subtract it from $\widehat{\theta}$ to obtain a new $\widehat{\theta}' = \widehat{\theta} - \widehat{b}$. The estimate $\widehat{b}$ may, however, be subject to uncertainties that are larger than the bias itself.

(c) *Efficiency* is the inverse of the ratio of the variance $V[\widehat{\theta}]$ to its minimum possible value. Under rather general conditions, the minimum variance is given by the Rao-Cramér-Frechet bound,

$$\sigma^2_{\rm min} = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / I(\theta) \ , \tag{31.1}$$

where

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \sum_i \ln f(x_i; \theta)\right)^2\right] \tag{31.2}$$

is the *Fisher information*. The sum is over all data, assumed independent and distributed according to the p.d.f. $f(x;\theta)$, b is the bias, if any, and the allowed range of x must not depend on θ.

The *mean-squared error*,

$$\text{MSE} = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + b^2 \ , \tag{31.3}$$

is a convenient quantity which combines the errors due to bias and variance.

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

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

31.1.1. ***Estimators for mean, variance and median***:

Suppose we have a set of N independent measurements x_i assumed to be unbiased measurements of the same unknown quantity μ with a common, but unknown, variance σ^2. Then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{31.4}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 \tag{31.5}$$

are unbiased estimators of μ and σ^2. The variance of $\widehat{\mu}$ is σ^2/N and the variance of $\widehat{\sigma^2}$ is

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right) , \tag{31.6}$$

where m_4 is the 4th central moment of x. For Gaussian distributed x_i this becomes $2\sigma^4/(N-1)$ for any $N \geq 2$, and for large N the standard deviation of $\widehat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. Again if the x_i are Gaussian, $\widehat{\mu}$ is an efficient estimator for μ and the estimators $\widehat{\mu}$ and $\widehat{\sigma^2}$ are uncorrelated. Otherwise the arithmetic mean (31.4) is not necessarily the most efficient estimator; this is discussed in more detail in [4] Sec. 8.7

If σ^2 is known, it does not improve the estimate $\widehat{\mu}$, as can be seen from Eq. (31.4); however, if μ is known, substitute it for $\widehat{\mu}$ in Eq. (31.5) and replace $N-1$ by N to obtain a somewhat better estimator of σ^2.

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

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{N} w_i x_i \tag{31.7}$$

is an unbiased estimator for μ with a smaller variance than an unweighted average; here $w_i = 1/\sigma_i^2$ and $w = \sum_i w_i$. The standard deviation of $\widehat{\mu}$ is $1/\sqrt{w}$.

As an estimator for the median $x_{\rm med}$ one can use the value $\widehat{x}_{\rm med}$ such that half the x_i are below and half above (the sample median). If the sample median lies between two observed values, it is set by convention halfway between them. If the p.d.f. of x has the form $f(x-\mu)$ and μ is both mean and median, then for large N the variance of the sample median approaches $1/[4Nf^2(0)]$, provided $f(0) > 0$. Although estimating the median can often be more difficult computationally than the mean, the resulting estimator is generally more robust, as it is insensitive to the exact shape of the tails of a distribution.

31.1.2. ***The method of maximum likelihood***:

"From a theoretical point of view, the most important general method of estimation so far known is the *method of maximum likelihood*" [5]. We suppose that a set of N independently measured quantities x_i came from a p.d.f. $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is set of n parameters whose values are unknown. The method of maximum likelihood takes the estimators $\widehat{\boldsymbol{\theta}}$ to be those values of $\boldsymbol{\theta}$ that maximize the *likelihood function*,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i;\boldsymbol{\theta}) \ . \tag{31.8}$$

The likelihood function is the joint p.d.f. for the data, evaluated with the data obtained in the experiment and regarded as a function of the parameters. Note that the likelihood function is *not* a p.d.f. for the parameters $\boldsymbol{\theta}$; in frequentist statistics this is not defined. In Bayesian statistics one can obtain from the likelihood the posterior p.d.f. for $\boldsymbol{\theta}$, but this requires multiplying by a prior p.d.f. (see Sec. 31.4.1).

It is usually easier to work with $\ln L$, and since both are maximized for the same parameter values $\boldsymbol{\theta}$, the maximum likelihood (ML) estimators can be found by solving the *likelihood equations*,

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \ , \qquad i = 1, \ldots, n \ . \tag{31.9}$$

Maximum likelihood estimators are important because they are approximately unbiased and efficient for large data samples, under quite general conditions, and the method has a wide range of applicability.

In evaluating the likelihood function, it is important that any normalization factors in the p.d.f. that involve $\boldsymbol{\theta}$ be included. However, we will only be interested in the maximum of L and in ratios of L at different values of the parameters; hence any multiplicative factors that do not involve the parameters that we want to estimate may be dropped, including factors that depend on the data but not on $\boldsymbol{\theta}$.

Under a one-to-one change of parameters from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$, the ML estimators $\widehat{\boldsymbol{\theta}}$ transform to $\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})$. That is, the ML solution is invariant under change of parameter. However, other properties of ML estimators, in particular the bias, are not invariant under change of parameter.

The inverse V^{-1} of the covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$ for a set of ML estimators can be estimated by using

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\theta}} \ . \tag{31.10}$$

For finite samples, however, Eq. (31.10) can result in an underestimate of the variances. In the large sample limit (or in a linear model with Gaussian errors), L has a Gaussian form and $\ln L$ is (hyper)parabolic. In this case it can be seen that a numerically equivalent

way of determining s-standard-deviation errors is from the contour given by the $\boldsymbol{\theta}'$ such that

$$\ln L(\boldsymbol{\theta}') = \ln L_{\max} - s^2/2 \ , \tag{31.11}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (31.46)). The extreme limits of this contour on the θ_i axis give an approximate s-standard-deviation confidence interval for θ_i (see Section 31.4.2.3).

In the case where the size n of the data sample $x_1, \ldots, x_n$ is small, the unbinned maximum likelihood method is preferred, since binning can only result in a loss of information. The sample size n can be regarded as fixed or the user can choose to treat it as a Poisson-distributed variable; this latter option is sometimes called "extended maximum likelihood" (see, *e.g.*, [6, 7, 8]). If the sample is large it can be convenient to bin the values in a histogram, so that one obtains a vector of data $\boldsymbol{n} = (n_1, \ldots, n_N)$ with expectation values $\boldsymbol{\nu} = E[\boldsymbol{n}]$ and probabilities $f(\boldsymbol{n}; \boldsymbol{\nu})$. Then one may maximize the likelihood function based on the contents of the bins (so i labels bins). This is equivalent to maximizing the likelihood ratio $\lambda(\boldsymbol{\theta}) = f(\boldsymbol{n}; \boldsymbol{\nu}(\boldsymbol{\theta}))/f(\boldsymbol{n}; \boldsymbol{n})$, or to minimizing the quantity [9]

$$-2 \ln \lambda(\boldsymbol{\theta}) = 2 \sum_{i=1}^{N} \left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln \frac{n_i}{\nu_i(\boldsymbol{\theta})} \right] \ , \tag{31.12}$$

where in bins where $n_i = 0$, the last term in (31.12) is zero.

The reason for minimizing Eq. (31.12) defined in this way is that in the large sample limit, the minimum of $-2 \ln \lambda$ follows a χ^2 distribution and can be used in a test of goodness-of-fit (see Sec. 31.3.2). If there are N bins and m fitted parameters, then the number of degrees of freedom for the χ^2 distribution is $N - m - 1$ if the data are treated as multinomially distributed and $N - m$ if the n_i are Poisson variables with $\nu_{\rm tot} = \sum_i \nu_i$ fixed. If the n_i are Poisson distributed and $\nu_{\rm tot}$ is also fitted, then by minimizing Eq. (31.12) one obtains that the area under the fitted function is equal to the sum of the histogram contents, *i.e.*, $\sum_i \nu_i = \sum_i n_i$. This is not the case for parameter estimation methods based on a least-squares procedure with traditional weights (see, *e.g.*, Ref. [8]).

31.1.3. *The method of least squares*:

The *method of least squares* (LS) coincides with the method of maximum likelihood in the following special case. Consider a set of N independent measurements y_i at known points x_i. The measurement y_i is assumed to be Gaussian distributed with mean $F(x_i; \boldsymbol{\theta})$ and known variance σ_i^2. The goal is to construct estimators for the unknown parameters $\boldsymbol{\theta}$. The likelihood function contains the sum of squares

$$\chi^2(\boldsymbol{\theta}) = -2 \ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - F(x_i; \boldsymbol{\theta}))^2}{\sigma_i^2} \ . \tag{31.13}$$

The set of parameters $\boldsymbol{\theta}$ which maximize L is the same as those which minimize χ^2.

The minimum of Equation (31.13) defines the least-squares estimators $\widehat{\boldsymbol{\theta}}$ for the more general case where the y_i are not Gaussian distributed as long as they are independent. If they are not independent but rather have a covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta})) \ , \tag{31.14}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)$ is the vector of measurements, $\boldsymbol{F}(\boldsymbol{\theta})$ is the corresponding vector of predicted values (understood as a column vector in (31.14)), and the superscript T denotes transposed (*i.e.*, row) vector.

In many practical cases one further restricts the problem to the situation where $F(x_i; \boldsymbol{\theta})$ is a linear function of the parameters, *i.e.*,

$$F(x_i; \boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) \ . \tag{31.15}$$

Here the $h_j(x)$ are m linearly independent functions, *e.g.*, $1, x, x^2, \ldots, x^{m-1}$, or Legendre polynomials. We require $m < N$ and at least m of the x_i must be distinct.

Minimizing χ^2 in this case with m parameters reduces to solving a system of m linear equations. Defining $H_{ij} = h_j(x_i)$ and minimizing χ^2 by setting its derivatives with respect to the θ_i equal to zero gives the LS estimators,

$$\widehat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y} \ . \tag{31.16}$$

The covariance matrix for the estimators $U_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$ is given by

$$U = DVD^T = (H^T V^{-1} H)^{-1} \ , \tag{31.17}$$

or equivalently, its inverse U^{-1} can be found from

$$(U^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\widehat{\theta}} = \sum_{k,l=1}^{N} h_i(x_k)(V^{-1})_{kl} h_j(x_l) \ . \tag{31.18}$$

The LS estimators can also be found from the expression

$$\widehat{\boldsymbol{\theta}} = U\boldsymbol{g} \ , \tag{31.19}$$

where the vector $\boldsymbol{g}$ is defined by

$$g_i = \sum_{j,k=1}^{N} y_j h_i(x_k)(V^{-1})_{jk} \ . \tag{31.20}$$

For the case of uncorrelated y_i, for example, one can use (31.19) with

$$(U^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k)h_j(x_k)}{\sigma_k^2} \ , \tag{31.21}$$

$$g_i = \sum_{k=1}^{N} \frac{y_k h_i(x_k)}{\sigma_k^2} \ . \tag{31.22}$$

Expanding $\chi^2(\boldsymbol{\theta})$ about $\widehat{\boldsymbol{\theta}}$, one finds that the contour in parameter space defined by

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\widehat{\boldsymbol{\theta}}) + 1 = \chi^2_{\rm min} + 1 \tag{31.23}$$

has tangent planes located at plus or minus one standard deviation $\sigma_{\widehat{\theta}}$ from the LS estimates $\widehat{\boldsymbol{\theta}}$.

In constructing the quantity $\chi^2(\boldsymbol{\theta})$, one requires the variances or, in the case of correlated measurements, the covariance matrix. Often these quantities are not known *a priori* and must be estimated from the data; an important example is where the measured value y_i represents a counted number of events in the bin of a histogram. If, for example, y_i represents a Poisson variable, for which the variance is equal to the mean, then one can either estimate the variance from the predicted value, $F(x_i;\boldsymbol{\theta})$, or from the observed number itself, y_i. In the first option, the variances become functions of the fitted parameters, which may lead to calculational difficulties. The second option can be undefined if y_i is zero, and in both cases for small y_i the variance will be poorly estimated. In either case one should constrain the normalization of the fitted curve to the correct value, *e.g.*, one should determine the area under the fitted curve directly from the number of entries in the histogram (see [8] Section 7.4). A further alternative is to use the method of maximum likelihood; for binned data this can be done by minimizing Eq. (31.12)

As the minimum value of the χ^2 represents the level of agreement between the measurements and the fitted function, it can be used for assessing the goodness-of-fit; this is discussed further in Section 31.3.2.

31.2. Propagation of errors

Consider a set of n quantities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and a set of m functions $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_m(\boldsymbol{\theta}))$. Suppose we have estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, using, say, maximum likelihood or least squares, and we also know or have estimated the covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$. The goal of *error propagation* is to determine the covariance matrix for the functions, $U_{ij} = \mathrm{cov}[\widehat{\eta}_i, \widehat{\eta}_j]$, where $\widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})$. In particular, the diagonal elements $U_{ii} = V[\widehat{\eta}_i]$ give the variances. The new covariance matrix can be found by expanding the functions $\boldsymbol{\eta}(\boldsymbol{\theta})$ about the estimates $\widehat{\boldsymbol{\theta}}$ to first order in a Taylor series. Using this one finds

$$U_{ij} \approx \sum_{k,l} \left.\frac{\partial \eta_i}{\partial \theta_k}\frac{\partial \eta_j}{\partial \theta_l}\right|_{\widehat{\theta}} V_{kl} \ . \tag{31.24}$$

This can be written in matrix notation as $U \approx AVA^T$ where the matrix of derivatives A is

$$A_{ij} = \left.\frac{\partial \eta_i}{\partial \theta_j}\right|_{\widehat{\theta}} \tag{31.25}$$

and A^T is its transpose. The approximation is exact if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is linear (it holds, for example, in equation (31.17)). If this is not the case the approximation can break down if, for example, $\boldsymbol{\eta}(\boldsymbol{\theta})$ is significantly nonlinear close to $\widehat{\boldsymbol{\theta}}$ in a region of a size comparable to the standard deviations of $\widehat{\boldsymbol{\theta}}$.

31.3. Statistical tests

In addition to estimating parameters, one often wants to assess the validity of certain statements concerning the data's underlying distribution. *Hypothesis tests* provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. In *goodness-of-fit tests* one gives the probability to obtain a level of incompatibility with a certain hypothesis that is greater than or equal to the level observed with the actual data.

31.3.1. *Hypothesis tests*:

Consider an experiment whose outcome is characterized by a vector of data $\boldsymbol{x}$. A *hypothesis* is a statement about the distribution of $\boldsymbol{x}$. It could, for example, define completely the p.d.f. for the data (a simple hypothesis) or it could specify only the functional form of the p.d.f., with the values of one or more parameters left open (a composite hypothesis).

A *statistical test* is a rule that states for which values of $\boldsymbol{x}$ a given hypothesis (often called the null hypothesis, H_0) should be rejected. This is done by defining a region of $\boldsymbol{x}$-space called the critical region; if the outcome if the experiment lands in this region, H_0 is rejected. Equivalently one can say that the hypothesis is accepted if $\boldsymbol{x}$ is observed in the acceptance region, *i.e.*, the complement of the critical region. Here 'accepted' is understood to mean simply that the test did not reject H_0.

Rejecting H_0 if it is true is called an error of the first kind. The probability for this to occur is called the *significance level* of the test, α, which is often chosen to be equal to some pre-specified value. It can also happen that H_0 is false and the true hypothesis is given by some alternative, H_1. If H_0 is accepted in such a case, this is called an error of the second kind. The probability for this to occur, β, depends on the alternative hypothesis, say, H_1, and $1-\beta$ is called the *power* of the test to reject H_1.

In High Energy Physics the components of $\boldsymbol{x}$ might represent the measured properties of candidate events, and the acceptance region is defined by the cuts that one imposes in order to select events of a certain desired type. That is, H_0 could represent the signal hypothesis, and various alternatives, H_1, H_2, etc., could represent background processes.

Often rather than using the full data sample $\boldsymbol{x}$ it is convenient to define a *test statistic*, t, which can be a single number or in any case a vector with fewer components than $\boldsymbol{x}$. Each hypothesis for the distribution of $\boldsymbol{x}$ will determine a distribution for t, and the acceptance region in $\boldsymbol{x}$-space will correspond to a specific range of values of t. In constructing t one attempts to reduce the volume of data without losing the ability to discriminate between different hypotheses.

In particle physics terminology, the probability to accept the signal hypothesis, H_0, is the selection efficiency, *i.e.*, one minus the significance level. The efficiencies for the various background processes are given by one minus the power. Often one tries to construct a test to minimize the background efficiency for a given signal efficiency. The *Neyman–Pearson lemma* states that this is done by defining the acceptance region such that, for $\boldsymbol{x}$ in that region, the ratio of p.d.f.s for the hypotheses H_0 and H_1,

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_0)}{f(\boldsymbol{x}|H_1)} \ , \tag{31.26}$$

is greater than a given constant, the value of which is chosen to give the desired signal efficiency. This is equivalent to the statement that (31.26) represents the test statistic with which one may obtain the highest purity sample for a given signal efficiency. It can be difficult in practice, however, to determine $\lambda(\boldsymbol{x})$, since this requires knowledge of the joint p.d.f.s $f(\boldsymbol{x}|H_0)$ and $f(\boldsymbol{x}|H_1)$. Instead, test statistics based on *neural networks* or *Fisher discriminants* are often used (see [10]).

31.3.2. *Goodness-of-fit tests*:

Often one wants to quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses. This can be done by defining a *goodness-of-fit statistic*, t, which is a function of the data whose value reflects in some way the level of agreement between the data and the hypothesis. The user must decide what values of the statistic correspond to better or worse levels of agreement with the hypothesis in question; for many goodness-of-fit statistics there is an obvious choice.

The hypothesis in question, say, H_0, will determine the p.d.f. $g(t|H_0)$ for the statistic. The goodness-of-fit is quantified by giving the p-value, defined as the probability to find t in the region of equal or lesser compatibility with H_0 than the level of compatibility observed with the actual data. For example, if t is defined such that large values correspond to poor agreement with the hypothesis, then the p-value would be

$$p = \int_{t_{\rm obs}}^{\infty} g(t|H_0)\, dt \ , \tag{31.27}$$

where $t_{\rm obs}$ is the value of the statistic obtained in the actual experiment. The p-value should not be confused with the significance level of a test or the confidence level of a confidence interval (Section 31.4), both of which are pre-specified constants.

The p-value is a function of the data and is therefore itself a random variable. If the hypothesis used to compute the p-value is true, then for continuous data, p will be

uniformly distributed between zero and one. Note that the p-value is not the probability for the hypothesis; in frequentist statistics this is not defined. Rather, the p-value is the probability, under the assumption of a hypothesis H_0, of obtaining data at least as incompatible with H_0 as the data actually observed.

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 (31.13), which can be used as a goodness-of-fit statistic. It may also happen that no parameters are estimated from the data, but that one simply wants to compare a histogram, *e.g.*, a vector of Poisson distributed numbers $\boldsymbol{n} = (n_1, \ldots, n_N)$, with a hypothesis for their expectation values $\nu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \nu_i$, the χ^2 (31.13) becomes *Pearson's* χ^2 *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \; . \tag{31.28}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct and if the measured values n_i in (31.28) are sufficiently large (in practice, this will be a good approximation if all $n_i > 5$), then the χ^2 statistic will follow the χ^2 p.d.f. with the number of degrees of freedom equal to the number of measurements N minus the number of fitted parameters. The same holds for the minimized χ^2 from Eq. (31.13) if the y_i are Gaussian.

Alternatively one may fit parameters and evaluate goodness-of-fit by minimizing $-2 \ln \lambda$ from Eq. (31.12). One finds that the distribution of this statistic approaches the asymptotic limit faster than does Pearson's χ^2 and thus computing the p-value with the χ^2 p.d.f. will in general be better justified (see [9] and references therein).

Assuming the goodness-of-fit statistic follows a χ^2 p.d.f., the p-value for the hypothesis is then

$$p = \int_{\chi^2}^{\infty} f(z; n_{\rm d})\, dz \; , \tag{31.29}$$

where $f(z; n_{\rm d})$ is the χ^2 p.d.f. and $n_{\rm d}$ is the appropriate number of degrees of freedom. Values can be obtained from Fig. 31.1 or from the CERNLIB routine `PROB`. If the conditions for using the χ^2 p.d.f. do not hold, the statistic can still be defined as before, but its p.d.f. must be determined by other means in order to obtain the p-value, *e.g.*, using a Monte Carlo calculation.

Since the mean of the χ^2 distribution is equal to $n_{\rm d}$, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_{\rm d}$. Hence the quantity $\chi^2/n_{\rm d}$ is sometimes reported. Since the p.d.f. of $\chi^2/n_{\rm d}$ depends on $n_{\rm d}$, however, one must report $n_{\rm d}$ as well in order to make a meaningful statement. The p-values obtained for different values of $\chi^2/n_{\rm d}$ are shown in Fig. 31.2.

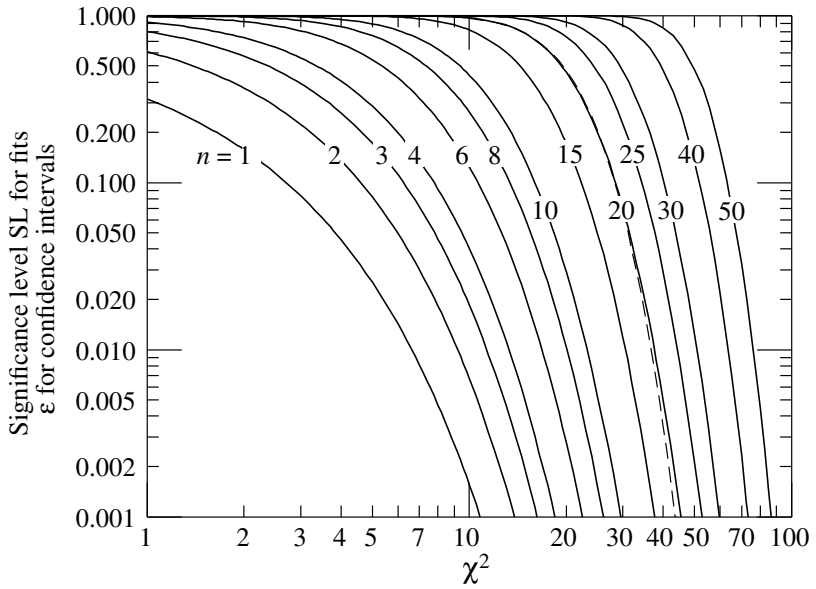

Figure 31.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 31.4.2.3).

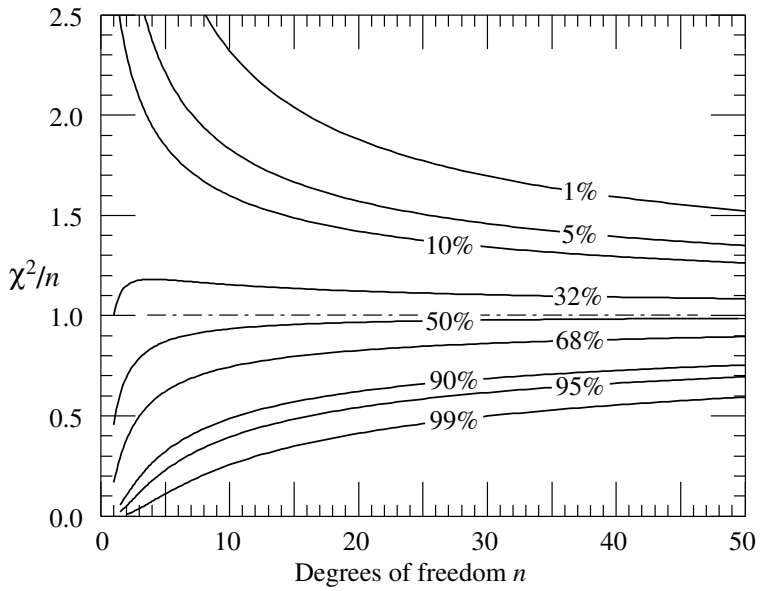

Figure 31.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

31.4. Confidence intervals and limits

When the goal of an experiment is to determine a parameter θ, the result is usually expressed by quoting, in addition to the point estimate, some sort of interval which reflects the statistical precision of the measurement. In the simplest case this can be given by the parameter's estimated value $\widehat{\theta}$ plus or minus an estimate of the standard deviation of $\widehat{\theta}$, $\sigma_{\widehat{\theta}}$. If, however, the p.d.f. of the estimator is not Gaussian or if there are physical boundaries on the possible values of the parameter, then one usually quotes instead an interval according to one of the procedures described below.

In reporting an interval or limit, the experimenter may wish to

- communicate as objectively as possible the result of the experiment;
- provide an interval that is constructed to cover the true value of the parameter with a specified probability;
- provide the information needed by the consumer of the result to draw conclusions about the parameter or to make a particular decision;
- draw conclusions about the parameter that incorporate the author's prior beliefs.

With a sufficiently large data sample, the point estimate and standard deviation (or for the multiparameter case, the parameter estimates and covariance matrix) satisfy essentially all of these goals. For finite data samples, no single method for quoting an interval will achieve all of them. In particular, drawing conclusions about the parameter in the framework of Bayesian statistics necessarily requires subjective input.

In addition to the goals listed above, the choice of method may be influenced by practical considerations such as ease of producing an interval from the results of several measurements. Of course the experimenter is not restricted to quoting a single interval or limit; one may choose, for example, first to communicate the result with a confidence interval having certain frequentist properties, and then in addition to draw conclusions about a parameter using Bayesian statistics. It is recommended, however, that there be a clear separation between these two aspects of reporting a result. In the remainder of this section we assess the extent to which various types of intervals achieve the goals stated here.

31.4.1. *The Bayesian approach*:

Suppose the outcome of the experiment is characterized by a vector of data $\boldsymbol{x}$, whose probability distribution depends on an unknown parameter (or parameters) $\boldsymbol{\theta}$ that we wish to determine. In Bayesian statistics, all knowledge about $\boldsymbol{\theta}$ is summarized by the posterior p.d.f. $p(\boldsymbol{\theta}|\boldsymbol{x})$, which gives the degree of belief for $\boldsymbol{\theta}$ to take on values in a certain region given the data $\boldsymbol{x}$. It is obtained by using Bayes' theorem,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'} \,, \tag{31.30}$$

where $L(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function, *i.e.*, the joint p.d.f. for the data given a certain value of $\boldsymbol{\theta}$, evaluated with the data actually obtained in the experiment, and $\pi(\boldsymbol{\theta})$ is the

prior p.d.f. for $\boldsymbol{\theta}$. Note that the denominator in (31.30) serves simply to normalize the posterior p.d.f. to unity.

Bayesian statistics supplies no fundamental rule for determining $\pi(\boldsymbol{\theta})$; this reflects the experimenter's subjective degree of belief about $\boldsymbol{\theta}$ before the measurement was carried out. By itself, therefore, the posterior p.d.f. is not a good way to report objectively the result of an observation, since it contains both the result (through the likelihood function) and the experimenter's prior beliefs. Without the likelihood function, someone with different prior beliefs would be unable to substitute these to determine his or her own posterior p.d.f. This is an important reason, therefore, to publish wherever possible the likelihood function or an appropriate summary of it. Often this can be achieved by reporting the ML estimate and one or several low order derivatives of L evaluated at the estimate.

In the single parameter case, for example, an interval (called a Bayesian or credible interval) $[\theta_{\rm lo}, \theta_{\rm up}]$ can be determined which contains a given fraction $1-\alpha$ of the probability, *i.e.*,

$$1 - \alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta \ . \tag{31.31}$$

Sometimes an upper or lower limit is desired, *i.e.*, $\theta_{\rm lo}$ can be set to zero or $\theta_{\rm up}$ to infinity. In other cases one might choose $\theta_{\rm lo}$ and $\theta_{\rm up}$ such that $p(\theta|\boldsymbol{x})$ is higher everywhere inside the interval than outside; these are called *highest posterior density* (HPD) intervals. Note that HPD intervals are not invariant under a nonlinear transformation of the parameter.

The main difficulty with Bayesian intervals is in quantifying the prior beliefs. Sometimes one attempts to construct $\pi(\boldsymbol{\theta})$ to represent complete ignorance about the parameters by setting it equal to a constant. A problem here is that if the prior p.d.f. is flat in $\boldsymbol{\theta}$, then it is not flat for a nonlinear function of $\boldsymbol{\theta}$, and so a different parametrization of the problem would lead in general to a different posterior p.d.f. In fact, one rarely chooses a flat prior as a true expression of degree of belief about a parameter; rather, it is used as a recipe to construct an interval, which in the end will have certain frequentist properties.

If a parameter is constrained to be non-negative, then the prior p.d.f. can simply be set to zero for negative values. An important example is the case of a Poisson variable n which counts signal events with unknown mean s as well as background with mean b, assumed known. For the signal mean s one often uses the prior

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \ . \tag{31.32}$$

As mentioned above, this is regarded as providing an interval whose frequentist properties can be studied, rather than as representing a degree of belief. In the absence of a clear discovery, (*e.g.*, if $n = 0$ or if in any case n is compatible with the expected background), one usually wishes to place an upper limit on s. Using the likelihood function for Poisson distributed n,

$$L(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{31.33}$$

along with the prior (31.32) in (31.30) gives the posterior density for s. An upper limit $s_{\rm up}$ at confidence level $1-\alpha$ can be obtained by requiring

$$1-\alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} L(n|s)\,\pi(s)\,ds}{\int_{-\infty}^{\infty} L(n|s)\,\pi(s)\,ds} \ , \tag{31.34}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (31.34) to incomplete gamma functions, the equation reduces to

$$\alpha = e^{-s_{\rm up}} \frac{\sum_{m=0}^{n} (s_{\rm up}+b)^m/m!}{\sum_{m=0}^{n} b^m/m!} \ . \tag{31.35}$$

This must be solved numerically for the limit $s_{\rm up}$. For the special case of $b = 0$, the sums can be related to the *quantile* $F^{-1}_{\chi^2}$ of the χ^2 distribution (inverse of the cumulative distribution) to give

$$s_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\alpha; n_{\rm d}) \ , \tag{31.36}$$

where the number of degrees of freedom is $n_{\rm d} = 2(n+1)$. The quantile of the χ^2 distribution can be obtained using the CERNLIB routine `CHISIN`. It so happens that for the case of $b = 0$, the upper limits from Eq. (31.36) coincide numerically with the values of the frequentist upper limits discussed in Section 31.4.2.4. Values for $1-\alpha = 0.9$ and 0.95 are given by the values $\nu_{\rm up}$ in Table 31.3. The frequentist properties of confidence intervals for the Poisson mean obtained in this way are discussed in Refs. [2] and [11].

Bayesian statistics provides a framework for incorporating systematic uncertainties into a result. Suppose, for example, that a model depends not only on parameters of interest $\boldsymbol{\theta}$ but on *nuisance parameters* $\boldsymbol{\nu}$, whose values are known with some limited accuracy. For a single nuisance parameter ν, for example, one might have a p.d.f. centered about its nominal value with a certain standard deviation σ_ν. Often a Gaussian p.d.f. provides a reasonable model for one's degree of belief about a nuisance parameter; in other cases more complicated shapes may be appropriate. The likelihood function, prior and posterior p.d.f.s then all depend on both $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ and are related by Bayes' theorem as usual. One can obtain the posterior p.d.f. for $\boldsymbol{\theta}$ alone by integrating over the nuisance parameters, *i.e.*,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu}|\boldsymbol{x})\, d\boldsymbol{\nu} \ . \tag{31.37}$$

If the prior joint p.d.f. for $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ factorizes, then integrating the posterior p.d.f. over $\boldsymbol{\nu}$ is equivalent to replacing the likelihood function by (see Ref. [12]),

$$L'(\boldsymbol{x}|\boldsymbol{\theta}) = \int L(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\nu})\pi(\boldsymbol{\nu})\,d\boldsymbol{\nu}\ . \tag{31.38}$$

The function $L'(\boldsymbol{x}|\boldsymbol{\theta})$ can also be used together with frequentist methods that employ the likelihood function such as ML estimation of parameters. The results then have a mixed frequentist/Bayesian character, where the systematic uncertainty due to limited knowledge of the nuisance parameters is built in. Although this may make it more difficult to disentangle statistical from systematic effects, such a hybrid approach may satisfy the objective of reporting the result in a convenient way.

Even if the subjective Bayesian approach is not used explicitly, Bayes' theorem represents the way that people evaluate the impact of a new result on their beliefs. One of the criteria in choosing a method for reporting a measurement, therefore, should be the ease and convenience with which the consumer of the result can carry out this exercise.

31.4.2. *Frequentist confidence intervals*:

The unqualified phrase "confidence intervals" refers to frequentist intervals obtained with a procedure due to Neyman [13], described below. These are intervals (or in the multiparameter case, regions) constructed so as to include the true value of the parameter with a probability greater than or equal to a specified level, called the *coverage probability.* In this section we discuss several techniques for producing intervals that have, at least approximately, this property.

31.4.2.1. *The Neyman construction for confidence intervals:*

Consider a p.d.f. $f(x;\theta)$ where x represents the outcome of the experiment and θ is the unknown parameter for which we want to construct a confidence interval. The variable x could (and often does) represent an estimator for θ. Using $f(x;\theta)$ we can find for a pre-specified probability $1-\alpha$ and for every value of θ a set of values $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ such that

$$P(x_1 < x < x_2;\theta) = 1-\alpha = \int_{x_1}^{x_2} f(x;\theta)\,dx\ . \tag{31.39}$$

This is illustrated in Fig. 31.3: a horizontal line segment $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ is drawn for representative values of θ. The union of such intervals for all values of θ, designated in the figure as $D(\alpha)$, is known as the *confidence belt.* Typically the curves $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ are monotonic functions of θ, which we assume for this discussion.

Upon performing an experiment to measure x and obtaining a value x_0, one draws a vertical line through x_0. The confidence interval for θ is the set of all values of θ for which the corresponding line segment $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ is intercepted by this vertical line. Such confidence intervals are said to have a *confidence level* (CL) equal to $1-\alpha$.

Now suppose that the true value of θ is θ_0, indicated in the figure. We see from the figure that θ_0 lies between $\theta_1(x)$ and $\theta_2(x)$ if and only if x lies between $x_1(\theta_0)$ and $x_2(\theta_0)$. The two events thus have the same probability, and since this is true for any value θ_0, we can drop the subscript 0 and obtain

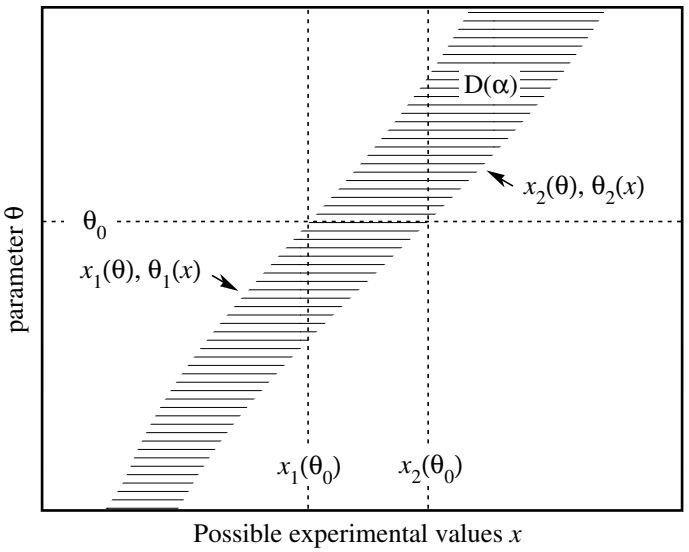

Figure 31.3: Construction of the confidence belt (see text).

$$1 - \alpha = P(x_1(\theta) < x < x_2(\theta)) = P(\theta_2(x) < \theta < \theta_1(x)) \; . \tag{31.40}$$

In this probability statement $\theta_1(x)$ and $\theta_2(x)$, *i.e.*, the endpoints of the interval, are the random variables and θ is an unknown constant. If the experiment were to be repeated a large number of times, the interval $[\theta_1, \theta_2]$ would vary, covering the fixed value θ in a fraction $1 - \alpha$ of the experiments.

The condition of coverage Eq. (31.39) does not determine x_1 and x_2 uniquely and additional criteria are needed. The most common criterion is to choose *central intervals* such that the probabilities excluded below x_1 and above x_2 are each $\alpha/2$. In other cases one may want to report only an upper or lower limit, in which case the probability excluded below x_1 or above x_2 can be set to zero. Another principle based on *likelihood ratio ordering* for determining which values of x should be included in the confidence belt is discussed in Sec. 31.4.2.2

When the observed random variable x is continuous, the coverage probability obtained with the Neyman construction is $1 - \alpha$, regardless of the true value of the parameter. If x is discrete, however, it is not possible to find segments $[x_1(\theta, \alpha), x_2(\theta, \alpha)]$ that satisfy (31.39) exactly for all values of θ. By convention one constructs the confidence belt requiring the probability $P(x_1 < x < x_2)$ to be *greater than or equal to* $1 - \alpha$. This gives confidence intervals that include the true parameter with a probability greater than or equal to $1 - \alpha$.

31.4.2.2. *Relationship between intervals and tests:*

An equivalent method of constructing confidence intervals is to consider a test (see Sec. 31.3) of the hypothesis that the parameter's true value is θ. One then excludes all values of θ where the hypothesis would be rejected at a significance level less than α. The remaining values constitute the confidence interval at confidence level $1-\alpha$.

In this procedure one is still free to choose the test to be used; this corresponds to the freedom in the Neyman construction as to which values of the data are included in the confidence belt. One possibility is use a test statistic based on the *likelihood ratio*,

$$\lambda = \frac{f(x;\theta)}{f(x;\widehat{\theta}\,)}\ , \tag{31.41}$$

where $\widehat{\theta}$ is the value of the parameter which, out of all allowed values, maximizes $f(x;\theta)$. This results in the intervals described in [14] by Feldman and Cousins. The same intervals can be obtained from the Neyman construction described in the previous section by including in the confidence belt those values of x which give the greatest values of λ.

Another technique that can be formulated in the language of statistical tests has been used to set limits on the Higgs mass from measurements at LEP [15]. For each value of the Higgs mass, a statistic called $\mathrm{CL_s}$ is determined from the ratio

$$\mathrm{CL_s} = \frac{p\text{-value of signal plus background hypothesis}}{p\text{-value of hypothesis of background only}}\ . \tag{31.42}$$

The p-values in (31.42) are themselves based on a goodness-of-fit statistic which depends in general on the signal being tested, *i.e.*, on the hypothesized Higgs mass. Smaller $\mathrm{CL_s}$ corresponds to a lesser level of agreement with the signal hypothesis.

In the usual procedure for constructing confidence intervals, one would exclude the signal hypothesis if the probability to obtain a value of $\mathrm{CL_s}$ less than the one actually observed is less than α. The LEP Higgs group has in fact followed a more conservative approach and excludes the signal at a confidence level $1-\alpha$ if $\mathrm{CL_s}$ itself (not the probability to obtain a lower $\mathrm{CL_s}$ value) is less than α. This results in a coverage probability that is in general greater than $1-\alpha$. The interpretation of such intervals is discussed in [15].

31.4.2.3. *Gaussian distributed measurements:*

An important example of constructing a confidence interval is when the data consists of a single random variable x that follows a Gaussian distribution; this is often the case when x represents an estimator for a parameter and one has a sufficiently large data sample. If there is more than one parameter being estimated, the multivariate Gaussian is used. For the univariate case with known σ,

$$1-\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\,dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{31.43}$$

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 for the interval $x \pm \delta$ to include μ. Fig. 31.4 shows a $\delta = 1.64\sigma$ confidence interval unshaded. The choice $\delta = \sigma$ gives an interval called the *standard error* which has $1 - \alpha = 68.27\%$ if σ is known. Values of α for other frequently used choices of δ are given in Table 31.1.

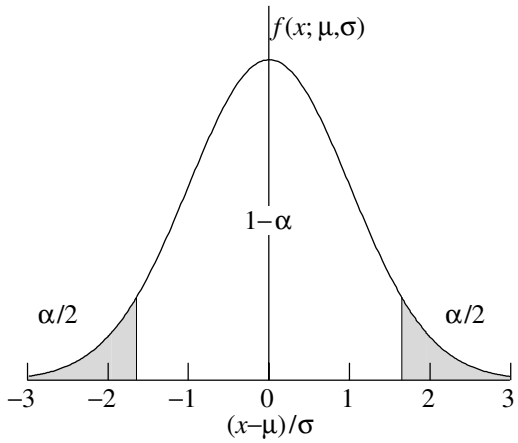

Figure 31.4: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by α, are as shown.

Table 31.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α (%)	δ	α (%)	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

We can set a one-sided (upper or lower) limit by excluding above $x + \delta$ (or below $x - \delta$). The values of α for such limits are half the values in Table 31.1.

In addition to Eq. (31.43), α and δ are also related by the cumulative distribution function for the χ^2 distribution,

$$\alpha = 1 - F(\chi^2; n) \,, \tag{31.44}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom. This can be obtained from Fig. 31.1 on the $n = 1$ curve or by using the CERNLIB routine `PROB`.

For multivariate measurements of, say, n parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, one requires the full covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$, which can be estimated as described in Sections 31.1.2 and 31.1.3. Under fairly general conditions with the methods of maximum-likelihood or least-squares in the large sample limit, the estimators will be distributed according to a multivariate Gaussian centered about the true (unknown) values $\boldsymbol{\theta}$, and furthermore the likelihood function itself takes on a Gaussian shape.

The standard error ellipse for the pair $(\widehat{\theta}_i, \widehat{\theta}_j)$ is shown in Fig. 31.5, corresponding to a contour $\chi^2 = \chi^2_{\mathrm{min}} + 1$ or $\ln L = \ln L_{\mathrm{max}} - 1/2$. The ellipse is centered about the estimated values $\widehat{\boldsymbol{\theta}}$, and the tangents to the ellipse give the standard deviations of the estimators, σ_i and σ_j. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2 - \sigma_j^2} \ , \tag{31.45}$$

where $\rho_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]/\sigma_i\sigma_j$ is the correlation coefficient.

The correlation coefficient can be visualized as the fraction of the distance σ_i from the ellipse's horizontal centerline at which the ellipse becomes tangent to vertical, *i.e.* at the distance $\rho_{ij}\sigma_i$ below the centerline as shown. As ρ_{ij} goes to $+1$ or -1, the ellipse thins to a diagonal line.

It could happen that one of the parameters, say, θ_j, is known from previous measurements to a precision much better than σ_j so that the current measurement contributes almost nothing to the knowledge of θ_j. However, the current measurement of of θ_i and its dependence on θ_j may still be important. In this case, instead of quoting both parameter estimates and their correlation, one sometimes reports the value of θ_i which minimizes χ^2 at a fixed value of θ_j, such as the PDG best value. This θ_i value lies along the dotted line between the points where the ellipse becomes tangent to vertical, and has statistical error σ_{inner} as shown on the figure. Instead of the correlation ρ_{ij}, one reports the dependency $d\widehat{\theta}_i/d\theta_j$ which is the slope of the dotted line. This slope is related to the correlation coefficient by $d\widehat{\theta}_i/d\theta_j = \rho_{ij} \times \frac{\sigma_i}{\sigma_j}$.

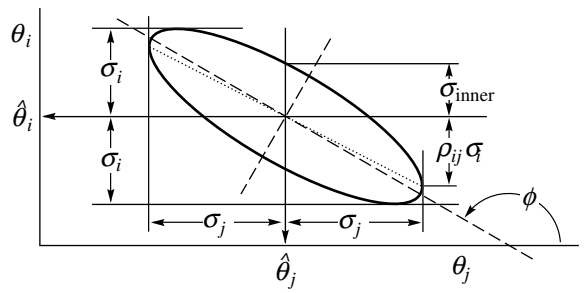

Figure 31.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

Table 31.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

As in the single-variable case, because of the symmetry of the Gaussian function between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, one finds that contours of constant $\ln L$ or χ^2 cover the true values with a certain, fixed probability. That is, the confidence region is determined by

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\max} - \Delta \ln L \ , \tag{31.46}$$

or where a χ^2 has been defined for use with the method of least squares,

$$\chi^2(\boldsymbol{\theta}) \leq \chi^2_{\min} + \Delta\chi^2 \ . \tag{31.47}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in Table 31.2 for several values of the coverage probability and number of fitted parameters.

For finite data samples, the probability for the regions determined by Equations (31.46) or (31.47) to cover the true value of $\boldsymbol{\theta}$ will depend on $\boldsymbol{\theta}$, so these are not exact confidence regions according to our previous definition. Nevertheless, they can still have a coverage probability only weakly dependent on the true parameter and approximately as given in Table 31.2. In any case the coverage probability of the intervals or regions obtained according to this procedure can in principle be determined as a function of the true parameter(s), for example, using a Monte Carlo calculation.

One of the practical advantages of intervals that can be constructed from the log-likelihood function or χ^2 is that it is relatively simple to produce the interval for the combination of several experiments. If N independent measurements result in log-likelihood functions $\ln L_i(\boldsymbol{\theta})$, then the combined log-likelihood function is simply the sum,

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln L_i(\boldsymbol{\theta}) \ . \tag{31.48}$$

This can then be used to determine an approximate confidence interval or region with Equation (31.46), just as with a single experiment.

31.4.2.4. *Poisson or binomial data:*

Another important class of measurements consists of counting a certain number of events n. In this section we will assume these are all events of the desired type, *i.e.*, there is no background. If n represents the number of events produced in a reaction with cross section σ, say, in a fixed integrated luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\nu = \sigma\mathcal{L}$. If, on the other hand, one has selected a larger sample of N events and found n of them to have a particular property, then n follows a binomial distribution where the parameter p gives the probability for the event to possess the property in question. This is appropriate, *e.g.*, for estimates of branching ratios or selection efficiencies based on a given total number of events.

For the case of Poisson distributed n, the upper and lower limits on the mean value ν can be found from the Neyman procedure to be

$$\nu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha_{\rm lo}; 2n) \,, \tag{31.49a}$$

$$\nu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{\rm up}; 2(n+1)) \,, \tag{31.49b}$$

where the upper and lower limits are at confidence levels of $1-\alpha_{\rm lo}$ and $1-\alpha_{\rm up}$, respectively, and $F^{-1}_{\chi^2}$ is the *quantile* of the χ^2 distribution (inverse of the cumulative distribution). The quantiles $F^{-1}_{\chi^2}$ can be obtained from standard tables or from the CERNLIB routine `CHISIN`. For central confidence intervals at confidence level $1-\alpha$, set $\alpha_{\rm lo} = \alpha_{\rm up} = \alpha/2$.

It happens that the upper limit from (31.49a) coincides numerically with the Bayesian upper limit for a Poisson parameter using a uniform prior p.d.f. for ν. Values for confidence levels of 90% and 95% are shown in Table 31.3.

For the case of binomially distributed n successes out of N trials with probability of success p, the upper and lower limits on p are found to be

$$p_{\rm lo} = \frac{nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]}{N-n+1 \,+\, nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]} \,, \tag{31.50a}$$

$$p_{\rm up} = \frac{(n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]}{(N-n) \,+\, (n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]} \,. \tag{31.50b}$$

Here F^{-1}_F is the quantile of the F distribution (also called the Fisher–Snedecor distribution; see Ref. [4]).

Table 31.3: Lower and upper limits for the mean ν of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	$\nu_{\rm lo}$	$\nu_{\rm up}$	$\nu_{\rm lo}$	$\nu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

31.4.2.5. *Difficulties with intervals near a boundary:*

A number of issues arise in the construction and interpretation of confidence intervals when the parameter can only take on values in a restricted range. An important example is where the mean of a Gaussian variable is constrained on physical grounds to be non-negative. This arises, for example, when the square of the neutrino mass is estimated from $\widehat{m}^2 = \widehat{E}^2 - \widehat{p}^2$, where $\widehat{E}$ and $\widehat{p}$ are independent, Gaussian distributed estimates of the energy and momentum. Although the true m^2 is constrained to be positive, random errors in $\widehat{E}$ and $\widehat{p}$ can easily lead to negative values for the estimate $\widehat{m}^2$.

If one uses the prescription given above for Gaussian distributed measurements, which says to construct the interval by taking the estimate plus or minus one standard deviation, then this can give intervals that are partially or entirely in the unphysical region. In fact, by following strictly the Neyman construction for the central confidence interval, one finds that the interval is truncated below zero; nevertheless an extremely small or even a zero-length interval can result.

An additional important example is where the experiment consists of counting a certain number of events, n, which is assumed to be Poisson distributed. Suppose the expectation value $E[n] = \nu$ is equal to $s + b$, where s and b are the means for signal and background processes, and assume further that b is a known constant. Then $\widehat{s} = n - b$ is an unbiased estimator for s. Depending on true magnitudes of s and b, the estimate $\widehat{s}$ can easily fall in the negative region. Similar to the Gaussian case with the positive mean, the central confidence interval or even the upper limit for s may be of zero length.

The confidence interval is in fact designed not to cover the parameter with a probability of at most α, and if a zero-length interval results, then this is evidently one of those experiments. So although the construction is behaving as it should, a null interval is an unsatisfying result to report and several solutions to this type of problem are possible.

An additional difficulty arises when a parameter estimate is not significantly far away from the boundary, in which case it is natural to report a one-sided confidence interval (often an upper limit). It is straightforward to force the Neyman prescription to produce only an upper limit by setting $x_2 = \infty$ in Eq. 31.39. Then x_1 is uniquely determined and the upper limit can be obtained. If, however, the data come out such that the parameter estimate is not so close to the boundary, one might wish to report a central (*i.e.*, two-sided) confidence interval. As pointed out by Feldman and Cousins [14], however, if the decision to report an upper limit or two-sided interval is made by looking at the data ("flip-flopping"), then the resulting intervals will not in general cover the parameter with the probability $1 - \alpha$.

With the confidence intervals suggested in [14], the prescription determines whether the interval is one- or two-sided in a way which preserves the coverage probability. Intervals with this property are said to be *unified*. Furthermore, the Feldman–Cousins prescription is such that null intervals do not occur. For a given choice of $1 - \alpha$, if the parameter estimate is sufficiently close to the boundary, then the method gives a one-sided limit. In the case of a Poisson variable in the presence of background, for example, this would occur if the number of observed events is compatible with the expected background. For parameter estimates increasingly far away from the boundary, i.e., for increasing signal significance, the interval makes a smooth transition from one- to two-sided, and far away from the boundary one obtains a central interval.

The intervals according to this method for the mean of Poisson variable in the absence of background are given in Table 31.4. (Note that α in [14] is defined following Neyman [13] as the coverage probability; this is opposite the modern convention used here in which the coverage probability is $1 - \alpha$.) The values of $1 - \alpha$ given here refer to the coverage of the true parameter by the whole interval $[\nu_1, \nu_2]$. In Table 31.3 for the one-sided upper and lower limits, however, $1 - \alpha$ refers to the probability to have individually $\nu_{\rm up} \geq \nu$ or $\nu_{\rm lo} \leq \nu$.

A potential difficulty with unified intervals arises if, for example, one constructs such an interval for a Poisson parameter s of some yet to be discovered signal process with, say, $1 - \alpha = 0.9$. If the true signal parameter is zero, or in any case much less than the expected background, one will usually obtain a one-sided upper limit on s. In a certain fraction of the experiments, however, a two-sided interval for s will result. Since, however, one typically chooses $1 - \alpha$ to be only 0.9 or 0.95 when searching for a new effect, the value $s = 0$ may be excluded from the interval before the existence of the effect is well established. It must then be communicated carefully that in excluding $s = 0$ from the interval, one is not necessarily claiming to have discovered the effect.

The intervals constructed according to the unified procedure in [14] for a Poisson variable n consisting of signal and background have the property that for $n = 0$ observed events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if $n = 0$ for the experiment in question, then no

Table 31.4: Unified confidence intervals $[\nu_1, \nu_2]$ for a the mean of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	ν_1	ν_2	ν_1	ν_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

background was observed, and therefore one may argue that the expected background should not be relevant. The extent to which one should regard this feature as a drawback is a subject of some controversy (see, *e.g.*, Ref. [17]).

Another possibility is to construct a Bayesian interval as described in Section 31.4.1. The presence of the boundary can be incorporated simply by setting the prior density to zero in the unphysical region. Priors based on invariance principles (rather than subjective degree of belief) for the Poisson mean are rarely used in high energy physics; they diverge for the case of zero events observed, and they give upper limits which undercover when evaluated by the frequentist definition of coverage [2]. Rather, priors uniform in the Poisson mean have been used, although as previsouly mentioned, this is generally not done to reflect the experimenter's degree of belief but rather as a procedure for obtaining an interval with certain frequentist properties. The resulting upper limits have a coverage probability that depends on the true value of the Poisson parameter and is everywhere greater than the stated probability content. Lower limits and two-sided intervals for the Poisson mean based on flat priors undercover, however, for some values of the parameter, although to an extent that in practical cases may not be too severe [2, 11]. Intervals constructed in this way have the advantage of being easy to derive; if several independent measurements are to be combined then one simply multiplies the likelihood functions (cf. Eq. (31.48)).

An additional alternative is presented by the intervals found from the likelihood function or χ^2 using the prescription of Equations (31.46) or (31.47). As in the case of

the Bayesian intervals, the coverage probability is not, in general, independent of the true parameter. Furthermore, these intervals can for some parameter values undercover. The coverage probability can of course be determined with some extra effort and reported with the result.

Also as in the Bayesian case, intervals derived from the value of the likelihood function from a combination of independent experiments can be determined simply by multiplying the likelihood functions. These intervals are also invariant under transformation of the parameter; this is not true for Bayesian intervals with a conventional flat prior, because a uniform distribution in, say, θ will not be uniform if transformed to θ^2. Use of the likelihood function to determine approximate confidence intervals is discussed further in [16].

In any case it is important always to report sufficient information so that the result can be combined with other measurements. Often this means giving an unbiased estimator and its standard deviation, even if the estimated value is in the unphysical region.

Regardless of the type of interval reported, the consumer of that result will almost certainly use it to derive some impression about the value of the parameter. This will inevitably be done, either explicitly or intuitively, with Bayes' theorem,

$$p(\theta|\text{result}) \propto L(\text{result}|\theta)\pi(\theta) \ , \tag{31.51}$$

where the reader supplies his or her own prior beliefs $\pi(\theta)$ about the parameter, and the 'result' is whatever sort of interval or other information the author has reported. For all of the intervals discussed, therefore, it is not sufficient to know the result; one must also know the probability to have obtained this result as a function of the parameter, *i.e.*, the likelihood. Contours of constant likelihood, for example, provide this information, and so an interval obtained from $\ln L = \ln L_{\max} - \Delta \ln L$ already takes one step in this direction.

It can also be useful with a frequentist interval to calculate its subjective probability content using the posterior p.d.f. based on one or several reasonable guesses for the prior p.d.f. If it turns out to be significantly less than the stated confidence level, this warns that it would be particularly misleading to draw conclusions about the parameter's value without further information from the likelihood.

References:

1. B. Efron, Am. Stat. **40**, 11 (1986).
2. R.D. Cousins, Am. J. Phys. **63**, 398 (1995).
3. A. Stuart, A.K. Ord, and Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2 *Classical Inference and Relationship* 6th Ed., (Oxford Univ. Press, 1998), and earlier editions by Kendall and Stuart. The likelihood-ratio ordering principle is described at the beginning of Ch. 23. Chapter 26 compares different schools of statistical inference.
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. H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press, New Jersey (1958).
6. L. Lyons, *Statistics for Nuclear and Particle Physicists* (Cambridge University Press, New York, 1986).
7. R. Barlow, Nucl. Inst. Meth. A **297**, 496 (1990).
8. G. Cowan, *Statistical Data Analysis* (Oxford University Press, Oxford, 1998).
9. For a review, see S. Baker and R. Cousins, Nucl. Instrum. Methods **221**, 437 (1984).
10. For information on neural networks and related topics, see *e.g.* C.M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford (1995); C. Peterson and T. Rögnvaldsson, An Introduction to Artificial Neural Networks, in *Proceedings of the 1991 CERN School of Computing*, C. Verkerk (ed.), CERN 92-02 (1992).
11. Byron P. Roe and Michael B. Woodroofe, Phys. Rev. **D63**, 13009 (2001).
12. Paul H. Garthwaite, Ian T. Jolliffe and Byron Jones, *Statistical Inference* (Prentice Hall, 1995).
13. 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).
14. G.J. Feldman and R.D. Cousins, Phys. Rev. **D57**, 3873 (1998). This paper does not specify what to do if the ordering principle gives equal rank to some values of x. Eq. 23.6 of Ref. 3 gives the rule: all such points are included in the acceptance region (the domain $D(\alpha)$). Some authors have assumed the contrary, and shown that one can then obtain null intervals.
15. A.L. Read, *Modified frequentist analysis of search results (the CL_s method)*, in F. James, L. Lyons and Y. Perrin (eds.), *Workshop on Confidence Limits*, CERN Yellow Report 2000-005, available through `weblib.cern.ch`.
16. F. Porter, Nucl. Inst. Meth. A **368**, 793 (1996).
17. Workshop on Confidence Limits, CERN, 17-18 Jan. 2000, `www.cern.ch/CERN/Divisions/EP/Events/CLW/`. The proceedings, F. James, L. Lyons, and Y. Perrin (eds.), CERN Yellow Report 2000-005, are available through `weblib.cern.ch`. See also the later Fermilab workshop linked to the CERN web page.