36. STATISTICS

Revised September 2011 by G. Cowan (RHUL).

This chapter gives an overview of statistical methods used in high-energy physics. In statistics, we are interested in using a given sample of data to make inferences about a probabilistic model, *e.g.*, to assess the model's validity or to determine the values of its parameters. 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. The most important tools in this framework are parameter estimation, covered in Section 36.1, and statistical tests, discussed in Section 36.2. Frequentist confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, are treated in Section 36.3.2. Note that in frequentist statistics one does not define a probability for a hypothesis or for a parameter.

Frequentist statistics provides the usual tools for reporting the outcome of an experiment objectively, without needing to incorporate prior beliefs concerning the parameter being measured or the theory being tested. As such, they are used for reporting most measurements and their statistical uncertainties in high-energy physics.

In Bayesian statistics, the interpretation of probability is more general and includes *degree of belief* (called subjective probability). One can then speak of a probability density function (p.d.f.) for a parameter, which expresses one's state of knowledge about where its true value lies. Bayesian methods allow for a natural way to input additional information, which in general may be subjective; in fact they *require* the *prior* p.d.f. as input for the parameters, *i.e.*, the degree of belief about the parameters' values before carrying out the measurement. Using Bayes' theorem Eq. (35.4), the prior degree of belief is updated by the data from the experiment. Bayesian methods for interval estimation are discussed in Sections 36.3.1 and 36.3.2.6

Bayesian techniques are often used to treat systematic uncertainties, where the author's 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. In some analyses, both the frequentist and Bayesian approaches are used together. One may, for example, treat systematic uncertainties in a model using Bayesian methods, but then construct a frequentist statistical test of that model.

For many inference problems, the frequentist and Bayesian approaches give similar 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.

Following common usage in physics, the word "error" is often used in this chapter to

mean "uncertainty." More specifically it can indicate the size of an interval as in "the standard error" or "error propagation," where the term refers to the standard deviation of an estimator.

36.1. Parameter estimation

Here we review *point estimation* of parameters, first with an overview of the frequentist approach and its two most important methods, maximum likelihood and least squares, treated in Sections 36.1.2 and 36.1.3. The Bayesian approach is outlined in Sec. 36.1.4.

An *estimator* $\widehat{\theta}$ (written with a hat) is a function of the data used to estimate the value of the parameter θ. Sometimes the word 'estimate' is used to denote the value of the estimator when evaluated with given data. 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 statistical or systematic uncertainties that are larger than the bias itself, so that the new $\widehat{\theta}'$ may not be better than the original.

(c) *Efficiency* is the ratio of the minimum possible variance for any estimator of θ to the variance $V[\widehat{\theta}]$ of the estimator actually used. Under rather general conditions, the minimum variance is given by the Rao-Cramér-Frechet bound,

$$\sigma^2_{\rm min} = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / I(\theta) \,, \tag{36.1}$$

where

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \sum_i \ln f(x_i;\theta)\right)^2\right] \tag{36.2}$$

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*,

$$\text{MSE} = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + b^2 \,, \tag{36.3}$$

is a measure of an estimator's quality which combines bias and variance.

(d) *Robustness* is the property of being insensitive to departures from assumptions in the p.d.f., *e.g.*, owing to uncertainties in the distribution's tails.

Simultaneously optimizing for all the measures of estimator quality described above can lead to conflicting requirements. For example, there is in general a trade-off between bias and variance. 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.

36.1.1. ***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

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{36.4}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 \tag{36.5}$$

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

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right) , \tag{36.6}$$

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 (36.4) is not necessarily the most efficient estimator; this is discussed further in Sec. 8.7 of Ref. 4.

If σ^2 is known, it does not improve the estimate $\widehat{\mu}$, as can be seen from Eq. (36.4); however, if μ is known, substitute it for $\widehat{\mu}$ in Eq. (36.5) and replace $N-1$ by N to obtain an estimator of σ^2 still with zero bias but smaller variance. If the x_i have different, known variances σ_i^2, then the weighted average

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{N} w_i x_i \tag{36.7}$$

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.

36.1.2. *The method of maximum likelihood* :

Suppose we have a set of N measured quantities $\boldsymbol{x} = (x_1, \ldots, x_N)$ described by a joint p.d.f. $f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is set of n parameters whose values are unknown. The *likelihood function* is given by the p.d.f. evaluated with the data $\boldsymbol{x}$, but viewed as a function of the parameters, *i.e.*, $L(\boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta})$. If the measurements x_i are statistically independent and each follow the p.d.f. $f(x; \boldsymbol{\theta})$, then the joint p.d.f. for $\boldsymbol{x}$ factorizes and the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i; \boldsymbol{\theta}) \; . \tag{36.8}$$

The method of maximum likelihood takes the estimators $\widehat{\boldsymbol{\theta}}$ to be those values of $\boldsymbol{\theta}$ that maximize $L(\boldsymbol{\theta})$.

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 the posterior p.d.f. for $\boldsymbol{\theta}$ from the likelihood, but this requires multiplying by a prior p.d.f. (see Sec. 36.3.1).

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*,

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \; , \qquad i = 1, \ldots, n \; . \tag{36.9}$$

Often the solution must be found numerically. 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

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\theta}} \; . \tag{36.10}$$

For finite samples, however, Eq. (36.10) 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

$$\ln L(\boldsymbol{\theta}') = \ln L_{\max} - s^2/2 \,, \tag{36.11}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (36.58)). The extreme limits of this contour on the θ_i axis give an approximate s-standard-deviation confidence interval for θ_i (see Section 36.3.2.4).

In the case where the size n of the data sample $x_1, \ldots, x_n$ is small, the unbinned maximum likelihood method, *i.e.*, use of equation (36.8), is preferred since binning can only result in a loss of information, and hence larger statistical errors for the parameter estimates. The sample size n can be regarded as fixed, or the analyst can choose to treat it as a Poisson-distributed variable; this latter option is sometimes called "extended maximum likelihood" (see, *e.g.*, Refs. [6–8]) .

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 equivalent quantity $-2 \ln \lambda(\boldsymbol{\theta})$. For independent Poisson distributed n_i this is [9]

$$-2 \ln \lambda(\boldsymbol{\theta}) = 2 \sum_{i=1}^{N} \left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln \frac{n_i}{\nu_i(\boldsymbol{\theta})} \right] \,, \tag{36.12}$$

where for bins with $n_i = 0$, the last term in (36.12) is zero. The expression (36.12) without the terms $\nu_i - n_i$ also gives $-2 \ln \lambda(\boldsymbol{\theta})$ for multinomially distributed n_i, *i.e.*, when the total number of entries is regarded as fixed. In the limit of zero bin width, maximizing (36.12) is equivalent to maximizing the unbinned likelihood function (36.8).

A benefit of binning is that it allows for a goodness-of-fit test (see Sec. 36.2.2). Assuming the model is correct, then according to Wilks' theorem, for sufficiently large ν_i and providing certain regularity conditions are met, the minimum of $-2 \ln \lambda$ as defined by Eq. (36.12) follows a χ^2 distribution (see, *e.g.*, Ref. 3). If there are N bins and m fitted parameters, then the number of degrees of freedom for the χ^2 distribution is $N - m$ if the data are treated as Poisson-distributed, and $N - m - 1$ if the n_i are multinomially distributed.

Suppose the n_i are Poisson-distributed and the overall normalization $\nu_{\text{tot}} = \sum_i \nu_i$ is taken as an adjustable parameter, so that $\nu_i = \nu_{\text{tot}} p_i(\boldsymbol{\theta})$, where the probability to be in the ith bin, $p_i(\boldsymbol{\theta})$, does not depend on ν_{tot}. Then by minimizing Eq. (36.12), 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).

36.1.3. *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

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - F(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} \ . \tag{36.13}$$

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

The minimum of Equation (36.13) 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 $V_{ij} = \text{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{y} - \boldsymbol{F}(\boldsymbol{\theta})) \ , \tag{36.14}$$

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 (36.14)), and the superscript T denotes the 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.*,

$$F(x_i;\boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) \ . \tag{36.15}$$

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,

$$\widehat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y} \ . \tag{36.16}$$

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

$$U = DVD^T = (H^T V^{-1} H)^{-1} \ , \tag{36.17}$$

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

$$(U^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\bigg|_{\theta=\widehat{\theta}} = \sum_{k,l=1}^{N} h_i(x_k)(V^{-1})_{kl} h_j(x_l) \ . \tag{36.18}$$

The LS estimators can also be found from the expression

$$\widehat{\boldsymbol{\theta}} = U\boldsymbol{g} \ , \tag{36.19}$$

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

$$g_i = \sum_{j,k=1}^{N} y_j h_i(x_k)(V^{-1})_{jk} \ . \tag{36.20}$$

For the case of uncorrelated y_i, for example, one can use (36.19) with

$$(U^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k)h_j(x_k)}{\sigma_k^2} \ , \tag{36.21}$$

$$g_i = \sum_{k=1}^{N} \frac{y_k h_i(x_k)}{\sigma_k^2} \ . \tag{36.22}$$

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

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\widehat{\boldsymbol{\theta}}) + 1 = \chi^2_{\rm min} + 1 \tag{36.23}$$

has tangent planes located at approximately 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, *i.e.*, one should determine the area under the fitted curve directly from the number of entries in the histogram (see Ref. 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. (36.12)

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 36.2.2.

36.1.4. *The Bayesian approach* :

In the frequentist methods discussed above, probability is associated only with data, not with the value of a parameter. This is no longer the case in Bayesian statistics, however, which we introduce in this section. Bayesian methods are considered further in Sec. 36.3.1 for interval estimation and in Sec. 36.2.3 for model selection. For general introductions to Bayesian statistics see, *e.g.*, Refs. [20–23].

Suppose the outcome of an 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. $p(\boldsymbol{\theta}|\boldsymbol{x})$, whose integral over any given region gives the degree of belief for $\boldsymbol{\theta}$ to take on values in that region, given the data $\boldsymbol{x}$. It is obtained by using Bayes' theorem,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'} \,, \tag{36.24}$$

where $L(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function, *i.e.*, the joint p.d.f. for the data viewed as a function 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 Eq. (36.24) serves to normalize the posterior p.d.f. to unity.

As it can be difficult to report the full posterior p.d.f. $p(\boldsymbol{\theta}|\boldsymbol{x})$, one would usually summarize it with statistics such as the mean (or median), and covariance matrix. In addition one may construct intervals with a given probability content, as is discussed in Sec. 36.3.1 on Bayesian interval estimation.

36.1.4.1. *Priors:*

Bayesian statistics supplies no unique rule for determining the prior $\pi(\boldsymbol{\theta})$; this reflects the experimenter's subjective degree of belief (or state of knowledge) about $\boldsymbol{\theta}$ before the measurement was carried out. For the result to be of value to the broader community, whose members may not share these beliefs, it is important to carry out a sensitivity analysis, that is, to show how the result changes under a reasonable variation of the prior probabilities.

One might like 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 non-equivalent posterior p.d.f.

For the special case of a constant prior, one can see from Bayes' theorem (36.24) that the posterior is proportional to the likelihood, and therefore the mode (peak position) of the posterior is equal to the ML estimator. The posterior mode, however, will change in general upon a transformation of parameter. A summary statistic other than the mode may be used as the Bayesian estimator, such as the median, which is invariant under parameter transformation. But this will not in general coincide with the ML estimator.

The difficult and subjective nature of encoding personal knowledge into priors has led to what is called *objective Bayesian statistics*, where prior probabilities are based not on an actual degree of belief but rather derived from formal rules. These give, for example,

priors which are invariant under a transformation of parameters or which result in a maximum gain in information for a given set of measurements. For an extensive review see, *e.g.*, Ref. 24.

Objective priors do not in general reflect degree of belief, but they could in some cases be taken as possible, although perhaps extreme, subjective priors. The posterior probabilities as well therefore do not necessarily reflect a degree of belief. However one may regard investigating a variety of objective priors to be an important part of the sensitivity analysis. Furthermore, use of objective priors with Bayes' theorem can be viewed as a recipe for producing estimators or intervals which have desirable frequentist properties.

An important procedure for deriving objective priors is due to Jeffreys. According to *Jeffreys' rule* one takes the prior as

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}\,, \tag{36.25}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln L(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\, L(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \tag{36.26}$$

is the *Fisher information matrix*. One can show that the Jeffreys prior leads to inference that is invariant under a transformation of parameters. One should note that the Jeffreys prior depends on the likelihood function, and thus contains information about the measurement model itself, which goes beyond one's degree of belief about the value of a parameter. As examples, the Jeffreys prior for the mean μ of a Gaussian distribution is a constant, and for the mean of a Poisson distribution one finds $\pi(\mu) \propto 1/\sqrt{\mu}$.

Neither the constant nor $1/\sqrt{\mu}$ priors can be normalized to unit area and are said to be *improper*. This can be allowed because the prior always appears multiplied by the likelihood function, and if the likelihood falls off sufficiently quickly then one may have a normalizable posterior density.

An important type of objective prior is the reference prior due to Bernardo and Berger [25]. To find the reference prior for a given problem one considers the Kullback-Leibler divergence $D_n[\pi, p]$ of the posterior $p(\boldsymbol{\theta}|\boldsymbol{x})$ relative to a prior $\pi(\boldsymbol{\theta})$, obtained from a set of data $\boldsymbol{x} = (x_1, \ldots, x_n)$, which are assumed to consist of n independent and identically distributed values of x:

$$D_n[\pi, p] = \int p(\boldsymbol{\theta}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\,. \tag{36.27}$$

This is effectively a measure of the gain in information provided by the data. The reference prior is chosen so that the expectation value of this information gain is maximized for the limiting case of $n \to \infty$, where the expectation is computed with respect to the marginal distribution of the data,

$$p(\boldsymbol{x}) = \int L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\,. \tag{36.28}$$

For a single, continuous parameter the reference prior is usually identical to Jeffreys' prior. In the multiparameter case an iterative algorithm exists, which requires sorting the parameters by order of inferential importance. Often the result does not depend on this order, but when it does, this can be part of a robustness analysis. Further discussion and applications to particle physics problems can be found in Ref. 26.

36.1.4.2. *Bayesian treatment of nuisance parameters:*

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. If, for example, the parameter represents a non-negative quantity then a log-normal or gamma p.d.f. can be a more natural choice than a Gaussian truncated at zero. 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.*,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu}|\boldsymbol{x})\, d\boldsymbol{\nu} \; . \tag{36.29}$$

Such integrals can often not be carried out in closed form, and if the number of nuisance parameters is large, then they can be difficult to compute with standard Monte Carlo methods. *Markov Chain Monte Carlo* (MCMC) is often used for computing integrals of this type (see Sec. 37.5).

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 the *marginal likelihood* (see Ref. 27),

$$L_{\rm m}(\boldsymbol{x}|\boldsymbol{\theta}) = \int L(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\nu})\pi(\boldsymbol{\nu})\, d\boldsymbol{\nu} \; . \tag{36.30}$$

The marginal likelihood 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 (see Ref. 28). 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. The marginal likelihood may be compared with the profile likelihood, which is discussed in Sec. 36.3.2.3.

36.1.5. ***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 estimated $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, 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

$$U_{ij} \approx \sum_{k,l} \left.\frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l}\right|_{\widehat{\theta}} V_{kl} \; . \tag{36.31}$$

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

$$A_{ij} = \left.\frac{\partial \eta_i}{\partial \theta_j}\right|_{\widehat{\theta}} \; , \tag{36.32}$$

and A^T is its transpose. The approximation is exact if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is linear (it holds, for example, in equation (36.17)). 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 $\widehat{\boldsymbol{\theta}}$.

36.2. Statistical tests

In addition to estimating parameters, one often wants to assess the validity of certain statements concerning the data's underlying distribution. Frequentist *hypothesis tests*, described in Sec. 36.2.1, provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. In *significance tests*, covered in Sec. 36.2.2, 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. In the Bayesian approach, the corresponding procedure is based fundamentally on the posterior probabilities of the competing hypotheses. In Sec. 36.2.3 we describe a related construct called the Bayes factor, which can be used to quantify the degree to which the data prefer one or another hypothesis.

36.2.1. ***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, w, such that there is no more than a specified

probability under H_0, α, called the *size* or *significance level* of the test, to find $\boldsymbol{x} \in w$. If the data are discrete, it may not possible to find a critical region with exact probability content α, and thus we require $P(\boldsymbol{x} \in w|H_0) \leq \alpha$. If the data are observed in the critical region, H_0 is rejected.

There are in general a large (often infinite) number of regions of the data space that have probability content α and thus qualify as possible critical regions. To choose one of them one should take into account the probabilities for the data predicted by some alternative hypothesis (or set of alternatives) H_1. Rejecting H_0 if it is true is called a *type-I error*, and occurs by construction with probability no greater than α. Not rejecting H_0 if an alternative H_1 is true is called a *type-II error*, and for a given test this will have a certain probability β. The quantity $1 - \beta$ is called the *power* of the test of H_0 with respect to the alternative H_1. A strategy for defining the critical region can therefore be to maximize the power with respect to some alternative (or alternatives) given a fixed size α.

In high-energy physics, the components of $\boldsymbol{x}$ might represent the measured properties of candidate events, and the critical region is defined by the cuts that one imposes in order to reject background and thus accept events likely to be of a certain desired type. Here H_0 could represent the background hypothesis and the alternative H_1 could represent the sought after signal. In other cases, H_0 could be the hypothesis that an entire event sample consists of background events only, and the alternative H_1 may represent the hypothesis of a mixture of background and signal.

Often rather than using the full set of quantities $\boldsymbol{x}$, 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.

To maximize the power of a test of H_0 with respect to the alternative H_1, the *Neyman–Pearson lemma* states that the critical region w should be chosen such that for all data values $\boldsymbol{x}$ inside w, the ratio

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \,, \tag{36.33}$$

is greater than a given constant, the value of which is determined by the size of the test α. Here H_0 and H_1 must be simple hypotheses, i.e., they should not contain undetermined parameters.

The lemma is equivalent to the statement that (36.33) represents the optimal test statistic where the critical region is defined by a single cut on λ. This test will lead to the maximum power (e.g., probability to reject the background hypothesis if the signal hypothesis is true) for a given probability α to reject the background hypothesis if it is in fact true. 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 $f(\boldsymbol{x}|H_1)$.

In the usual case where the likelihood ratio (36.33) cannot be used explicitly, there exist a variety of other multivariate classifiers that effectively separate different types of events. Methods often used in HEP include *neural networks* or *Fisher discriminants*

(see Ref. 10). Recently, further classification methods from machine-learning have been applied in HEP analyses; these include *probability density estimation (PDE)* techniques, *kernel-based PDE* (*KDE* or *Parzen window*), *support vector machines*, and *decision trees*. Techniques such as "boosting" and "bagging" can be applied to combine a number of classifiers into a stronger one with greater stability with respect to fluctuations in the training data. Descriptions of these methods can be found in [11–13], and *Proceedings of the PHYSTAT* conference series [14]. Software for HEP includes the `TMVA` [15] and `StatPatternRecognition` [16] packages.

36.2.2. *Significance 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 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 analyst 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 significance of a discrepancy between the data and what one expects under the assumption of H_0 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

$$p = \int_{t_{\rm obs}}^{\infty} g(t|H_0)\, dt \ , \tag{36.34}$$

where $t_{\rm obs}$ is the value of the statistic obtained in the actual experiment.

The p-value should not be confused with the size (significance level) of a test, or the confidence level of a confidence interval (Section 36.3), both of which are pre-specified constants. We may formulate a hypothesis test, however, by defining the critical region to correspond to the data outcomes that give the lowest p-values, so that finding $p < \alpha$ implies that the data outcome was in the critical region. When constructing a p-value, one generally takes the region of data space deemed to have lower compatibility with the model being tested to have higher compatibility with a given alternative, and thus the corresponding test will have a high power with respect to this alternative.

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 searching for a new phenomenon, one tries to reject the hypothesis H_0 that the data are consistent with known, *e.g.*, Standard Model processes. If the p-value of H_0 is sufficiently low, then one is willing to accept that some alternative hypothesis is

true. Often one converts the p-value into an equivalent significance Z, defined so that a Z standard deviation upward fluctuation of a Gaussian random variable would have an upper tail area equal to p, *i.e.*,

$$Z = \Phi^{-1}(1 - p) \, . \tag{36.35}$$

Here Φ is the cumulative distribution of the Standard Gaussian, and Φ^{-1} is its inverse (quantile) function. Often in HEP, the level of significance where an effect is said to qualify as a discovery is $Z = 5$, *i.e.*, a 5σ effect, corresponding to a p-value of 2.87×10^{-7}. One's actual degree of belief that a new process is present, however, will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data, one's confidence in the model that led to the observed p-value, and possible corrections for multiple observations out of which one focuses on the smallest p-value obtained (the "look-elsewhere effect"). For a review of how to incorporate systematic uncertainties into p-values see, *e.g.*, Ref. 17; a computationally fast method that provides an approximate correction for the look-elsewhere effect is described in Ref. 18.

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 (36.13). This statistic can be used to test the *goodness-of-fit*, *i.e.*, the test provides a measure of the significance of a discrepancy between the data and the hypothesized functional form used in the fit. 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 (36.13) becomes *Pearson's* χ^2 *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \, . \tag{36.36}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct, and if the expected values ν_i in (36.36) are sufficiently large (or equivalently, if the measurements n_i can be treated as following a Gaussian distribution), 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.

Alternatively, one may fit parameters and evaluate goodness-of-fit by minimizing $-2 \ln \lambda$ from Eq. (36.12). 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 Ref. 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

$$p = \int_{\chi^2}^{\infty} f(z; n_{\rm d}) \, dz \, , \tag{36.37}$$

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. 36.1 or from the ROOT function `TMath::Prob`. If the

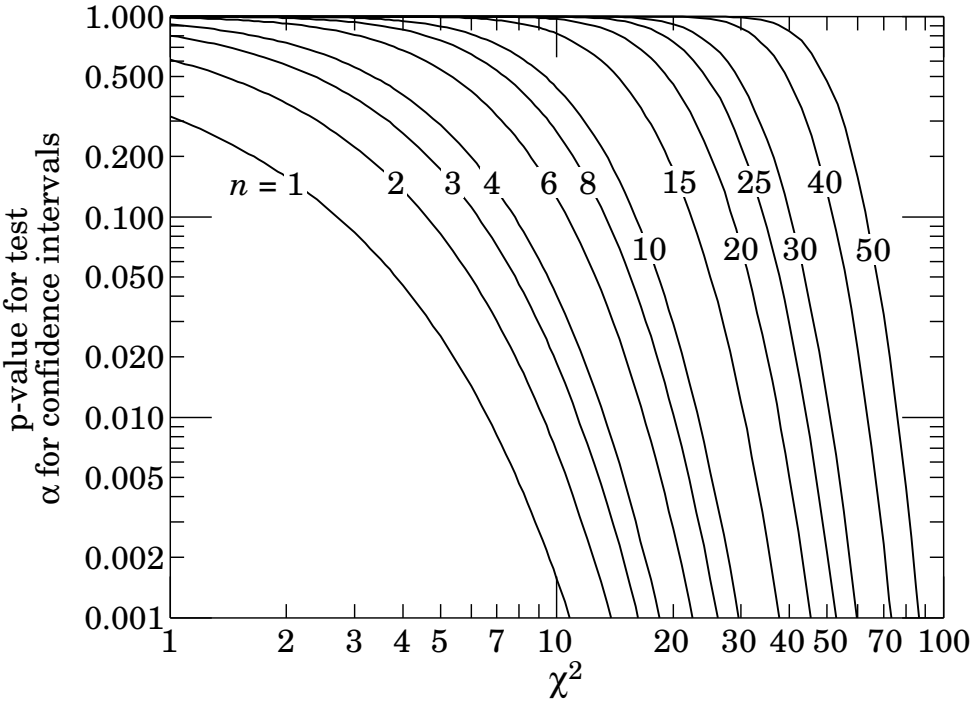

Figure 36.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 36.3.2.4).

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_d, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_\mathrm{d}$. Hence the quantity χ^2/n_d is sometimes reported. Since the p.d.f. of χ^2/n_d depends on n_d, however, one must report n_d as well if one wishes to determine the p-value. The p-values obtained for different values of χ^2/n_d are shown in Fig. 36.2.

If one finds a χ^2 value much greater than n_d, and a correspondingly small p-value, one may be tempted to expect a high degree of uncertainty for any fitted parameters. Poor goodness-of-fit, however, does not mean that one will have large statistical errors for parameter estimates. If, for example, the error bars (or covariance matrix) used in constructing the χ^2 are underestimated, then this will lead to underestimated statistical errors for the fitted parameters. The standard deviations of estimators that one finds from, say, Eq. (36.11) reflect how widely the estimates would be distributed if one were to repeat the measurement many times, assuming that the hypothesis and measurement errors used in the χ^2 are also correct. They do not include the systematic error which may result from an incorrect hypothesis or incorrectly estimated measurement errors in the χ^2.

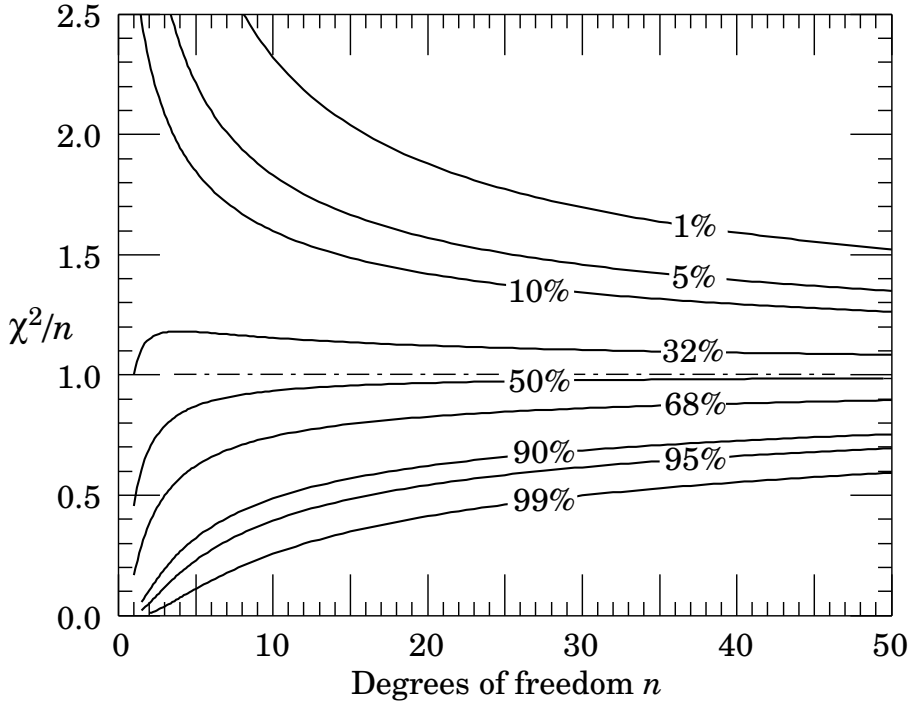

Figure 36.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

36.2.3. ***Bayesian model selection*** :

In Bayesian statistics, all of one's knowledge about a model is contained in its posterior probability, which one obtains using Bayes' theorem (36.24). Thus one could reject a hypothesis H if its posterior probability $P(H|\boldsymbol{x})$ is sufficiently small. The difficulty here is that $P(H|\boldsymbol{x})$ is proportional to the prior probability $P(H)$, and there will not be a consensus about the prior probabilities for the existence of new phenomena. Nevertheless one can construct a quantity called the Bayes factor (described below), which can be used to quantify the degree to which the data prefer one hypothesis over another, and is independent of their prior probabilities.

Consider two models (hypotheses), H_i and H_j, described by vectors of parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, respectively. Some of the components will be common to both models and others may be distinct. The full prior probability for each model can be written in the form

$$\pi(H_i, \boldsymbol{\theta}_i) = P(H_i)\pi(\boldsymbol{\theta}_i|H_i) \; , \tag{36.38}$$

Here $P(H_i)$ is the overall prior probability for H_i, and $\pi(\boldsymbol{\theta}_i|H_i)$ is the normalized p.d.f. of its parameters. For each model, the posterior probability is found using Bayes' theorem,

$$P(H_i|\boldsymbol{x}) = \frac{\int L(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)P(H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{P(\boldsymbol{x})} \; , \tag{36.39}$$

where the integration is carried out over the internal parameters $\boldsymbol{\theta}_i$ of the model. The

ratio of posterior probabilities for the models is therefore

$$\frac{P(H_i|\boldsymbol{x})}{P(H_j|\boldsymbol{x})} = \frac{\int L(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int L(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \frac{P(H_i)}{P(H_j)} \,. \tag{36.40}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int L(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int L(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \,. \tag{36.41}$$

This gives what the ratio of posterior probabilities for models i and j would be if the overall prior probabilities for the two models were equal. If the models have no nuisance parameters *i.e.*, no internal parameters described by priors, then the Bayes factor is simply the likelihood ratio. The Bayes factor therefore shows by how much the probability ratio of model i to model j changes in the light of the data, and thus can be viewed as a numerical measure of evidence supplied by the data in favour of one hypothesis over the other.

Although the Bayes factor is by construction independent of the overall prior probabilities $P(H_i)$ and $P(H_j)$, it does require priors for all internal parameters of a model, *i.e.*, one needs the functions $\pi(\boldsymbol{\theta}_i|H_i)$ and $\pi(\boldsymbol{\theta}_j|H_j)$. In a Bayesian analysis where one is only interested in the posterior p.d.f. of a parameter, it may be acceptable to take an unnormalizable function for the prior (an improper prior) as long as the product of likelihood and prior can be normalized. But improper priors are only defined up to an arbitrary multiplicative constant, and so the Bayes factor would depend on this constant. Furthermore, although the range of a constant normalized prior is unimportant for parameter determination (provided it is wider than the likelihood), this is not so for the Bayes factor when such a prior is used for only one of the hypotheses. So to compute a Bayes factor, all internal parameters must be described by normalized priors that represent meaningful probabilities over the entire range where they are defined.

An exception to this rule may be considered when the identical parameter appears in the models for both numerator and denominator of the Bayes factor. In this case one can argue that the arbitrary constants would cancel. One must exercise some caution, however, as parameters with the same name and physical meaning may still play different roles in the two models.

Both integrals in equation (36.41) are of the form

$$m = \int L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \,, \tag{36.42}$$

which is called the *marginal likelihood* (or in some fields called the *evidence*). A review of Bayes factors including a discussion of computational issues can be found in Ref. 30.

36.3. 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 stated 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 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 a judiciously chosen subjective Bayesian prior.

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.

36.3.1. *Bayesian intervals* :

As described in Sec. 36.1.4, a Bayesian posterior probability may be used to determine regions that will have a given probability of containing the true value of a parameter. 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 posterior probability, *i.e.*,

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta \ . \tag{36.43}$$

Sometimes an upper or lower limit is desired, *i.e.*, $\theta_{\rm lo}$ or $\theta_{\rm up}$ can be set to a physical boundary or to plus or minus 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.

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

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \quad . \tag{36.44}$$

This prior 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 (see, however, Sec. 36.3.2.6 on "flip-flopping" concerning frequentist coverage). Using the likelihood function for Poisson distributed n,

$$L(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{36.45}$$

along with the prior (36.44) in (36.24) gives the posterior density for s. An upper limit $s_{\rm up}$ at confidence level (or here, rather, credibility level) $1 - \alpha$ can be obtained by requiring

$$1 - \alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} L(n|s)\, \pi(s)\, ds}{\int_{-\infty}^{\infty} L(n|s)\, \pi(s)\, ds} \ , \tag{36.46}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (36.46) to incomplete gamma functions, the equation reduces to

$$\alpha = e^{-s_{\rm up}} \frac{\sum_{m=0}^{n} (s_{\rm up} + b)^m / m!}{\sum_{m=0}^{n} b^m / m!} \ . \tag{36.47}$$

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

$$s_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; n_{\rm d}) \ , \tag{36.48}$$

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 ROOT function `TMath::ChisquareQuantile`. It so happens that for the case of $b = 0$, the upper limits from Eq. (36.48) coincide numerically with the values of the frequentist upper limits discussed in Section 36.3.2.5. Values for $1 - \alpha = 0.9$ and 0.95 are given by the values $\nu_{\rm up}$ in Table 36.3. The frequentist properties of confidence intervals for the Poisson mean in this way are discussed in Refs. [2] and [19].

As in any Bayesian analysis, it is important to show how the result would change if one uses different prior probabilities. For example, one could consider the Jeffreys prior as described in Sec. 36.1.4. For this problem one finds the Jeffreys prior $\pi(s) \propto 1/\sqrt{s+b}$ for $s \geq 0$ and zero otherwise. As with the constant prior, one would not regard this as representing one's prior beliefs about s, both because it is improper and also as it depends on b. Rather it is used with Bayes' theorem to produce an interval whose frequentist properties can be studied.

36.3.2. *Frequentist confidence intervals* :

The unqualified phrase "confidence intervals" refers to frequentist intervals obtained with a procedure due to Neyman [29], 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.

36.3.2.1. *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

$$P(x_1 < x < x_2;\theta) = 1-\alpha = \int_{x_1}^{x_2} f(x;\theta)\,dx\ . \tag{36.49}$$

This is illustrated in Fig. 36.3: 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.

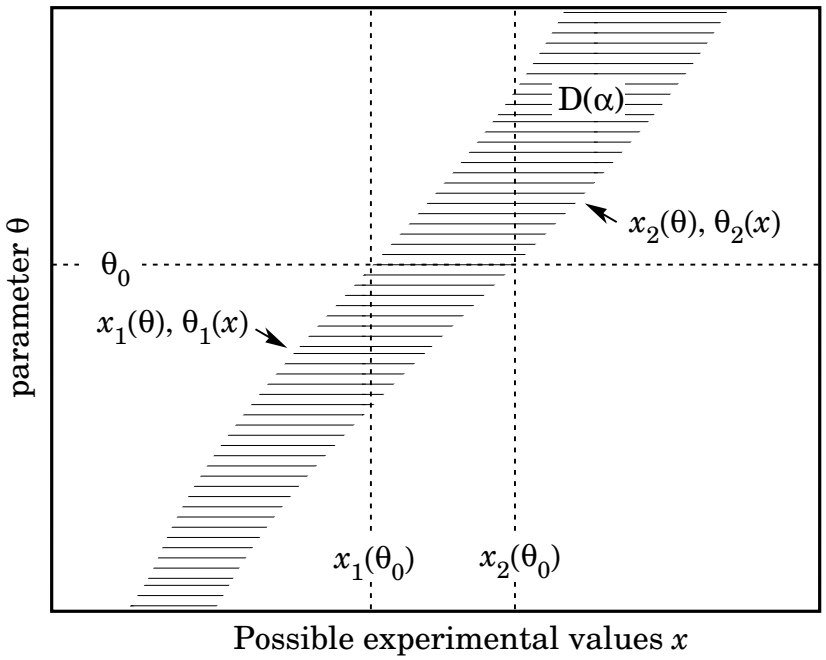

Figure 36.3: Construction of the confidence belt (see text).

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

$$1-\alpha = P(x_1(\theta) < x < x_2(\theta)) = P(\theta_2(x) < \theta < \theta_1(x)) \; . \tag{36.50}$$

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 in Eq. (36.49) does not determine x_1 and x_2 uniquely, and additional criteria are needed. One possibility 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 below.

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 Eq. (36.49) 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 $1-\alpha$.

An equivalent method of constructing confidence intervals is to consider a test (see Sec. 36.2) of the hypothesis that the parameter's true value is θ (assume one constructs a test for all physical values of 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 to use a test statistic based on the *likelihood ratio*,

$$\lambda = \frac{f(x;\theta)}{f(x;\widehat{\theta}\,)} \; , \tag{36.51}$$

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 Ref. 31 by Feldman and Cousins. The same intervals can be obtained from the Neyman construction described above by including in the confidence belt those values of x which give the greatest values of λ.

36.3.2.2. *Parameter exclusion in cases of low sensitivity:*

An important example of a statistical test arises in the search for a new signal process. Suppose the parameter μ is defined such that it is proportional to the signal cross section. A statistical test may be carried out for hypothesized values of μ, which may be done by computing a p-value, p_μ, for each hypothesized μ. Those values not rejected in a test of size α, *i.e.*, for which one does not find $p_\mu < \alpha$, constitute a confidence interval with confidence level $1 - \alpha$.

In general one will find that for some regions in the parameter space of the signal model, the predictions for data are almost indistinguishable from those of the background-only model. This corresponds to the case where μ is very small, as would occur, *e.g.*, if one searches for a Higgs boson with a mass so high that its production rate in a given experiment is negligible. That is, one has essentially no experimental sensitivity to such a model.

One would prefer that if the sensitivity to a model (or a point in a model's parameter space) is very low, then it should not be excluded. Even if the outcomes predicted with or without signal are identical, however, the probability to reject the signal model will equal α, the type-I error rate. As one often takes α to be 5%, this would mean that in a large number of searches covering a broad range of a signal model's parameter space, there would inevitably be excluded regions in which the experimental sensitivity is very small, and thus one may question whether it is justified to regard such parameter values as disfavored.

Exclusion of models to which one has little or no sensitivity occurs, for example, if the data fluctuate very low relative to the expectation of the background-only hypothesis. In this case the resulting upper limit on the predicted rate (cross section) of a signal model may be anomalously low. As a means of controlling this effect one often determines the mean or median limit under assumption of the background-only hypothesis using a simplified Monte Carlo simulation of the experiment. An upper limit found significantly below the background-only expectation may indicate a strong downward fluctuation of the data, or perhaps as well an incorrect estimate of the background rate.

The $\mathrm{CL_s}$ method aims to mitigate the problem of excluding models to which one is not sensitive by effectively penalizing the p-value of a tested parameter by an amount that increases with decreasing sensitivity [32,33]. The procedure is based on a statistic called $\mathrm{CL_s}$, which is defined as

$$\mathrm{CL_s} = \frac{p_\mu}{1 - p_0} \,, \tag{36.52}$$

where p_0 is the p-value of the background-only hypothesis. In the usual formulation of the method, the p-values for μ and 0 are defined using a single test statistic, and the definition of $\mathrm{CL_s}$ above assumes this statistic is continuous; more details can be found in Refs. [32,33].

A point in a model's parameter space is regarded as excluded if one finds $\mathrm{CL_s} < \alpha$. As the denominator in Eq. (36.52) is always less than or equal to unity, the exclusion criterion based on $\mathrm{CL_s}$ is more stringent than the usual requirement $p_\mu < \alpha$. In this sense the $\mathrm{CL_s}$ procedure is conservative, and the coverage probability of the corresponding intervals will

exceed the nominal confidence level $1-\alpha$. If the experimental sensitivity to a given value of μ is very low, then one finds that as p_μ decreases, so does the denominator $1-p_0$, and thus the condition $\mathrm{CL_s} < \alpha$ is effectively prevented from being satisfied. In this way the exclusion of parameters in the case of low sensitivity is suppressed.

The $\mathrm{CL_s}$ procedure has the attractive feature that the resulting intervals coincide with those obtained from the Bayesian method in two important cases: the mean value of a Poisson or Gaussian distributed measurement with a constant prior. The $\mathrm{CL_s}$ intervals overcover for all values of the parameter μ, however, by an amount that depends on μ.

The problem of excluding parameter values to which one has little sensitivity is particularly acute when one wants to set a one-sided limit, *e.g.*, an upper limit on a cross section. Here one tests a value of a rate parameter μ against the alternative of a lower rate, and therefore the critical region of the test is taken to correspond to data outcomes with a low event yield. If the number of events found in the search region fluctuates low enough, however, it can happen that all physically meaningful signal parameter values, including those to which one has very little sensitivity, are rejected by the test. Another solution to the problem, therefore, is to replace the one-sided test by one based on the likelihood ratio, where the critical region is not restricted to low rates. This is the approach followed in the Feldman-Cousins procedure described above. Further properties of Feldman-Cousins intervals are discussed below in Section 36.3.2.6.

36.3.2.3. *Profile likelihood and treatment of nuisance parameters:*

As mentioned in Section 36.3.1, one may have a model containing parameters that must be determined from data, but which are not of any interest in the final result (nuisance parameters). Suppose the likelihood $L(\boldsymbol{\theta}, \boldsymbol{\nu})$ depends on parameters of interest $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\nu}$. The nuisance parameters can be effectively removed from the problem by constructing the *profile likelihood*, defined by

$$L_{\mathrm{p}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\nu}}}(\boldsymbol{\theta})) \,, \tag{36.53}$$

where $\widehat{\widehat{\boldsymbol{\nu}}}(\boldsymbol{\theta})$ is given by the $\boldsymbol{\nu}$ that maximizes the likelihood for fixed $\boldsymbol{\theta}$. The profile likelihood may then be used to construct tests of intervals for the parameters of interest. This is in contrast to the marginal likelihood (36.30) used in the Bayesian approach. For example, one may construct the profile likelihood ratio,

$$\lambda_p(\boldsymbol{\theta}) = \frac{L_{\mathrm{p}}(\boldsymbol{\theta})}{L(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\nu}})} \,, \tag{36.54}$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\nu}}$ are the ML estimators. The ratio λ_{p} can be used in place of the likelihood ratio (36.51) for inference about $\boldsymbol{\theta}$. The resulting intervals for the parameters of interest are not guaranteed to have the exact coverage probability for all values of the nuisance parameters, but in cases of practical interest the approximation is found to be very good. Further discussion on use of the profile likelihood can be found in, *e.g.*, Refs. [37–39] and other contributions to the PHYSTAT conferences [14].

36.3.2.4. *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,

$$1 - \alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\, dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{36.55}$$

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. 36.4 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 36.1.

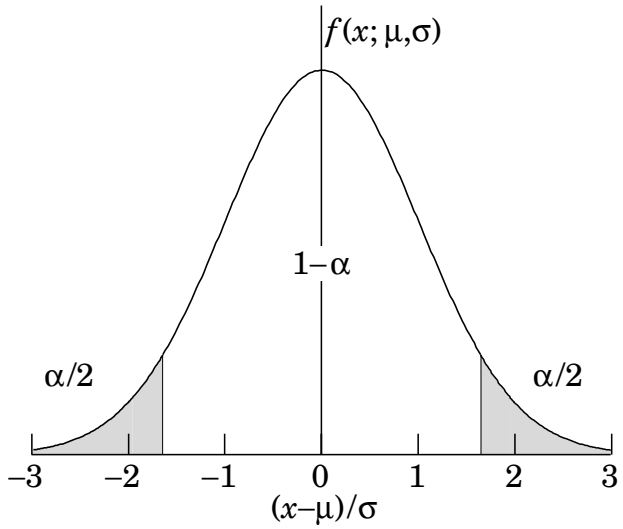

Figure 36.4: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by $\alpha = 0.1$, are as shown.

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 36.1.

The relation (36.55) can be re-expressed using the cumulative distribution function for the χ^2 distribution as

$$\alpha = 1 - F(\chi^2; n)\ , \tag{36.56}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom. This can be obtained from Fig. 36.1 on the $n = 1$ curve or by using the ROOT function `TMath::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

Table 36.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
$4.55\ \times 10^{-2}$	2σ	0.1	1.64σ
$2.7\ \times 10^{-3}$	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

in Sections 36.1.2 and 36.1.3. 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. 36.5, corresponding to a contour $\chi^2 = \chi^2_{\rm min} + 1$ or $\ln L = \ln L_{\rm max} - 1/2$. 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

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}\ , \tag{36.57}$$

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

The correlation coefficient can be visualized as the fraction of the distance σ_i from the ellipse's horizontal center-line at which the ellipse becomes tangent to vertical, *i.e.*, at the distance $\rho_{ij}\sigma_i$ below the center-line 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 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 $\sigma_{\rm inner}$ as shown on the figure, where $\sigma_{\rm inner} = (1 - \rho_{ij}^2)^{1/2}\sigma_i$. 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 $d\widehat{\theta}_i/d\theta_j = \rho_{ij} \times \frac{\sigma_i}{\sigma_j}$.

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

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\rm max} - \Delta \ln L\ , \tag{36.58}$$

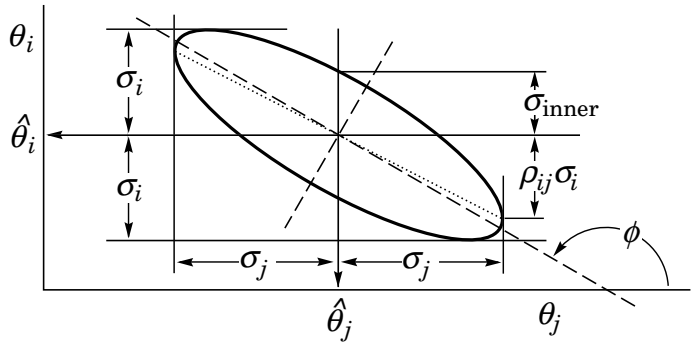

Figure 36.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

Table 36.2: $\Delta\chi^2$ 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

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

$$\chi^2(\boldsymbol{\theta}) \leq \chi^2_{\text{min}} + \Delta\chi^2 \ . \tag{36.59}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in Table 36.2 for several values of the coverage probability and number of fitted parameters.

For finite non-Gaussian data samples, the probability for the regions determined by equations (36.58) or (36.59) 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 36.2. 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,

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln L_i(\boldsymbol{\theta}) \ . \tag{36.60}$$

This can then be used to determine an approximate confidence interval or region with Eq. (36.58), just as with a single experiment.

36.3.2.5. *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

$$\nu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha_{\rm lo}; 2n) \ , \tag{36.61a}$$

$$\nu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{\rm up}; 2(n+1)) \ , \tag{36.61b}$$

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 ROOT routine `TMath::ChisquareQuantile`. 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 Eq. (36.61b) 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 36.3. 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

$$p_{\rm lo} = \frac{nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]}{N-n+1 \ + \ nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]} \ , \tag{36.62a}$$

$$p_{\rm up} = \frac{(n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]}{(N-n) \ + \ (n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]} \ . \tag{36.62b}$$

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

Table 36.3: Lower and upper (one-sided) 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

36.3.2.6. *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 interval that gives the upper limit for s may be of zero length.

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. (36.49). 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 confidence interval (*i.e.*, an interval based on a two-sided test with equal upper and lower tail areas). As pointed out by Feldman and Cousins [31], however, if the decision to report an upper limit or two-sided interval is made by looking at the data ("flip-flopping"), then in general there will be parameter values for which the resulting intervals have a coverage probability less than $1 - \alpha$.

With the confidence intervals suggested in [31], the prescription determines whether the interval is one- or two-sided in a way which preserves the coverage probability (and are thus said to be *unified*) and in addition they avoid the problem of null intervals. The intervals based on the Feldman-Cousins prescription are of this type. For a given choice of $1 - \alpha$, if the parameter estimate is sufficiently close to the boundary, 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 36.4. (Note that α in Ref. 31 is defined following Neyman [29] 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 36.3 for the one-sided upper limit, however, $1 - \alpha$ refers to the probability to have $\nu_{\rm up} \geq \nu$ (or $\nu_{\rm lo} \leq \nu$ for lower limits).

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 setting limits, the value $s = 0$ may be found below the lower edge of 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.

It must then be communicated carefully that in excluding $s = 0$ at, say, 90 or 95% confidence level from the interval, one is not necessarily claiming to have discovered the effect, for which one would usually require a higher level of significance (*e.g.*, 5 σ).

Table 36.4: 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

The intervals constructed according to the unified procedure in Ref. 31 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 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. 36).

Another possibility is to construct a Bayesian interval as described in Section 36.3.1. The presence of the boundary can be incorporated simply by setting the prior density to zero in the unphysical region. More specifically, the prior may be chosen using formal rules such as the reference prior or Jeffreys prior mentioned in Sec. 36.1.4. The use of such priors is currently receiving increased attention in HEP.

In HEP a widely used prior for the mean μ of a Poisson distributed measurement has been uniform for $\mu \geq 0$. This prior does not follow from any fundamental rule nor can it be regarded as reflecting a reasonable degree of belief, since the prior probability for μ to lie between any two finite limits is zero. It is more appropriately regarded as a procedure for obtaining intervals with frequentist properties that can be investigated. The resulting upper limits have a coverage probability that depends on the true value of the Poisson parameter, and is nowhere smaller 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 [2,19]. 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. (36.60)).

An additional alternative is presented by the intervals found from the likelihood function or χ^2 using the prescription of Equations (36.58) or (36.59). However, the coverage probability is not, in general, independent of the true parameter, and 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. 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 $1/\theta$. A study of the coverage of different intervals for a Poisson parameter can be found in [34]. Use of the likelihood function to determine approximate confidence intervals is discussed further in [35].

In any case, it is important to always 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.

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 from the interval alone.

References:

1. B. Efron, Am. Stat. **40**, 11 (1986).
2. R.D. Cousins, Am. J. Phys. **63**, 398 (1995).
3. A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model*, 6th ed., Oxford Univ. Press (1999), 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. F.E. James, *Statistical Methods in Experimental Physics*, 2nd ed., (World Scientific, Singapore, 2007).
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. Instrum. Methods **A297**, 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. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (2nd edition, Springer, New York, 2009).
12. A. Webb, *Statistical Pattern Recognition*, 2nd ed., (Wiley, New York, 2002).
13. L.I. Kuncheva, *Combining Pattern Classifiers*, (Wiley, New York, 2004).
14. Links to the *Proceedings of the PHYSTAT* conference series (Durham 2002, Stanford 2003, Oxford 2005, and Geneva 2007) can be found at `phystat.org`.
15. A. Höcker *et al.*, *TMVA Users Guide*, `physics/0703039`(2007); software available from `tmva.sf.net`.
16. I. Narsky, *StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data*, `physics/0507143`(2005); software avail. from `sourceforge.net/projects/statpatrec`.
17. L. Demortier, *P-Values and Nuisance Parameters*, *Proceedings of PHYSTAT 2007*, CERN-2008-001, p. 23.
18. E. Gross and O. Vitells, Eur. Phys. J. **C70**, 525 (2010); `arXiv:1005.1891`.
19. B.P. Roe and M.B. Woodroofe, Phys. Rev. **D63**, 13009 (2000).
20. A. O'Hagan and J.J. Forster, *Bayesian Inference*, (2nd edition, volume 2B of *Kendall's Advanced Theory of Statistics*, Arnold, London, 2004).
21. Devinderjit Sivia and John Skilling, *Data Analysis: A Bayesian Tutorial*, (Oxford University Press, 2006).
22. P.C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, (Cambridge University Press, 2005).
23. J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, (Wiley, 2000).
24. Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc. **91**, 1343 (1996).
25. J.M. Bernardo, J. R. Statist. Soc. **B41**, 113 (1979); J.M. Bernardo and J.O. Berger, J. Am. Stat. Assoc. **84**, 200 (1989). See also J.M. Bernardo, *Reference Analysis*, in *Handbook of Statistics*, 25 (D.K. Dey and C.R. Rao, eds.), 17-90, Elsevier (2005) and references therein.
26. L. Demortier, S. Jain, and H. Prosper, Phys. Rev. **D 82**, 034002 (2010); `arXiv:1002.1111`.
27. P.H. Garthwaite, I.T. Jolliffe, and B. Jones, *Statistical Inference*, (Prentice Hall, 1995).
28. R.D. Cousins and V.L. Highland, *Incorporating systematic uncertainties into an upper limit*, Nucl. Instrum. Methods **A320**, 331 (1992).
29. 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).
30. Robert E. Kass and Adrian E. Raftery, *Bayes Factors*, J. Am. Stat. Assoc. **90**, 773 (1995).
31. 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. 21.6 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.

32. 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 `cdsweb.cern.ch`.
33. T. Junk, Nucl. Instrum. Methods **A434**, 435 (1999).
34. Joel Heinrich, *Coverage of Error Bars for Poisson Data*, CDF Statistics Note 6438, `www-cdf.fnal.gov/publications/cdf6438_coverage.pdf` (2003).
35. F. Porter, Nucl. Instrum. Methods **A368**, 793 (1996).
36. 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 `cdsweb.cern.ch`. See also the Fermilab workshop at `conferences.fnal.gov/cl2k/`.
37. N. Reid, *Likelihood Inference in the Presence of Nuisance Parameters*, *Proceedings of PHYSTAT2003*, L. Lyons, R. Mount, and R. Reitmeyer, eds., eConf C030908, Stanford, 2003.
38. W.A. Rolke, A.M. Lopez, and J. Conrad, Nucl. Instrum. Methods **A551**, 493 (2005); `physics/0403059`.
39. Glen Cowan *et al.*, Eur. Phys. J. **C71**, 1554 (2011).