1. STATISTICS

Revised September 2013 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 1.2, statistical tests, discussed in Section 1.3, and confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, as described in Section 1.4.2. Note that in frequentist statistics one does not define a probability for a hypothesis or for the value of a parameter.

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. Using Bayes' theorem (Eq. (31.4)), the prior degree of belief is updated by the data from the experiment. Bayesian methods for interval estimation are discussed in Sections 1.4.1 and 1.4.2.5.

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.

1.1. Fundamental concepts

Consider an experiment whose outcome is characterized by one or more data values, which we can write as a vector $\boldsymbol{x}$. A *hypothesis* H is a statement about the probability for the data, often written $P(\boldsymbol{x}|H)$. (We will usually use a capital letter for a probability and lower case for a probability density. Often the term p.d.f. is used loosely to refer to either a probability or a probability density.) This 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 not determined (a composite hypothesis).

If the probability $P(\boldsymbol{x}|H)$ for data $\boldsymbol{x}$ is regarded as a function of the hypothesis H, then it is called the *likelihood* of H, usually written $L(H)$. Often the hypothesis is characterized by one or more parameters $\boldsymbol{\theta}$, in which case $L(\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})$ is called the likelihood function.

In some cases one can obtain at least approximate frequentist results using the likelihood evaluated only with the data obtained. In general, however, the frequentist approach requires a full specification of the probability model $P(\boldsymbol{x}|H)$ both as a function of the data $\boldsymbol{x}$ and hypothesis H.

In the Bayesian approach, inference is based on the posterior probability for H given the data $\boldsymbol{x}$, which represents one's degree of belief that H is true given the data. This is obtained from Bayes'

theorem (31.4), which can be written

$$P\left(H|\boldsymbol{x}\right) = \frac{P\left(\boldsymbol{x}|H\right)\pi\left(H\right)}{\int P\left(\boldsymbol{x}|H'\right)\pi\left(H'\right)\,dH'}\ . \tag{1.1}$$

Here $P(\boldsymbol{x}|H)$ is the likelihood for H, which depends only on the data actually obtained. The quantity $\pi(H)$ is the prior probability for H, which represents one's degree of belief for H before carrying out the measurement. The integral in the denominator (or sum, for discrete hypotheses) serves as a normalization factor. If H is characterized by a continuous parameter θ then the posterior probability is a p.d.f. $p(\theta|\boldsymbol{x})$. Note that the likelihood function itself is not a p.d.f. for θ.

1.2. 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 1.2.2 and 1.2.3. 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.

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

Suppose we have a set of n independent measurements, $x_1, \ldots, x_n$, each assumed to follow a p.d.f. with unknown mean μ and unknown variance σ^2. The measurements do not necessarily have to follow a Gaussian distribution. Then

$$\widehat{\mu} \ = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.5}$$

$$\widehat{\sigma^2} \ = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}\right)^2 \tag{1.6}$$

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

$$V\left[\widehat{\sigma^2}\right] \ = \ \frac{1}{n}\left(m_4 \ - \ \frac{n-3}{n-1}\sigma^4\right)\ , \tag{1.7}$$

where m_4 is the 4th central moment of x (see Eq. (37.8b)). 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 $\sigma/\sqrt{2n}$. For any n and Gaussian x_i, $\widehat{\mu}$ is an efficient estimator for μ, and the estimators $\widehat{\mu}$ and $\widehat{\sigma^2}$ are uncorrelated. Otherwise the arithmetic mean (1.5) is not necessarily the most efficient estimator.

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

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{n} w_i x_i\ , \tag{1.8}$$

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

1.2.2. *The method of maximum likelihood*:

Suppose we have a set of measured quantities $\boldsymbol{x}$ and the likelihood $L(\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})$ for a set of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. The

maximum likelihood (ML) estimators for $\boldsymbol{\theta}$ are defined as the values that give the maximum of L. Because of the properties of the logarithm, it is usually easier to work with $\ln L$, and since both are maximized for the same parameter values $\boldsymbol{\theta}$, the ML estimators can be found by solving the *likelihood equations*,

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \ , \qquad i = 1, \ldots, N \ . \tag{1.9}$$

In evaluating the likelihood function, it is important that any normalization factors in the p.d.f. that involve $\boldsymbol{\theta}$ be included.

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

$$\left(\widehat{V}^{-1}\right)_{ij} = - \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\widehat{\theta}} \ ; \tag{1.10}$$

for finite samples, however, Eq. (1.10) 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 hypersurface defined by the $\boldsymbol{\theta}'$ such that

$$\ln L\left(\boldsymbol{\theta}'\right) = \ln L_{\max} - s^2/2 \ , \tag{1.11}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (1.61)). The minimum and maximum values of θ_i on the hypersurface then give an approximate s-standard deviation confidence interval for θ_i (see Section 32.3.2.4).

1.2.2.1. *ML with binned data:*

1.2.3. *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 $\mu(x_i; \boldsymbol{\theta})$ 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

$$\chi^2\left(\boldsymbol{\theta}\right) = -2 \ln L\left(\boldsymbol{\theta}\right) + \text{ constant } = \sum_{i=1}^{N} \frac{\left(y_i - \mu\left(x_i; \boldsymbol{\theta}\right)\right)^2}{\sigma_i^2} \ . \tag{1.13}$$

The parameter values that maximize L are the same as those which minimize χ^2.

The minimum of Equation (1.13) defines the least-squares estimators $\widehat{\boldsymbol{\theta}}$ for the more general case where the y_i are not Gaussian distributed as long as they are independent. If they are not independent but rather have a covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2\left(\boldsymbol{\theta}\right) = \left(\boldsymbol{y} - \boldsymbol{\mu}\left(\boldsymbol{\theta}\right)\right)^T V^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\left(\boldsymbol{\theta}\right)\right) \ , \tag{1.14}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)$ is the (column) vector of measurements, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the corresponding vector of predicted values, and the superscript T denotes the transpose.

Often one further restricts the problem to the case where $\mu(x_i; \boldsymbol{\theta})$ is a linear function of the parameters, *i.e.*,

$$\mu\left(x_i; \boldsymbol{\theta}\right) = \sum_{j=1}^{m} \theta_j h_j\left(x_i\right) \ . \tag{1.15}$$

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,

$$\widehat{\boldsymbol{\theta}} = \left(H^T V^{-1} H\right)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y} \ . \tag{1.16}$$

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

$$U = DVD^T = \left(H^T V^{-1} H\right)^{-1} \ . \tag{1.17}$$

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

$$\chi^2\left(\boldsymbol{\theta}\right) = \chi^2\left(\widehat{\boldsymbol{\theta}}\right) + 1 = \chi^2_{\mathrm{min}} + 1 \tag{1.23}$$

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

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 32.2.2.

1.2.4. *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. For general introductions to Bayesian statistics see, *e.g.*, Refs. [22–25].

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,

$$p\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \frac{P\left(\boldsymbol{x}|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}'\right)\pi\left(\boldsymbol{\theta}'\right)\, d\boldsymbol{\theta}'} \ , \tag{1.24}$$

where $P(\boldsymbol{x}|\boldsymbol{\theta})$ 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. (1.24) 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) values, and covariance matrix. In addition one may construct intervals with a given probability content, as is discussed in Sec. 1.4.1 on Bayesian interval estimation.

For the special case of a constant prior, one can see from Bayes' theorem (1.24) 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. One may use as the Bayesian estimator a summary statistic other than the mode, such as the median, which is

invariant under parameter transformation. But this will not in general coincide with the ML estimator.

1.2.4.1. *Bayesian treatment of nuisance parameters:*

As discussed in Sec. 1.2.2, a model may depend on parameters of interest $\boldsymbol{\theta}$ as well as on nuisance parameters $\boldsymbol{\nu}$, which must be included for an accurate description of the data. Knowledge about the values of $\boldsymbol{\nu}$ may be supplied by control measurements, theoretical insights, physical constraints, etc. Suppose, for example, one has data $\boldsymbol{y}$ from a control measurement which is characterized by a probability $P_y(\boldsymbol{y}|\boldsymbol{\nu})$. Suppose further that before carrying out the control measurement one's state of knowledge about $\boldsymbol{\nu}$ is described by an initial prior $\pi_0(\boldsymbol{\nu})$, which in practice is often taken to be a constant or in any case very broad. By using Bayes' theorem (1.1) one obtains the updated prior $\pi(\boldsymbol{\nu})$ (*i.e.*, now $\pi(\boldsymbol{\nu}) = \pi(\boldsymbol{\nu}|\boldsymbol{y})$, the probability for $\boldsymbol{\nu}$ given $\boldsymbol{y}$),

$$\pi\left(\boldsymbol{\nu}|\boldsymbol{y}\right) \propto P\left(\boldsymbol{y}|\boldsymbol{\nu}\right)\pi_0\left(\boldsymbol{\nu}\right) \ . \tag{1.27}$$

In the absence of a model for $P(\boldsymbol{y}|\boldsymbol{\nu})$ one may make some reasonable but ad hoc choices. For a single nuisance parameter ν, for example, one might characterize the uncertainty in a nuisance parameter ν by a p.d.f. $\pi(\nu)$ 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. Note also that truncation of the prior of a nuisance parameter ν at zero will in general make $\pi(\nu)$ nonzero at $\nu = 0$, which can lead to an unnormalizable posterior for a parameter of interest that appears multiplied by ν.

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. Note that the likelihood here only refers to the primary measurement $\boldsymbol{x}$. Once any control measurements $\boldsymbol{y}$ are used to find the updated prior $\pi(\boldsymbol{\nu})$ for the nuisance parameters, this information is fully encapsulated in $\pi(\boldsymbol{\nu})$ and the control measurements do not appear further.

One can obtain the posterior p.d.f. for $\boldsymbol{\theta}$ alone by integrating over the nuisance parameters, *i.e.*,

$$p\left(\boldsymbol{\theta}|\boldsymbol{x}\right) = \int p\left(\boldsymbol{\theta},\boldsymbol{\nu}|\boldsymbol{x}\right)\, d\boldsymbol{\nu} \ . \tag{1.28}$$

1.2.5. ***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

$$U_{ij} \approx \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l}\bigg|_{\widehat{\theta}} V_{kl} \ . \tag{1.29}$$

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

$$A_{ij} = \left.\frac{\partial \eta_i}{\partial \theta_j}\right|_{\widehat{\theta}} , \tag{1.30}$$

and A^T is its transpose. The approximation is exact if $\boldsymbol{\eta}(\boldsymbol{\theta})$ is linear.

1.3. Statistical tests

1.3.1. *Hypothesis tests*:

A frequentist *test* of a hypothesis (often called the null hypothesis, H_0) is a rule that states for which data values $\boldsymbol{x}$ the hypothesis is rejected. A region of $\boldsymbol{x}$-space called the critical region, w, is specified such that such that there is no more than a given 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 be 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.

The critical region is not unique. Choosing 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 $\beta = P(\boldsymbol{x} \notin w|H_1)$. 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.

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

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} , \tag{1.31}$$

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 (1.31) 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 (*i.e.*, the maximum probability to reject H_0 if H_1 is true) for a given probability α to reject H_0 if H_0 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 (1.31) 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.

1.3.2. *Tests of significance (goodness-of-fit)*:

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 hypothesis in question, H_0, will determine the p.d.f. $f(t|H_0)$ 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

$$p = \int_{t_{\rm obs}}^{\infty} f\left(t|H_0\right)\, dt \ , \tag{1.32}$$

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 1.4), 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 $p \leq \alpha$ implies that the data outcome was in the critical region. When constructing a p-value, one generally chooses the region of data space deemed to have lower compatibility with the model being tested as one having higher compatibility with a given alternative, such that 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.*,

$$Z = \Phi^{-1}\left(1-p\right) \ . \tag{1.33}$$

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", discussed in Section 1.3.2.1).

1.3.2.1. *The look-elsewhere effect:*

The "look-elsewhere effect" relates to multiple measurements used to test a single hypothesis. The classic example is when one searches in a distribution for a peak whose position is not predicted in advance. Here the no-peak hypothesis is tested using data in a given range of the distribution. In the frequentist approach the correct p-value of the no-peak hypothesis is the probability, assuming background only, to find a signal as significant as the one found or more so anywhere in the search region. This can be substantially higher than the probability to find a peak of equal or greater significance in the particular place where it appeared. There is in general some ambiguity as to what constitutes the relevant search region or even the broader set of relevant measurements. Although the desired p-value is well defined once the search region has been fixed, an exact treatment can require extensive computation.

The "brute-force" solution to this problem by Monte Carlo involves generating data under the background-only hypothesis and for each data set, fitting a peak of unknown position and recording a measure of its significance. To establish a discovery one often requires a p-value less than 2.9×10^{-7}, corresponding to a 5σ or larger effect. Determining this with Monte Carlo thus requires generating and fitting a very large number of experiments, perhaps several times 10^7. In contrast, if the position of the peak is fixed, then the fit to the distribution is much easier, and furthermore one can in many cases use formulae valid for sufficiently large samples that bypass completely the need for Monte Carlo (see, e.g., [38]) . But this fixed-position or "local" p-value would not be correct in general, as it assumes the position of the peak was known in advance.

A method that allows one to modify the local p-value computed under assumption of a fixed position to obtain an approximation to the correct "global" value using a relatively simple calculation is described in Ref. 18. Suppose a test statistic q_0, defined so that larger values indicate increasing disagreement with the data, is observed to have a value u. Furthermore suppose the model contains a nuisance parameter θ (such as the peak position) which is only defined under the signal model (there is no peak in the background-only model). An approximation for the global p-value is found to be

$$p_{\text{global}} \approx p_{\text{local}} + \langle N_u \rangle \ , \tag{1.36}$$

where $\langle N_u \rangle$ is the mean number of "upcrossings" of the the statistic q_0 above the level u in the range of the nuisance parameter considered (e.g., the mass range).

The value of $\langle N_u \rangle$ can be estimated from the number of upcrossings $\langle N_{u_0} \rangle$ above some much lower value, u_0, by using a relation due to Davis [19],

$$\langle N_u \rangle \approx \langle N_{u_0} \rangle e^{-(u-u_0)/2} \ . \tag{1.37}$$

By choosing u_0 sufficiently low, the value of $\langle N_u \rangle$ can be estimated by simulating only a very small number of experiments or even from the observed data, rather than the 10^7 needed if one is dealing with a 5σ effect.

1.3.2.2. *Goodness-of-fit with the method of Least Squares:*

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity χ^2 (1.13). 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 $\mu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \mu_i$, the χ^2 (1.13) becomes *Pearson's* χ^2 *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \mu_i)^2}{\mu_i} \ . \tag{1.38}$$

If the hypothesis $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$ is correct, and if the expected values μ_i in (1.38) 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.

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

$$p = \int_{\chi^2}^{\infty} f\left(z; n_{\mathrm{d}}\right) \, dz \ , \tag{1.39}$$

where $f(z; n_{\mathrm{d}})$ is the χ^2 p.d.f. and n_{d} is the appropriate number of degrees of freedom. Values are shown in Fig. 32.1 or obtained from the ROOT function `TMath::Prob`.

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. 1.2.

1.3.3. ***Bayes factors***:

In Bayesian statistics, all of one's knowledge about a model is contained in its posterior probability, which one obtains using Bayes' theorem (1.24). 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

$$\pi\left(H_i, \boldsymbol{\theta}_i\right) = P\left(H_i\right) \pi\left(\boldsymbol{\theta}_i | H_i\right) \ . \tag{1.40}$$

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,

$$P\left(H_i|\boldsymbol{x}\right) = \frac{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) P\left(H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right) \, d\boldsymbol{\theta}_i}{P\left(\boldsymbol{x}\right)} \ , \tag{1.41}$$

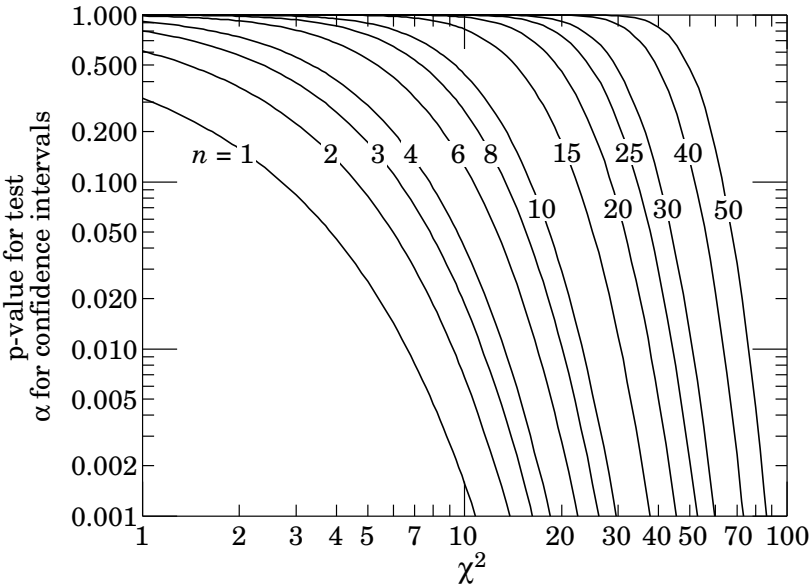

Figure 1.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 32.3.2.4).

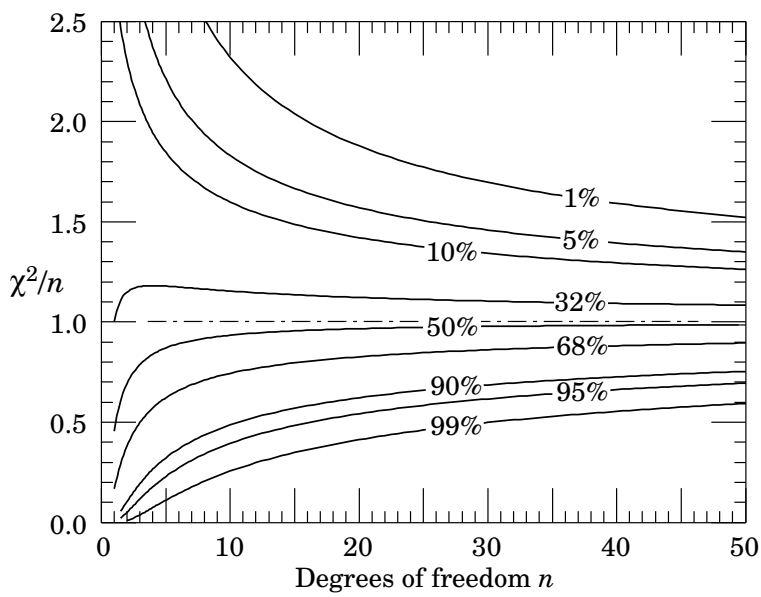

Figure 1.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

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

$$\frac{P\left(H_i|\boldsymbol{x}\right)}{P\left(H_j|\boldsymbol{x}\right)} = \frac{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right)\, d\boldsymbol{\theta}_i}{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j\right) \pi\left(\boldsymbol{\theta}_j|H_j\right)\, d\boldsymbol{\theta}_j} \frac{P\left(H_i\right)}{P\left(H_j\right)} \,. \tag{1.42}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i\right) \pi\left(\boldsymbol{\theta}_i|H_i\right)\, d\boldsymbol{\theta}_i}{\int P\left(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j\right) \pi\left(\boldsymbol{\theta}_j|H_j\right)\, d\boldsymbol{\theta}_j} \ . \tag{1.43}$$

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 (1.43) are of the form

$$m = \int P\left(\boldsymbol{x}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right)\, d\boldsymbol{\theta} \ , \tag{1.44}$$

which is the marginal likelihood seen previously in Eq. (38.44) (in some fields this quantity is called the *evidence*). A review of Bayes factors can be found in Ref. 30. Computing marginal likelihoods can be difficult; in many cases it can be done with the nested sampling algorithm [31] as implemented, *e.g.*, in the program `MultiNest` [32].

1.4. 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 $\widehat{\theta}$, $\widehat{\sigma}_{\widehat{\theta}}$. 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.

1.4.1. *Bayesian intervals*:

As described in Sec. 1.2.4, 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.*,

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p\left(\theta|\boldsymbol{x}\right)\, d\theta \ . \tag{1.45}$$

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 be interested in the set of θ values for which $p(\theta|\boldsymbol{x})$ is higher than for any θ not belonging to the set, which may constitute a single interval or a set of disjoint regions; 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

$$\pi\left(s\right) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \ . \tag{1.46}$$

For example, to obtain an upper limit on s, one may proceed as follows. The likelihood for s is given by the Poisson distribution for n with mean $s+b$,

$$P\left(n|s\right) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{1.47}$$

along with the prior (1.46) in (1.24) 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

$$1-\alpha = \int_{-\infty}^{s_{\rm up}} p\left(s|n\right) ds = \frac{\int_{-\infty}^{s_{\rm up}} P\left(n|s\right)\, \pi\left(s\right)\, ds}{\int_{-\infty}^{\infty} P\left(n|s\right)\, \pi\left(s\right)\, ds} \ , \tag{1.48}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (1.48) to incomplete gamma functions, the solution for the upper limit is found to be

$$s_{\rm up} = \tfrac{1}{2} F_{\chi^2}^{-1}\left[p, 2\left(n+1\right)\right] - b \ , \tag{1.49}$$

where $F_{\chi^2}^{-1}$ is the quantile of the χ^2 distribution (inverse of the cumulative distribution). Here the quantity p is

$$p = 1-\alpha\left(F_{\chi^2}\left[2b, 2\left(n+1\right)\right]\right) \ , \tag{1.50}$$

where F_{χ^2} is the cumulative χ^2 distribution. For both F_{χ^2} and $F_{\chi^2}^{-1}$ above, the argument $2(n+1)$ gives the number of degrees of freedom. For the special case of $b=0$, the limit reduces to

$$s_{\rm up} = \tfrac{1}{2} F_{\chi^2}^{-1}\left(1-\alpha; 2\left(n+1\right)\right) \ . \tag{1.51}$$

It happens that for the case of $b=0$, the upper limit from Eq. (1.51) coincides numerically with the frequentist upper limit discussed in Section 1.4.2.4. Values for $1-\alpha = 0.9$ and 0.95 are given by the values $\mu_{\rm up}$ in Table 1.3.

As in any Bayesian analysis, it is important to show how the result changes under assumption of different prior probabilities. For example, one could consider the Jeffreys prior as described in Sec. 1.2.4. 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.

If the model contains nuisance parameters then these are eliminated by marginalizing, as in Eq. (1.28), to obtain the p.d.f. for the parameters of interest. For example, if the parameter b in the Poisson counting problem above were to be characterized by a prior p.d.f. $\pi(b)$, then one would first use Bayes' theorem to find $p(s, b|n)$. This is then marginalized to find $p(s|n) = \int p(s, b|n)\pi(b)\, db$, from which one may determine an interval for s. One may not be certain whether to extend a model by including more nuisance parameters. In this case, a Bayes factor may be used to determine to what extent the data prefer a model with additional parameters, as described in Section 1.3.3.

1.4.2. *Frequentist confidence intervals*:

1.4.2.1. *The Neyman construction for confidence intervals:*

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

$$P\left(x_1 < x < x_2; \theta\right) = 1 - \alpha = \int_{x_1}^{x_2} f\left(x;\theta\right)\, dx\ . \tag{1.53}$$

This is illustrated in Fig. 1.3: 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.

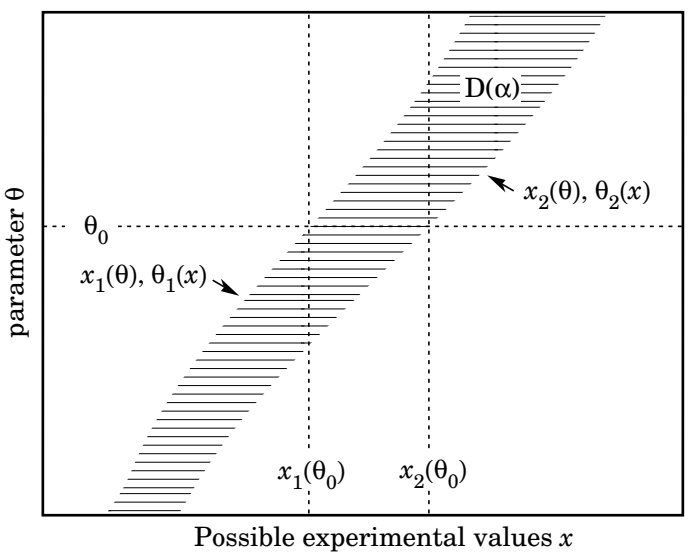

Figure 1.3: Construction of the confidence belt (see text).

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

$$1-\alpha = P\left(x_1\left(\theta\right) < x < x_2\left(\theta\right)\right) = P\left(\theta_2\left(x\right) < \theta < \theta_1\left(x\right)\right) \ . \qquad (1.54)$$

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. (1.53) 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. (1.53) 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 $1-\alpha$.

If the model contains nuisance parameters $\boldsymbol{\nu}$, then these can be incorporated into the test (or the p-values) used to determine the limit by profiling as discussed in Section 38.3.2.1. As mentioned there, the strict frequentist approach is to regard the parameter of interest θ as excluded only if it is rejected for all possible values of $\boldsymbol{\nu}$. The resulting interval for θ will then cover then cover the true value with a probability greater than or equal to the nominal confidence level for all points in $\boldsymbol{\nu}$-space.

If the p-value is based on the profiled values of the nuisance parameters, *i.e.*, with $\boldsymbol{\nu} = \widehat{\widehat{\boldsymbol{\nu}}}(\theta)$ used in Eq. (38.42), then the resulting interval for the parameter of interest will have the correct coverage if the true values of $\boldsymbol{\nu}$ are equal to the profiled values. Otherwise the coverage probability may be too high or too low. This procedure has been called *profile construction* in HEP [20](see also [17]) .

1.4.2.3. *Gaussian distributed measurements:*

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

the measured value x will fall within $\pm\delta$ of the true value μ is

$$1-\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\,dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) = 2\Phi\left(\frac{\sigma}{\delta}\right) - 1\,, \tag{1.58}$$

where erf is the Gaussian error function, which is rewritten in the final equality using Φ, the Gaussian cumulative distribution. Fig. 1.4 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 1.1.

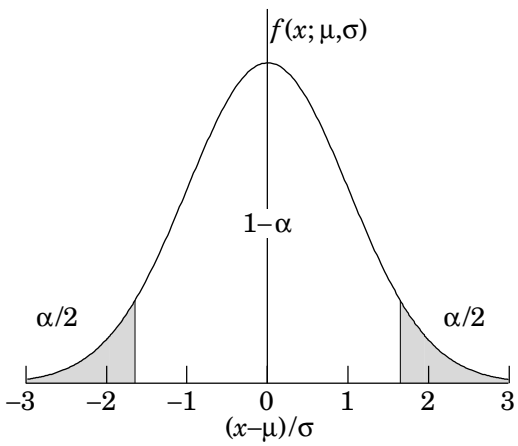

Figure 1.4: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by $\alpha = 0.1$, are as shown.

Table 1.1: 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
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-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

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 1.1.

The relation (1.58) can be re-expressed using the cumulative distribution function for the χ^2 distribution as

$$\alpha = 1 - F\left(\chi^2; n\right)\,, \tag{1.59}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom. This can be seen as the $n = 1$ curve in Fig. 1.1 or obtained 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 in Sections 1.2.2 and 1.2.3. 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. 1.5, corresponding to a contour $\chi^2 = \chi^2_{\mathrm{min}} + 1$ or $\ln L = \ln L_{\mathrm{max}} - 1/2$. The ellipse is centered about the estimated values $\widehat{\boldsymbol{\theta}}$, and the tangents to the ellipse give the standard deviations of the estimators, σ_i and σ_j. The angle of the major axis of the ellipse is given by

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2} \ , \tag{1.60}$$

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

The correlation coefficient can be visualized as the fraction of the distance σ_i from the ellipse's horizontal 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.

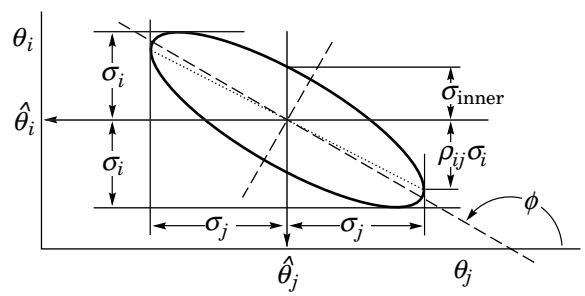

Figure 1.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

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

$$\ln L\,(\boldsymbol{\theta}) \geq \ln L_{\mathrm{max}} - \Delta \ln L \ , \tag{1.61}$$

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

$$\chi^2\,(\boldsymbol{\theta}) \leq \chi^2_{\mathrm{min}} + \Delta\chi^2 \ . \tag{1.62}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in Table 1.2 for several values of the coverage probability and number of fitted parameters.

For non-Gaussian data samples, the probability for the regions determined by equations (1.61) or (1.62) to cover the true value of $\boldsymbol{\theta}$ becomes independent of $\boldsymbol{\theta}$ only in the large-sample limit. So for a finite data sample these are not exact confidence regions according to our previous definition.

Table 1.2: Values of $\Delta\chi^2$ 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

1.4.2.4. *Poisson or binomial data:*

Another important class of measurements consists of counting a certain number of events, n. In this section, we will assume these are all events of the desired type, *i.e.*, there is no background. If n represents the number of events produced in a reaction with cross section σ, say, in a fixed integrated luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\mu = \sigma\mathcal{L}$. 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 μ can be found from the Neyman procedure to be

$$\mu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}\left(\alpha_{\rm lo}; 2n\right) , \tag{1.64a}$$

$$\mu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}\left(1-\alpha_{\rm up}; 2\left(n+1\right)\right) , \tag{1.64b}$$

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. (1.64*b*) coincides numerically with the Bayesian upper limit for a Poisson parameter, using a uniform prior p.d.f. for μ. Values for confidence levels of 90% and 95% are shown in Table 1.3. 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

$$p_{\rm lo} = \frac{nF^{-1}_F\left[\alpha_{\rm lo}; 2n, 2\left(N-n+1\right)\right]}{N-n+1 + nF^{-1}_F\left[\alpha_{\rm lo}; 2n, 2\left(N-n+1\right)\right]} , \tag{1.65a}$$

$$p_{\rm up} = \frac{\left(n+1\right)F^{-1}_F\left[1-\alpha_{\rm up}; 2\left(n+1\right), 2\left(N-n\right)\right]}{\left(N-n\right) + \left(n+1\right)F^{-1}_F\left[1-\alpha_{\rm up}; 2\left(n+1\right), 2\left(N-n\right)\right]} . \tag{1.65b}$$

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

A number of issues arise in the construction and interpretation of confidence intervals when the parameter can only take on values

Table 1.3: Lower and upper (one-sided) limits for the mean μ 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	$\mu_{\rm lo}$	$\mu_{\rm up}$	$\mu_{\rm lo}$	$\mu_{\rm up}$
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

in a restricted range. Important examples are where the mean of a Gaussian variable is constrained on physical grounds to be non-negative and where the experiment finds a Poisson-distributed number of events, n, which includes both signal and background. Application of some standard recipes can lead to intervals that are partially or entirely in the unphysical region. Furthermore, if the decision whether to report a one- or two-sided interval is based on the data, then the resulting intervals will not in general cover the parameter with the stated probability $1-\alpha$.

Several problems with such intervals are overcome by using the unified approach of Feldman and Cousins [33]. Properties of these intervals are described further in the *Review*. Table 1.4 gives the unified confidence intervals $[\mu_1, \mu_2]$ for the mean of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%. The values of $1-\alpha$ given here refer to the coverage of the true parameter by the whole interval $[\mu_1, \mu_2]$. In Table 1.3 for the one-sided upper and lower limits, however, $1-\alpha$ referred to the probability to have individually $\mu_{\rm up} \geq \mu$ or $\mu_{\rm lo} \leq \mu$.

Another possibility is to construct a Bayesian interval as described in Section 1.4.1. The presence of the boundary can be incorporated simply by setting the prior density to zero in the unphysical region. Advantages and pitfalls of this approach are discussed further in the *Review*.

Another alternative is presented by the intervals found from the likelihood function or χ^2 using the prescription of Equations (1.61) or (1.62). As in the case of the Bayesian intervals, the coverage probability is not, in general, independent of the true parameter. Furthermore, these intervals can for some parameter values undercover.

In any case it is important to report sufficient information so that the result can be combined with other measurements. Often this

Table 1.4: Unified confidence intervals $[\mu_1, \mu_2]$ 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	μ_1	μ_2	μ_1	μ_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.99	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

means giving an unbiased estimator and its standard deviation, even if the estimated value is in the unphysical region. It is also useful to report the likelihood function or an appropriate summary of it. Although this by itself is not sufficient to construct a frequentist confidence interval, it can be used to find the Bayesian posterior probability density for any desired prior p.d.f.

Further discussion and all references may be found in the full *Review of Particle Physics;* the equation and reference numbering corresponds to that version.