40. Statistics

Revised August 2023 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 limiting frequency of the outcome of a repeatable experiment. The most important tools in this framework are parameter estimation, covered in Section 40.2, statistical tests, discussed in Section 40.3, and confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, as described in Section 40.4.2. 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 subjective interpretation of probability is used to quantify one's *degree of belief* in a hypothesis. This allows one to define a probability density function (p.d.f.) for a parameter, which reflects one's knowledge about where its true value lies.

Bayesian methods provide a natural means to include additional information, which in general may be subjective; in fact they *require* prior probabilities for the hypotheses (or parameters) in question, *i.e.*, the degree of belief about the parameters' values, before carrying out the measurement. Using Bayes' theorem (Eq. (39.4)), the prior degree of belief is updated by the data from the experiment. Bayesian methods for interval estimation are discussed in Sections 40.4.1 and 40.4.2.4.

For many inference problems, the frequentist and Bayesian approaches give similar numerical values, even though they answer different questions and are based on fundamentally different interpretations of probability. In some important cases, however, the two approaches may yield very different results. For a discussion of Bayesian vs. non-Bayesian methods, see references written by a statistician [1], by a physicist [2], or the detailed comparison in Ref. [3].

40.1 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, for example, when determining confidence regions with Eq. (40.79). 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 (39.4), which can be written

$$P(H|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|H)\pi(H)}{\int P(\boldsymbol{x}|H')\pi(H')\,dH'}\,. \tag{40.1}$$

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.

40.2 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 40.2.2 and 40.2.3. The Bayesian approach is outlined in Sec. 40.2.6.

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 in probability (see Ref. [3]) 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.
(c) *Efficiency* is the ratio of the minimum possible variance for any estimator of θ to the variance $V[\widehat{\theta}]$ of the estimator $\widehat{\theta}$. For the case of a single parameter, under rather general conditions the minimum variance is given by the Rao-Cramér-Fréchet bound,

$$\sigma^2_{\rm min} = \left(1 + \frac{\partial b}{\partial \theta}\right)^2 / I(\theta) \,, \tag{40.2}$$

where

$$I(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] \tag{40.3}$$

is the *Fisher information*, L is the likelihood, and the operator $E[]$ in (40.3) is the expectation value with respect to the data. For the final equality to hold, the range of allowed data values must not depend on θ.

The *mean-squared error*,

$$\text{MSE} = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + b^2 \,, \tag{40.4}$$

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 about the p.d.f., *e.g.*, owing to uncertainties in the distribution's tails.

It is not in general possible to optimize simultaneously for all the measures of estimator quality described above. 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 in general depend on the unknown θ.

40.2.1 *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

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{40.5}$$

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 \tag{40.6}$$

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

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{n}\left(m_4 - \frac{n-3}{n-1}\sigma^4\right) , \tag{40.7}$$

where m_4 is the 4^{th} central moment of x (see Eq. (39.8)). 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 $\widehat{\sigma}$ 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 (40.5) 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. (40.5); however, if μ is known, one can substitute it for $\widehat{\mu}$ in Eq. (40.6) 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

$$\widehat{\mu} = \frac{1}{w}\sum_{i=1}^{n} w_i x_i , \tag{40.8}$$

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 $1/\sqrt{w}$.

As an estimator for the median x_{med}, one can use the value $\widehat{x}_{\text{med}}$ such that half the x_i are below and half above (the sample median). If there are an even number of observations and the sample median lies between two observed values, the estimator is set by convention to their arithmetic average. 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 $f(0) > 0$ [5]. 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.

40.2.2 *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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$. The *maximum likelihood estimators* (MLEs) 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 MLEs can be found by solving the *likelihood equations*,

$$\frac{\partial \ln L}{\partial \theta_i} = 0 , \qquad i = 1, \ldots, M . \tag{40.9}$$

Often the solution must be found numerically. Maximum likelihood estimators are important because they are asymptotically (*i.e.*, for large data samples) unbiased, efficient and have a Gaussian sampling distribution under quite general conditions, and the method has a wide range of applicability.

In general the likelihood function is obtained from the probability of the data under assumption of the parameters. An important special case is when the data consist of *i.i.d.* (independent

and identically distributed) values. Here one has a set of n statistically independent quantities $\boldsymbol{x} = (x_1, \ldots, x_n)$, where each component follows the same p.d.f. $f(x; \boldsymbol{\theta})$. In this case the joint p.d.f. of the data sample factorizes and the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \; . \tag{40.10}$$

Here the number n of observations (usually individual "events" in particle physics) is regarded as fixed. If, however, the probability to observe n events itself depends on the parameters $\boldsymbol{\theta}$, then this dependence should be included in the likelihood. For example, if n follows a Poisson distribution with mean μ and the independent x values all follow $f(x; \boldsymbol{\theta})$, then the likelihood becomes

$$L(\boldsymbol{\theta}) = \frac{\mu^n}{n!} e^{-\mu} \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \; . \tag{40.11}$$

Equation (40.11) is often called the *extended likelihood* (see, *e.g.*, Refs. [6–8]). If μ is given as a function of $\boldsymbol{\theta}$, then including the probability for n given $\boldsymbol{\theta}$ in the likelihood provides additional information about the parameters. This therefore leads to a reduction in their statistical uncertainties and in general changes their estimated values.

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 MLEs $\widehat{\boldsymbol{\theta}}$ transform to $\boldsymbol{\eta}(\widehat{\boldsymbol{\theta}})$. That is, the maximum-likelihood solution is invariant under change of parameter. However, other properties of MLEs, 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 MLEs can be estimated by using

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\boldsymbol{\theta}}} \; . \tag{40.12}$$

Equation (40.12) holds for a sufficiently large data sample (or for a small sample only in special cases, e.g., where the means of Gaussian distributed data are linear functions of the parameters); otherwise it can result in a misestimation of the variances. Under the conditions where the equation is valid, L has a Gaussian form and $\ln L$ is (hyper)parabolic. In this case, s times the standard deviations σ_i of the estimators for the parameters can be obtained from the hypersurface defined by the $\boldsymbol{\theta}$ such that

$$\ln L(\boldsymbol{\theta}) = \ln L_{\max} - s^2/2 \; , \tag{40.13}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (40.79)). 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 40.4.2.2).

40.2.2.1 *Maximum likelihood with binned data*

If the total number of data values x_i, $(i = 1, \ldots, n_{\mathrm{tot}})$, is small, the unbinned maximum-likelihood method, *i.e.*, use of Equation (40.10) (or (40.11) for extended maximum likelihood), is preferred since binning can only result in a loss of information, and hence larger statistical errors for the parameter estimates. If the sample is large, it can be convenient to bin the values in a histogram with N bins, so that one obtains a vector of data $\boldsymbol{n} = (n_1, \ldots, n_N)$ with expectation

values $\boldsymbol{\mu} = E[\boldsymbol{n}]$ and probabilities $f(\boldsymbol{n};\boldsymbol{\mu})$. Suppose the mean values $\boldsymbol{\mu}$ can be determined as a function of a set of parameters $\boldsymbol{\theta}$. Then one may maximize the likelihood function based on the contents of the bins.

As mentioned in Sec. 40.2.2, the total number of events $n_{\text{tot}} = \sum_i n_i$ can be regarded either as fixed or as a random variable. If it is fixed, the histogram follows a multinomial distribution,

$$f_{\text{M}}(\boldsymbol{n};\boldsymbol{\theta}) = \frac{n_{\text{tot}}!}{n_1! \cdots n_N!} p_1^{n_1} \cdots p_N^{n_N} \ , \tag{40.14}$$

where we assume the probabilities p_i are given functions of the parameters $\boldsymbol{\theta}$. The distribution can be written equivalently in terms of the expected number of events in each bin, $\mu_i = n_{\text{tot}} p_i$. If the n_i are regarded as independent and Poisson distributed, then the data are instead described by a product of Poisson probabilities,

$$f_{\text{P}}(\boldsymbol{n};\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{\mu_i^{n_i}}{n_i!} e^{-\mu_i} \ , \tag{40.15}$$

where the mean values μ_i are given functions of $\boldsymbol{\theta}$. The total number of events n_{tot} thus follows a Poisson distribution with mean $\mu_{\text{tot}} = \sum_i \mu_i$.

When using maximum likelihood with binned data, one can find the estimators and at the same time obtain a statistic usable for a test of goodness of fit (see Sec. 40.3.3.1). For independent Poisson distributed data, maximizing the likelihood $L(\boldsymbol{\theta}) = f_{\text{P}}(\boldsymbol{n};\boldsymbol{\theta})$ is equivalent to maximizing the likelihood ratio $\lambda(\boldsymbol{\theta}) = f_{\text{P}}(\boldsymbol{n};\boldsymbol{\theta}) / f_{\text{sat}}(\boldsymbol{n};\hat{\boldsymbol{\mu}})$, where in the denominator $f_{\text{sat}}(\boldsymbol{n};\boldsymbol{\mu})$ is the corresponding model with an adjustable parameter for each bin, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$, with estimators $\hat{\boldsymbol{\mu}} = (n_1, \ldots, n_N)$ (called the "saturated model"). Equivalently one often minimizes the quantity $-2 \ln \lambda(\boldsymbol{\theta})$. For independent Poisson distributed n_i this is [9]

$$-2 \ln \lambda(\boldsymbol{\theta}) = 2 \sum_{i=1}^{N} \left[\mu_i(\boldsymbol{\theta}) - n_i + n_i \ln \frac{n_i}{\mu_i(\boldsymbol{\theta})} \right] \ , \tag{40.16}$$

where for bins with $n_i = 0$, the last term in (40.16) is zero. The expression (40.16) without the terms $\mu_i - n_i$ 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, minimizing (40.16) is equivalent to maximizing the unbinned extended likelihood function (40.11); in the corresponding multinomial case without the $\mu_i - n_i$ terms one obtains Eq. (40.10).

A smaller value of $-2 \ln \lambda(\widehat{\boldsymbol{\theta}})$ corresponds to better agreement between the data and the hypothesized form of $\boldsymbol{\mu}(\boldsymbol{\theta})$. The value of $-2 \ln \lambda(\widehat{\boldsymbol{\theta}})$ can thus be translated into a p-value as a measure of goodness-of-fit, as described in Sec. 40.3.3.1. Assuming the model is correct, then according to Wilks' theorem [10], for sufficiently large μ_i and provided certain regularity conditions are met, the minimum of $-2 \ln \lambda$ as defined by Eq. (40.16) follows a χ^2 distribution (see, *e.g.*, Ref. [9]). 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 $\mu_{\text{tot}} = \sum_i \mu_i$ is taken as an adjustable parameter, so that $\mu_i = \mu_{\text{tot}} p_i(\boldsymbol{\theta})$, where the probability to be in the ith bin, $p_i(\boldsymbol{\theta})$, does not depend on μ_{tot}. Then by minimizing Eq. (40.16), one obtains that the area under the fitted function is equal to the sum of the histogram contents, *i.e.*, $\sum_i \hat{\mu}_i = \sum_i n_i$. This is a property not possessed by the estimators from the method of least squares (see, *e.g.*, Sec. 40.2.3 and Ref. [8]).

40.2.2.2 *Frequentist treatment of nuisance parameters*

Suppose we want to determine the values of parameters $\boldsymbol{\theta}$ using a set of measurements $\boldsymbol{x}$ described by a probability model $P(\boldsymbol{x}|\boldsymbol{\theta})$. In general the model is not perfect, which is to say it cannot provide an accurate description of the data even at the most optimal point of its parameter space. As a result, the estimated parameters can have a systematic bias.

One can improve the model by including in it additional parameters. That is, $P(\boldsymbol{x}|\boldsymbol{\theta})$ is replaced by a more general model $P(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\nu})$, which depends on parameters of interest $\boldsymbol{\theta}$ and *nuisance parameters* $\boldsymbol{\nu}$. The additional parameters are not of intrinsic interest but must be included for the model to be sufficiently accurate for some point in the enlarged parameter space.

Although including additional parameters may eliminate or at least reduce the effect of systematic uncertainties, their presence will result in increased statistical uncertainties for the parameters of interest. This occurs because the estimators for the nuisance parameters and those of interest will in general be correlated, which results in an enlargement of the contour defined by Eq. (40.13).

To reduce the impact of the nuisance parameters one often tries to constrain their values by means of control or calibration measurements, say, having data $\boldsymbol{y}$. For example, some components of $\boldsymbol{y}$ could represent estimates of the nuisance parameters, often from separate experiments. Suppose the measurements $\boldsymbol{y}$ are statistically independent from $\boldsymbol{x}$ and are described by a model $P(\boldsymbol{y}|\boldsymbol{\nu})$. The joint model for both $\boldsymbol{x}$ and $\boldsymbol{y}$ is in this case therefore the product of the probabilities for $\boldsymbol{x}$ and $\boldsymbol{y}$, and thus the likelihood function for the full set of parameters is

$$L(\boldsymbol{\theta},\boldsymbol{\nu}) = P(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\nu})P(\boldsymbol{y}|\boldsymbol{\nu}) \,. \tag{40.17}$$

Note that in this case if one wants to simulate the experiment by means of Monte Carlo, both the primary and control measurements, $\boldsymbol{x}$ and $\boldsymbol{y}$, must be generated for each repetition under assumption of fixed values for the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$.

Using all of the parameters $(\boldsymbol{\theta},\boldsymbol{\nu})$ in Eq. (40.13) to find the statistical errors in the parameters of interest $\boldsymbol{\theta}$ is equivalent to using the *profile likelihood*, which depends only on $\boldsymbol{\theta}$. It is defined as

$$L_{\mathrm{p}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta},\widehat{\widehat{\boldsymbol{\nu}}}(\boldsymbol{\theta})), \tag{40.18}$$

where the double-hat notation indicates the profiled values of the parameters $\boldsymbol{\nu}$, defined as the values that maximize L for the specified $\boldsymbol{\theta}$. The profile likelihood is discussed further in Section 40.3.2.1 in connection with hypothesis tests.

40.2.3 *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 log-likelihood function contains the sum of squares

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} \,. \tag{40.19}$$

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

The minimum of the chi-square function in Equation (40.19) defines the least-squares estimators $\widehat{\boldsymbol{\theta}}$ for the more general case where the y_i are not Gaussian distributed as long as they are independent. If they are not independent but rather have a covariance matrix $V_{ij} = \mathrm{cov}[y_i, y_j]$, then the LS estimators are determined by the minimum of

$$\chi^2(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \,, \tag{40.20}$$

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. If the y_i are not Gaussian distributed, then the least-squares and maximum-likelihood estimators will not in general coincide.

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

$$\mu(x_i; \boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) \ . \tag{40.21}$$

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,

$$\widehat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y} \ . \tag{40.22}$$

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

$$U = DVD^T = (H^T V^{-1} H)^{-1} \ , \tag{40.23}$$

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

$$(U^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \sum_{k,l=1}^{N} h_i(x_k)(V^{-1})_{kl} h_j(x_l) \ . \tag{40.24}$$

The LS estimators can also be found from the expression

$$\widehat{\boldsymbol{\theta}} = U\boldsymbol{g} \ , \tag{40.25}$$

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

$$g_i = \sum_{j,k=1}^{N} y_j h_i(x_k)(V^{-1})_{jk} \ . \tag{40.26}$$

For the case of uncorrelated y_i, for example, one can use (40.25) with

$$(U^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k) h_j(x_k)}{\sigma_k^2} \ , \tag{40.27}$$

$$g_i = \sum_{k=1}^{N} \frac{y_k h_i(x_k)}{\sigma_k^2} \ . \tag{40.28}$$

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

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\widehat{\boldsymbol{\theta}}) + 1 = \chi^2_{\mathrm{min}} + 1 \tag{40.29}$$

has tangent planes located at plus-or-minus-one standard deviation $\sigma_{\widehat{\boldsymbol{\theta}}}$ from the LS estimates $\widehat{\boldsymbol{\theta}}$ (the relation is approximate if the fit function $\mu(x; \boldsymbol{\theta})$ is nonlinear in the parameters).

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. In this case the least-squares and maximum-likelihood methods are no longer exactly equivalent even for Gaussian distributed measurements. An important example is where the measured value y_i represents the event count in a histogram bin. 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, $\mu(x_i;\boldsymbol{\theta})$, or from the observed number itself, y_i. In the first option, the variances become functions of the parameters, and as a result the estimators may need to be found numerically. The second option can be undefined if y_i is zero, and 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). As noted in Sec. 40.2.2.1, this issue is avoided when using the method of extended maximum likelihood with binned data by minimizing Eq. (40.16). In that case if the expected number of events $\mu_{\rm tot}$ does not depend on the other fitted parameters $\boldsymbol{\theta}$, then its extended MLE is equal to the observed total number of events.

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 40.3.3.1.

40.2.4 *Parameter estimation with constraints*

In some applications one is interested in using a set of measured quantities $\boldsymbol{y} = (y_1, \ldots, y_N)$ to estimate parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ subject to a number of constraints. For example, one may have measured coordinates from two tracks, and the goal is to estimate their momentum vectors subject to the constraint that the tracks have a common vertex. The parameters can also include momenta of undetected particles such as neutrinos, as long as the constraints from conservation of energy and momentum and from known masses of particles involved in the reaction chain provide enough information for these quantities to be inferred.

A set of K constraints can be given in the form of equations

$$c_k(\boldsymbol{\theta}) = 0 \ , \quad k = 1, \ldots, K \ . \tag{40.30}$$

In some problems it may be possible to define a new set of parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)$ with $L = M - K$ such that every point in $\boldsymbol{\eta}$-space automatically satisfies the constraints. If this is possible then the problem reduces to one of estimating $\boldsymbol{\eta}$ with, *e.g.*, maximum likelihood or least squares and then transforming the estimators back into $\boldsymbol{\theta}$-space.

In many cases it may be difficult or impossible to find an appropriate transformation $\boldsymbol{\eta}(\boldsymbol{\theta})$. Suppose that the parameters are determined by minimizing an objective function such as $\chi^2(\boldsymbol{\theta})$ in the method of least squares. Here one may enforce the constraints by finding the stationary points of the *Lagrange function*

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{y}) = \chi^2(\boldsymbol{\theta}, \boldsymbol{y}) + \sum_{k=1}^{K} \lambda_k c_k(\boldsymbol{\theta}) \tag{40.31}$$

with respect to both the parameters $\boldsymbol{\theta}$ and a set of *Lagrange multipliers* $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. Combining the parameters and Lagrange multipliers into an $(M+K)$-component vector $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_M, \lambda_1, \ldots, \lambda_K)$, the solutions for $\boldsymbol{\gamma}$, *i.e.*, the estimators $\hat{\boldsymbol{\gamma}}$, are found (*e.g.*, numerically) from the system of equations

$$F_i(\boldsymbol{\gamma}, \boldsymbol{y}) \equiv \frac{\partial \mathcal{L}}{\partial \gamma_i} = 0 \ , \quad i = 1, \ldots, M + K \ . \tag{40.32}$$

To obtain the covariance matrix of the estimated parameters one can find solutions $\tilde{\boldsymbol{\gamma}}$ corresponding to the expectation values of the data $\langle \boldsymbol{y} \rangle$ and expand $F_i(\hat{\boldsymbol{\gamma}}, \boldsymbol{y})$ to first order about these values.

This gives (see, e.g., Sec. 11.6 of Ref. [8]) linearized approximations for the estimators, $\hat{\boldsymbol{\gamma}}(\boldsymbol{y}) \approx \tilde{\boldsymbol{\gamma}} + C(\boldsymbol{y} - \langle \boldsymbol{y} \rangle)$, where the matrix $C = -A^{-1}B$, and A and B are given by

$$A_{ij} = \left[\frac{\partial F_i}{\partial \gamma_j}\right]_{\tilde{\boldsymbol{\gamma}}, \langle \boldsymbol{y} \rangle} \quad \text{and} \quad B_{ij} = \left[\frac{\partial F_i}{\partial y_j}\right]_{\tilde{\boldsymbol{\gamma}}, \langle \boldsymbol{y} \rangle} \; . \tag{40.33}$$

In practice the values $\langle \boldsymbol{y} \rangle$ and corresponding solutions $\tilde{\boldsymbol{\gamma}}$ are estimated using the data from the actual measurement. Using this approximation for $\hat{\boldsymbol{\gamma}}(\boldsymbol{y})$, one can find the covariance matrix $U_{ij} = \mathrm{cov}[\hat{\gamma}_i, \hat{\gamma}_j]$ of the estimators for the γ_i in terms of that of the data $V_{ij} = \mathrm{cov}[y_i, y_j]$ using error propagation (cf. Eqs. (39.17) and (39.18)),

$$U = CVC^T \; . \tag{40.34}$$

The upper-left $M \times M$ block of the matrix U gives the covariance matrix for the estimated parameters $\mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. One can show for linear constraints that $\mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ is also given by the upper-left $M \times M$ block of $2A^{-1}$. If the parameters are estimated using the method of least squares, then the number of degrees of freedom for the distribution of the minimized χ^2 is increased by the number of constraints, *i.e.*, it becomes $N - M + K$. Further details can be found in, *e.g.*, Ch. 8 of Ref. [4] and Ch. 7 of Ref. [11].

40.2.5 *Unfolding*

An important class of parameter estimation problem involves measurement of the differential distribution of a kinematic variable in the form of a histogram with N bins. The data thus consist of the vector of measured data values $\boldsymbol{n} = (n_1, \ldots n_N)$, with expectation values $\nu_i = E[n_i]$. The n_i are usually independent and often modeled as Poisson distributed, from which it follows that the maximum-likelihood estimators are $\hat{\nu}_i = n_i$ for all i.

Because of the limited acceptance and resolution of the experiment, however, the measured values of the kinematic variable in question differ in general from their true values, and as a consequence the form of the data histogram is distorted relative to what would be obtained with perfect resolution. The desired parameters are not, therefore, the ν_i but rather one wants to estimate the expected number of entries in a given bin that would be found with a perfect detector. We call this the "true histogram" and denote it with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$, where the number of bins M is not required to equal that of the observed histogram. The μ_i are related to the expected numbers of events in the observed histogram by

$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + \beta_i \; , \tag{40.35}$$

where β_i here represents the expected number of events in bin i due to background processes and the $N \times M$ *response matrix* R_{ij} gives the probability for an event to be observed in bin i of $\boldsymbol{\nu}$ given that the true value of the variable was in bin j of $\boldsymbol{\mu}$. For purposes of this discussion let us suppose that the response matrix and expected background values are known. There are two main approaches to this type of problem, which we can call *unfolding* and *folding*.

In unfolding, one treats the true histogram $\boldsymbol{\mu}$ as the parameters of interest. The result is thus given by estimators $\hat{\boldsymbol{\mu}}$ and the corresponding covariance matrix $U_{ij} = \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j]$. Provided the response matrix can be inverted, the maximum-likelihood solution is easily found as $\hat{\boldsymbol{\mu}} = R^{-1}(\boldsymbol{n} - \boldsymbol{\beta})$. If the response matrix allows for significant migration of events between bins, then the variances of these estimators can be very large, sometimes to the point where the $\hat{\boldsymbol{\mu}}$ bear essentially no resemblance to the true $\boldsymbol{\mu}$. In such cases the estimators can be found by maximizing

a linear combination of the log-likelihood and a regularization function that imposes some degree of smoothness on the unfolded distribution. In achieving a reduction in variance one inevitably introduces some bias into the estimators.

In the approach of folding, by contrast, to test a given model prediction for the true histogram $\boldsymbol{\mu}$ it is first "folded" with the response matrix and corrected for expected background to give the corresponding $\boldsymbol{\nu}$ according to Eq. (40.35), and these are compared to the corresponding $\boldsymbol{n}$ using, e.g., a likelihood ratio as shown in Sec. 40.2.2.1. In this way, one avoids any bias due to regularization.

To account for systematic uncertainties, the response matrix and background values may not be expressed as constants but rather as functions of nuisance parameters $\boldsymbol{\theta}$. These may be constrained by auxiliary measurements $\boldsymbol{u}$ with p.d.f. $p(\boldsymbol{u}|\boldsymbol{\theta})$ so that the full likelihood becomes (as in Eq. (40.17) but with different notation)

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\boldsymbol{n}|\boldsymbol{\nu}(\boldsymbol{\mu}, \boldsymbol{\theta}))p(\boldsymbol{u}|\boldsymbol{\theta}) \; . \tag{40.36}$$

Here the mean values $\boldsymbol{\nu} = R(\boldsymbol{\theta})\boldsymbol{\mu} + \boldsymbol{\beta}(\boldsymbol{\theta})$ now depend on both the true histogram parameters $\boldsymbol{\mu}$ as well as the nuisance parameters $\boldsymbol{\theta}$. Thus to record enough information for a future analysis using folding one must provide all of the ingredients used above: the primary data $\boldsymbol{n}$, the auxiliary measurements $\boldsymbol{u}$, and also the response matrix $R(\boldsymbol{\theta})$ and expected backgrounds $\boldsymbol{\beta}(\boldsymbol{\theta})$ as functions of the nuisance parameters. If unfolding is used, then future model tests or comparisons with other experiments can be carried out directly using the estimators $\hat{\boldsymbol{\mu}}$ and their covariance matrix.

If several distributions are unfolded, then to combine these in a test of a given model one should know how estimators for bins of different distributions are correlated, as can arise, e.g., through common systematic effects. If only the unfolded distributions and their separate covariance matrices are reported, however, then information on such correlations is not retained. In folding, one can include information on correlated systematic effects if the ingredients (R and $\boldsymbol{\beta}$) are known in terms of nuisance parameters that are common to different distributions.

In unfolding, the estimators for $\boldsymbol{\mu}$ can be constructed either by maximizing the log-likelihood using Eq. (40.36) or in the case where regularization is required one can maximize $\varphi(\boldsymbol{\mu}, \boldsymbol{\theta}) = \ln L(\boldsymbol{\mu}, \boldsymbol{\theta}) + \tau S(\boldsymbol{\mu})$, where the *regularization function* $S(\boldsymbol{\mu})$ reflects the smoothness of the true histogram and serves to reduce the variances of the estimators. Possible functions are based on the mean squared second derivative (a commonly used type of Tikhonov regularization) or the entropy of the true histogram. The parameter τ fixes the relative weighting of the log-likelihood and the regularization function and thus determines the balance between the bias and variance of the resulting estimators. Further discussion of the unfolding problem including methods for choosing the regularization function and parameter, as well as techniques that employ other types of regularization such as the iterative Bayes (Richardson-Lucy) method, can be found in Refs. [8,11–13] and references therein.

40.2.6 *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. [14–17].

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,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int P(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'} \; , \tag{40.37}$$

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. (40.37) 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, mode or median value and covariance matrix. In addition one may construct intervals with a given probability content, as is discussed in Sec. 40.4.1 on Bayesian interval estimation.

40.2.6.1 *Priors*

Bayesian statistics supplies no unique rule for determining the prior $\pi(\boldsymbol{\theta})$; this reflects the analyst'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 (40.37) that the posterior is proportional to the likelihood, and therefore the mode (peak position) of the posterior is equal to the maximum-likelihood 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 MLE.

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 ones which result in a maximum gain in information for a given set of measurements. For an extensive review see, *e.g.*, Ref. [18].

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

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}\,, \tag{40.38}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln P(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] \tag{40.39}$$

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 does not in general correspond to 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 therefore said to be *improper*. This can be allowed because the prior always appears multiplied by the likelihood

function, and if the likelihood falls to zero 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 [19]. 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 i.i.d. data $\boldsymbol{x} = (x_1, \ldots, x_n)$:

$$D_n[\pi, p] = \int p(\boldsymbol{\theta}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{\pi(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \; . \tag{40.40}$$

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,

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \; . \tag{40.41}$$

For a single, continuous parameter the reference prior is usually identical to the 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 sensitivity analysis. Further discussion and applications to particle physics problems can be found in Ref. [20].

40.2.6.2 *Bayesian treatment of nuisance parameters*

As discussed in Sec. 40.2.2, 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 $p(\boldsymbol{y}|\boldsymbol{\nu})$. 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 (40.1) 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 $\boldsymbol{y}$),

$$\pi(\boldsymbol{\nu}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\nu}) \pi_0(\boldsymbol{\nu}) \; . \tag{40.42}$$

In the absence of a model for $P(\boldsymbol{y}|\boldsymbol{\nu})$ one may make some reasonable but *ad hoc* choices in order to approximate $\pi(\boldsymbol{\nu})$. For a single nuisance parameter ν, for example, one might characterize the uncertainty 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 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.*,

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu}|\boldsymbol{x}) \, d\boldsymbol{\nu} \; . \tag{40.43}$$

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) techniques are often used for computing integrals of this type (see Sec. 42.6).

40.3 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. 40.3.1, provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. In *significance tests*, covered in Sec. 40.3.2, 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. Goodness-of-fit tests that quantify the general level of compatibility between data and a hypothesis are described in Sec. 40.3.3. In the Bayesian approach, the corresponding procedure is based fundamentally on the posterior probabilities of the competing hypotheses. In Sec. 40.3.4 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.

40.3.1 *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 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 data $\boldsymbol{x}$ used to construct a test could be, for example, a set of values that characterizes an individual event. In this case the test corresponds to classification as, *e.g.*, signal or background. Alternatively the data could represent a set of values from a collection of events. Often one is interested in knowing whether all of the events are of a certain type (background), or whether the sample contains at least some events of a new type (signal). Here the background-only hypothesis plays the role of H_0, and in the alternative H_1 both signal and background are present. Rejecting H_0 is, from the standpoint of frequentist statistics, the required step to establish discovery of the signal process.

The critical region is not unique. Its choice 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 likelihood ratio

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|H_1)}{f(\boldsymbol{x}|H_0)} \tag{40.44}$$

is greater than or equal to a given constant c_α, and everywhere outside the critical region one has $\lambda(\boldsymbol{x}) < c_\alpha$, where the value of c_α is determined by the size of the test α. The hypotheses H_0 and H_1 must be simple, *i.e.*, they should not contain undetermined parameters.

It is convenient to define the test using a scalar function of the data $\boldsymbol{x}$ called a *test statistic*, $t(\boldsymbol{x})$, such that the boundary of the critical region is given by a surface of constant $t(\boldsymbol{x})$. The Neyman–Pearson lemma is equivalent to the statement that the likelihood ratio (40.44) represents

the optimal test statistic. 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 $f(\boldsymbol{x}|H_1)$. Often one does not have explicit formulae for these, but rather Monte Carlo models that allow one to generate instances of $\boldsymbol{x}$ that follow the p.d.f.s.

In the case where the likelihood ratio (40.44) cannot be used explicitly, there exist a variety of other multivariate methods for constructing a test statistic that may approach its performance. These are based on machine-learning algorithms that use samples of *training data* corresponding to the hypotheses in question, often generated from Monte Carlo models. Further information on Machine Learning can be found in Sec. 41 of this Review.

The multivariate algorithms designed to classify events into signal and background types also form the basis of tests of the hypothesis that a sample of events consists of background only. Such a test can be constructed using the distributions of the test statistic $t(\boldsymbol{x})$ for event classification obtained from a multivariate algorithm such as a Neural Network output. The distributions $p(t|s)$ and $p(t|b)$ for signal and background events, respectively, are used to construct the likelihood ratio of the signal-plus-background hypothesis relative to that of background only. To the extent that the test statistic $t(\boldsymbol{x})$ approximates the likelihood ratio (or a monotonic function thereof) for individual events given by (40.44), the resulting test of the background-only hypothesis for the event sample will have maximum power with respect to the signal-plus-background alternative (see Ref. [21]).

40.3.2 *Tests of significance, p-values*

A frequentist *significance test* allows one to assign a numerical value, called the p-value, that reflects the level of agreement between a hypothesis H_0 and the observed data $\boldsymbol{x}$. To carry out a significance test one must specify a subset of the data space, whose boundary includes the observed data point $\boldsymbol{x}_{\rm obs}$, that is deemed to have equal or lesser compatibility with H_0 relative to $\boldsymbol{x}_{\rm obs}$. The p-value of H_0 is the probability, assuming data distributed according to H_0, to find $\boldsymbol{x}$ in the region of equal or lesser compatibility.

For continuous data, one can show that the p-value of H_0, here called p_0, follows a uniform distribution for data generated under H_0. Therefore one has $P(p_0 \leq \alpha|H_0) = \alpha$ for any constant α in $[0, 1]$. In this way the region of data space where $p_0 \leq \alpha$ gives a critical region w of a test of size α, as it satisfies $P(\boldsymbol{x} \in w|H_0) \leq \alpha$. Thus rejecting H_0 if $p_0 \leq \alpha$ is equivalent to the hypothesis test of Sec. 40.3.1 above. For the critical region w defined by $p_0 \leq \alpha$ one can find the power $P(\boldsymbol{x} \in w|H_1)$ with respect to an alternative hypothesis H_1, as defined in Sec. 40.3.1. In general the region of low compatibility with H_0 is defined so as to have a high probability for data generated under H_1, so that the distribution of p_0, assuming data $\boldsymbol{x}$ that follows H_1, will be concentrated at low values, and the critical region w derived from $p_0 \leq \alpha$ will have a high power with respect to H_1.

Similar to construction of a test's critical region in Sec. 40.3.1, finding the p-value is generally done by defining a statistic $t(\boldsymbol{x})$ whose value corresponds to better or worse levels of agreement with the hypothesis. The hypothesis being tested, H_0, will determine the p.d.f. $f(t|H_0)$ for the statistic. If, for example, t is defined such that large values correspond to poor agreement with the hypothesis, then the p-value is

$$p = \int_{t_{\rm obs}}^{\infty} f(t|H_0)\, dt\ , \tag{40.45}$$

where $t_{\rm obs} = t(\boldsymbol{x}_{\rm obs})$ is the value of the statistic obtained in the actual experiment.

Note that the p-value is not the probability of the hypothesis, which in frequentist statistics is not defined. The p-value should also not be confused with the size (significance level) of a test, or the confidence level of a confidence interval (Section 40.4), both of which are pre-specified constants.

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 effective significance Z, defined as an equivalent number of standard deviations of a Gaussian distributed random variable. In a search for an intrinsically positive signal, *i.e.*, where only upward fluctuations of the estimated rate appear signal like, this is defined as

$$Z = \Phi^{-1}(1-p) \; . \tag{40.46}$$

Here Φ is the cumulative distribution of the standard Gaussian, and Φ^{-1} is its inverse (quantile) function. In this way, a p-value of $1/2$ gives $Z = 0$, *i.e.*, having an estimated signal rate of zero corresponds to zero significance. If either positive or negative data fluctuations would indicate evidence of the signal, then one defines

$$Z = \Phi^{-1}(1-p/2) \; . \tag{40.47}$$

In this case an estimated signal rate of zero gives $p = 1$ and $Z = 0$.

Often in particle physics 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 40.3.2.2).

40.3.2.1 *Frequentist treatment of nuisance parameters and asymptotic methods for tests*

Suppose one wants to test hypothetical values of parameters $\boldsymbol{\theta}$, but the model also contains nuisance parameters $\boldsymbol{\nu}$. To find a p-value for $\boldsymbol{\theta}$ we can construct a test statistic $q_{\boldsymbol{\theta}}$ such that larger values constitute increasing incompatibility between the data and the hypothesis. Then for an observed value of the statistic $q_{\boldsymbol{\theta},\mathrm{obs}}$, the p-value of $\boldsymbol{\theta}$ is

$$p_{\boldsymbol{\theta}}(\boldsymbol{\nu}) = \int_{q_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(q_{\boldsymbol{\theta}}|\boldsymbol{\theta},\boldsymbol{\nu})\, dq_{\boldsymbol{\theta}} \; , \tag{40.48}$$

which depends in general on the nuisance parameters $\boldsymbol{\nu}$. In the strict frequentist approach, $\boldsymbol{\theta}$ is rejected only if the p-value is less than α for all possible values of the nuisance parameters.

The difficulty described above is effectively solved if we can define the test statistic $q_{\boldsymbol{\theta}}$ in such a way that its distribution $f(q_{\boldsymbol{\theta}}|\boldsymbol{\theta})$ is independent of the nuisance parameters. Although exact independence is only found in special cases, it can be achieved approximately by use of the *profile likelihood ratio*. This is given by the profile likelihood from Eq.(40.18) divided by the value of the likelihood at its maximum, *i.e.*, when evaluated with the maximum-likelihood estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\nu}}$:

$$\lambda_p(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\nu}}}(\boldsymbol{\theta}))}{L(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\nu}})} \; . \tag{40.49}$$

Wilks' theorem [10] states that, providing certain general conditions are satisfied, the distribution of $-2\ln\lambda_{\mathrm{p}}(\boldsymbol{\theta})$, under assumption of $\boldsymbol{\theta}$, approaches a χ^2 distribution in the limit where the data sample is very large, independent of the values of the nuisance parameters $\boldsymbol{\nu}$. Here the number of degrees of freedom is equal to the number of components of $\boldsymbol{\theta}$. More details on use of the profile likelihood are given in Refs. [22,23] and in contributions to the PHYSTAT conferences [24]; explicit formulae for special cases can be found in Ref. [25]. Further discussion on how to incorporate systematic uncertainties into p-values can be found in Ref. [26].

Even with use of the profile likelihood ratio, for a finite data sample the p-value of hypothesized parameters $\boldsymbol{\theta}$ will retain in general some dependence on the nuisance parameters $\boldsymbol{\nu}$. Ideally one would find the the maximum of $p_\theta(\boldsymbol{\nu})$ from Eq. (40.48) explicitly, but that is often impractical. An approximate and computationally feasible technique is to use $p_\theta(\widehat{\widehat{\boldsymbol{\nu}}}(\theta))$, where $\widehat{\widehat{\boldsymbol{\nu}}}(\theta)$ are the profiled values of the nuisance parameters as defined in Section 40.2.2.2. The resulting p-value is correct if the true values of the nuisance parameters are equal to the profiled values used; otherwise it could be either too high or too low. This is discussed further in Section 40.4.2 on confidence intervals.

The methods above based on the profile likelihood ratio are useful in practice because, provided one has a sufficiently large data sample, the distributions of the test statistics are related through Wilks' theorem to the chi-square p.d.f. If the data sample is not large enough to justify use of the asymptotic formulae, then some adjustments can be made to the test statistics so that their distributions are better approximated by the asymptotic ones; these improvements go under the general name of "higher-order asymptotics". Using these methods one may obtain p-values using the asymptotic formulae even for smaller data samples, avoiding costly Monte Carlo simulation to find distributions of test statistics.

Roughly speaking, the chi-square distribution for the profile likelihood ratio relies on having a Gaussian distribution for estimators of the model's parameters. An important type of corrected statistic relies on an improved distribution for the maximum-likelihood estimator due to Barndorff-Nielsen [27] (the p^* approximation). Another class of improved statistic due to Bartlett [28, 29] starts with a statistic t which, asymptotically, should be chi-square distributed with n_d degrees of freedom. A new statistic is defined as $t' = tn_\mathrm{d}/E[t]$, which by construction has mean $E[t'] = n_\mathrm{d}$, as for the chi-square distribution. Further details and applications of these methods are described in Refs. [30, 31].

One may also treat model uncertainties in a Bayesian manner but then use the resulting model in a frequentist test. Suppose the uncertainty in a set of nuisance parameters $\boldsymbol{\nu}$ is characterized by a Bayesian prior p.d.f. $\pi(\boldsymbol{\nu})$. This can be used to construct the marginal (also called the prior predictive) model for the data $\boldsymbol{x}$ and parameters of interest $\boldsymbol{\theta}$,

$$P_\mathrm{m}(\boldsymbol{x}|\boldsymbol{\theta}) = \int P(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\nu})\pi(\boldsymbol{\nu})\, d\boldsymbol{\nu}\ . \tag{40.50}$$

The marginal model does not represent the probability of data that would be generated if one were really to repeat the experiment, as in that case one would assume that the nuisance parameters do not vary. Rather, the marginal model represents a situation in which every repetition of the experiment is carried out with new values of $\boldsymbol{\nu}$, randomly sampled from $\pi(\boldsymbol{\nu})$. It is in effect an average of models each with a given $\boldsymbol{\nu}$, where the average is carried out with respect to the prior p.d.f. $\pi(\boldsymbol{\nu})$.

The marginal model for the data $\boldsymbol{x}$ can be used to determine the distribution of a test statistic Q, which can be written

$$P_\mathrm{m}(Q|\boldsymbol{\theta}) = \int P(Q|\boldsymbol{\theta},\boldsymbol{\nu})\pi(\boldsymbol{\nu})\, d\boldsymbol{\nu}\ . \tag{40.51}$$

In a search for a new signal process, the test statistic can be based on the ratio of likelihoods corresponding to the experiments where signal and background events are both present, L_{s+b}, to that of background only, L_b. Often the likelihoods are evaluated with the profiled values of the nuisance parameters, which may give improved performance. It is important to note, however, that it is through use of the marginal model for the distribution of Q that the uncertainties related to the nuisance parameters are incorporated into the result of the test. Different choices for the test statistic itself only result in variations of the power of the test with respect to different alternatives.

Studies of marginalization versus profiling of nuisance parameters for specific problems, e.g., related to a Poisson counting experiment, can be found in Refs. [32, 33].

40.3.2.2 *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 smaller than 2.87×10^{-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.*, Ref. [25]). However, 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. [34]. Suppose a model contains a nuisance parameter such as the peak position that is only defined under the signal model (there is no peak in the background-only model). Furthermore, suppose a test statistic q_0 is defined using the profile likelihood ratio, so that by Wilk's theorem it would be asymptotically chi-square distributed if the peak position were to be fixed. The asymptotic distribution no longer holds if the peak position is adjustable, however, as this violates the regularity conditions of Wilks' theorem. If the statistic has an observed value u, then an approximation for the global p-value is found to be

$$p_{\text{global}} \approx p_{\text{local}} + \langle N_u \rangle \ , \tag{40.52}$$

where $\langle N_u \rangle$, which is much smaller than one in cases of interest, is the mean number of "upcrossings" (as defined in Ref. [34]) of 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 [35],

$$\langle N_u \rangle \approx \langle N_{u_0} \rangle e^{-(u-u_0)/2} \ . \tag{40.53}$$

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 of the order 10^7 needed if one is dealing with a 5σ effect.

40.3.3 *Goodness of fit*

At times one wants to quantify the level of agreement between the data $\boldsymbol{x}$ and a hypothesis H_0 without explicit reference to alternative hypotheses. In the frequentist approach this is done using a *goodness-of-fit test*. The result is quantified with a p-value, in general obtained from a statistic

$t(\boldsymbol{x})$ as done in Sec. 40.3.2 with a significance test. Here, however, the *goodness-of-fit statistic* is defined using general considerations of what constitutes greater or lesser compatibility between data and hypothesis and not with reference to alternative hypotheses or power. Nevertheless, for a given statistic $t(\boldsymbol{x})$, the power of the resulting test with respect to an alternative can be found, so that types of alternatives to which one is sensitive (*i.e.*, to which the test has high power) can be investigated.

The Neyman-Pearson lemma (see Sec. 40.3.1) states that a test of H_0 that has maximum power with respect to an alternative H_1 should be based on a statistic that is monotonic in the likelihood ratio $P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0)$. Therefore a goodness-of-fit statistic derived without reference to a specific H_1 will not in general attain this maximum power. And even if a given $t(\boldsymbol{x})$ happens to be optimal or nearly so for some alternative H_1, there may be other relevant alternatives for which its power is lower. In practice, goodness-of-fit statistics are defined so as to provide sensitivity to a broad class of alternatives, e.g., those that differ in the location of the bulk of the distribution's probability or the distribution's width or the nature of the tails. It may therefore be useful to use several goodness-of-fit tests that are sensitive to different aspects of the data distribution in question.

The most important application of goodness-of-fit tests is no doubt in conjunction with the method of least squares, which we describe in Sec. 40.3.3.1 below. Other types of tests, particularly those based on the *empirical distribution function* are described in Sec. 40.3.3.2. Further discussion of goodness of fit can be found in Refs. [4,36] and references therein.

40.3.3.1 *Goodness-of-fit with chi-square or likelihood ratio*

An important type of goodness-of-fit test arises when a set of measurements $\boldsymbol{y} = (y_1, \ldots, y_N)$ are used to test the hypothesis that the values are Gaussian distributed with given mean values $E[y_i] = \mu_i$ and variances $V[y_i] = \sigma_i^2$ (or covariance $V_{ij} = \mathrm{cov}[y_i, y_j]$ if the values are correlated). The general procedures also hold when the data are Poisson or multinomially distributed with sufficiently large mean values so that the Gaussian represents an adequate approximation. The Gaussian hypothesis may be simple (all parameters fully specified), or the mean values μ_i may be given as functions of some other parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$, with $M < N$. The (co)variances are treated as fixed.

The test statistic derived from the likelihood ratio has already been introduced in Sec. 40.2.3 in connection with the method of least squares, and is given by the minimum of the function $\chi^2(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, i.e., $t = \chi^2_{\mathrm{min}} = \chi^2(\hat{\boldsymbol{\theta}})$, *e.g.*, in the case of uncorrelated measurements

$$t = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \; . \tag{40.54}$$

For Gaussian distributed $\boldsymbol{y}$, the statistic t follows a chi-square distribution for $N - M$ degrees of freedom. From Eq. (40.54) one sees that greater t corresponds to increasing incompatibility between the measured y_i and predicted μ_i. The p-value of the Gaussian hypothesis is therefore taken to be

$$p = \int_{\chi^2_{\mathrm{min}}}^{\infty} f(t; n_{\mathrm{d}})\, dt \; , \tag{40.55}$$

where $f(t; n_{\mathrm{d}})$ is the χ^2 p.d.f. and $n_{\mathrm{d}} = N - M$ is the appropriate number of degrees of freedom. Values are shown in Fig. 40.1 or obtained from standard computer libraries. If the asymptotic 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.

Often the data represent numbers of events in N bins of a histogram, *i.e.*, $\boldsymbol{n} = (n_1, \ldots, n_N)$. An important case is when $\boldsymbol{n}$ follows a multinomial distribution with $n_{\text{tot}} = \sum_{i=1}^{N} n_i$ total entries and mean values μ_i, or equivalently bin probabilities $p_i = \mu_i / n_{\text{tot}}$. Alternatively one may model the n_i as independent and Poisson distributed with means μ_i. If there are no adjustable parameters in the hypotheses, then the goodness-of-fit can be quantified with *Pearson's* χ^2 *statistic*, defined for multinomial data as

$$t_{\text{M}} = \sum_{i=1}^{N} \frac{(n_i - n_{\text{tot}} p_i)^2}{n_{\text{tot}} p_i} \,, \tag{40.56}$$

or for the independent Poisson data as

$$t_{\text{P}} = \sum_{i=1}^{N} \frac{(n_i - \mu_i)^2}{\mu_i} \,. \tag{40.57}$$

In the limit where the means μ_i are large, the statistics t_{M} and t_{P} are found to follow chi-square distributions for $N-1$ and N degrees of freedom, respectively. Having one fewer degrees of freedom in the multinomial case is related to the fact that n_{tot} is fixed and thus the n_i are correlated.

Alternatively one can use the test statistic based on the likelihood ratio $t = -2 \ln \lambda(\hat{\boldsymbol{\theta}})$ given previously in Eq. (40.16). Here the likelihood ratio λ is chosen to correspond to the multinomial or Poisson model as appropriate, as described in Sec. 40.2.2.1. One finds that the distribution of the likelihood-ratio statistic approaches the asymptotic limit faster than does Pearson's chi-square and thus when using the chi-square p.d.f. to compute a p-value one obtains in general a more accurate result from $-2 \ln \lambda$ (see [9]).

Since the mean of the chi-square distribution is equal to n_{d}, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_{\text{d}}$ (here χ^2 refers to the minimized value). 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. 40.2.

If the minimized χ^2 value indicates a low level of agreement between data and hypothesis, 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 and an increased value of the minimized χ^2. The standard deviations of estimators that one finds from, say, Eq. (40.13) 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.

40.3.3.2 *Goodness-of-fit with the empirical distribution function*

Suppose a measurement yields a sample of independent and identically distributed (i.i.d.) values $\boldsymbol{x} = (x_1, \ldots, x_n)$. Often the values may be summarized by creating a histogram, but this inevitably results in a loss of information, since the position of the x values within each bin is not retained. Particularly with small numbers of observations n one therefore may prefer to keep the values of each of the x_i (*i.e.*, "unbinned"). The sample may be displayed graphically using the *empirical distribution function* (e.d.f.)

$$F_n(x) = \frac{\text{number of elements in sample with value } \leq x}{n} \,. \tag{40.58}$$

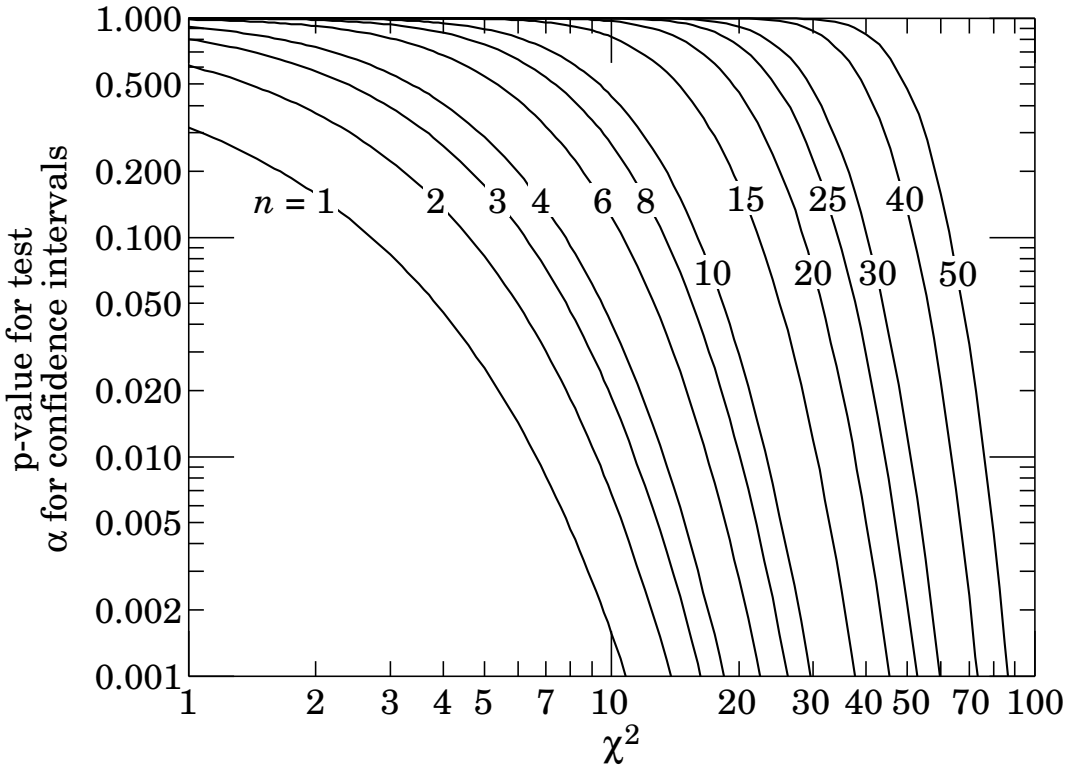

Figure 40.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 40.4.2.2).

The e.d.f. $F_n(x)$ begins at zero for x less than the lowest observed value, increases by $1/n$ at each observed x_i, and saturates at unity for x equal to the highest value in the sample.

Given an i.i.d. data sample $\boldsymbol{x} = (x_1, \ldots, x_n)$, suppose one wants to test the hypothesis H_0 that x follows the p.d.f. $f(x)$ using a goodness-of-fit test, *i.e.*, without reference to a specific alternative. The hypothesis of the p.d.f. $f(x)$ is equivalent to saying the cumulative distribution function (c.d.f.) is $F(x) = \int_{-\infty}^{x} f(x')\,dx'$. As the e.d.f. $F_n(x)$ can be viewed as an estimator for the c.d.f. $F(x)$, we can derive a goodness-of-fit statistic by some appropriately defined measure of the "distance" between the two.

A widely used example is the Kolmogorov-Smirnov (K-S) test, based on

$$D_n = \sup_x |F_n(x) - F(x)| \,, \tag{40.59}$$

where sup denotes the supremum. That is, D_n is the greatest vertical distance between $F_n(x)$ and $F(x)$ for any x. In the limit of large sample size n, one can show that the distribution of D_n, assuming the data follows the c.d.f. $F(x)$, approaches the Kolmogorov distribution, which is independent of the specific form of $F(x)$ and is provided in many software packages. A larger value of D_n corresponds to greater incompatibility between data and hypothesis, so the p-value is the probability $P(D_n \geq D_{n,\text{obs}}|H_0)$. The K-S test is sensitive in particular to differences in location (*i.e.*, in the mean of x), as this corresponds to a horizontal shift between $F_n(x)$ and $F(x)$, giving a large maximum vertical distance D_n.

The Cramér-von Mises family of tests is based on the squared difference between $F_n(x)$ and $F(x)$ through an integral of the form

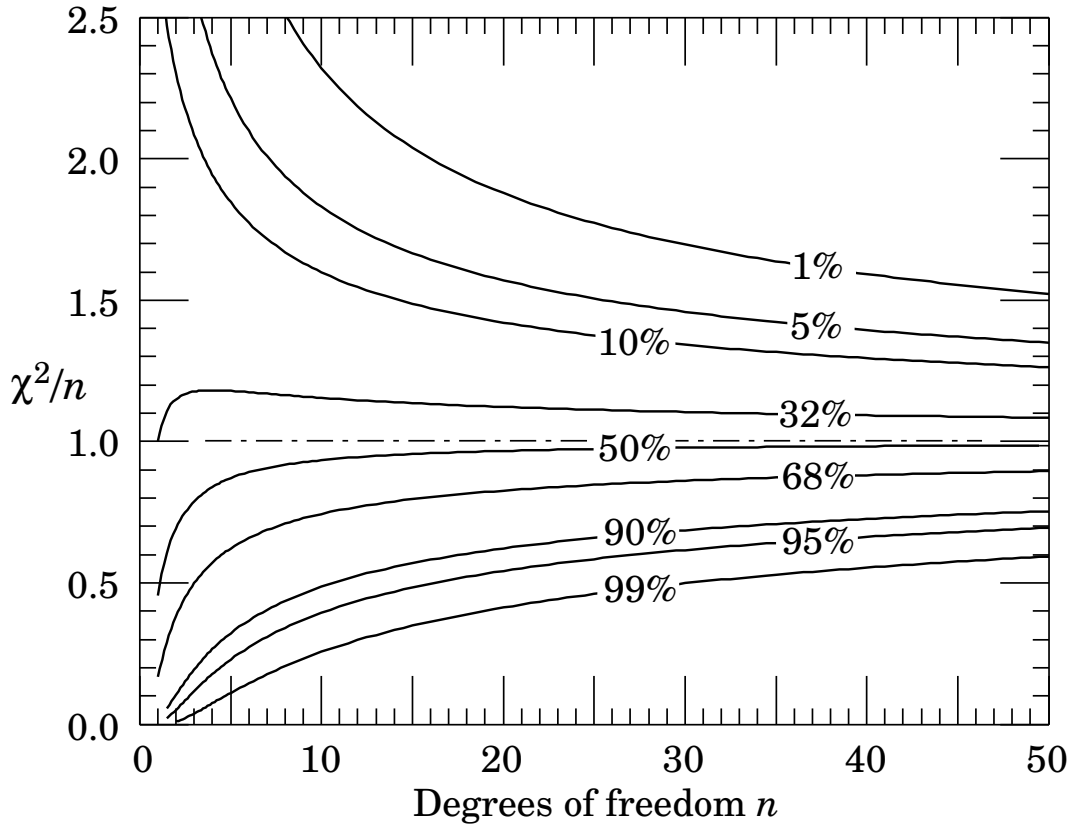

Figure 40.2: The 'reduced' χ^2, equal to χ^2/n, for n degrees of freedom. The curves show as a function of n the χ^2/n that corresponds to a given p-value.

$$W^2 = \int_{-\infty}^{\infty} (F(x) - F_n(x))^2 w(x) f(x)\, dx \ , \tag{40.60}$$

where $f(x) = dF/dx$ is the p.d.f. of x and $w(x)$ is a weight function that must be specified. Two important choices are the classic Cramér-von Mises test with $w(x) = 1$, and the Anderson-Darling test with $w(x) = [F(x)(1 - F(x))]^{-1}$, which gives increased sensitivity to departures in the tails of the distribution. For cases such as these where the weight function depends on x only through the cumulative distribution $F(x)$, one can show (see, *e.g.*, Refs. [4, 36]) that the distribution of W^2, assuming data that follows $F(x)$, is independent of the form of $F(x)$. The distributions of the Smirnov-Cramér-von Mises and Anderson-Darling statistics are available in standard computer libraries.

40.3.4 *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 (Eq. (40.37)). 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

$$\pi(H_i, \boldsymbol{\theta}_i) = P(H_i)\pi(\boldsymbol{\theta}_i|H_i) \; . \tag{40.61}$$

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,

$$P(H_i|\boldsymbol{x}) = \frac{\int P(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)P(H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{P(\boldsymbol{x})} \; , \tag{40.62}$$

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

$$\frac{P(H_i|\boldsymbol{x})}{P(H_j|\boldsymbol{x})} = \frac{\int P(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int P(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \frac{P(H_i)}{P(H_j)} \; . \tag{40.63}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int P(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int P(\boldsymbol{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \; . \tag{40.64}$$

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 favor 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. Improper priors are, however, 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 (40.64) are of the form

$$m = \int P(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \; , \tag{40.65}$$

which is similar to the marginal likelihood seen previously in Eq. (40.50) (in some fields this quantity is called the *evidence*). Computing marginal likelihoods can be difficult; in many cases it can be done with the nested sampling algorithm [37] as implemented, *e.g.*, in the program `MultiNest` [38]. A review of Bayes factors can be found in Ref. [39].

40.4 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.

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 on average 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 small 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.

40.4.1 *Bayesian intervals*

As described in Sec. 40.2.6, 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.*,

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\boldsymbol{x})\, d\theta \; . \tag{40.66}$$

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

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \; . \tag{40.67}$$

This prior may be regarded as providing an interval whose frequentist properties can be studied,

rather than as representing a degree of belief. 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 $s+b$,

$$P(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{40.68}$$

along with the prior (40.67) in (40.37) 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

$$1-\alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} P(n|s)\,\pi(s)\,ds}{\int_{-\infty}^{\infty} P(n|s)\,\pi(s)\,ds} \ , \tag{40.69}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (40.69) to incomplete gamma functions, the solution for the upper limit is found to be

$$s_{\rm up} = \frac{1}{2} F_{\chi^2}^{-1}\left[p, 2(n+1)\right] - b \ , \tag{40.70}$$

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

$$p = 1-\alpha\left(1 - F_{\chi^2}\left[2b, 2(n+1)\right]\right) \ , \tag{40.71}$$

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

$$s_{\rm up} = \frac{1}{2} F_{\chi^2}^{-1}(1-\alpha; 2(n+1)) \ . \tag{40.72}$$

It happens that for the case of $b=0$, the upper limit from Eq. (40.72) coincides numerically with the frequentist upper limit discussed in Section 40.4.2.3. Values for $1-\alpha = 0.9$ and 0.95 are given by the values $\mu_{\rm up}$ in Table 40.3. The frequentist properties of confidence intervals for the Poisson mean found in this way are discussed in Refs. [2] and [40].

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. 40.2.6. 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. (40.43), 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 40.3.4.

40.4.2 *Frequentist confidence intervals*

The unqualified phrase "confidence intervals" refers to frequentist intervals obtained with a procedure due to Neyman [41], described below. The boundary of the interval (or in the multiparameter case, region) is given by a specific function of the data, which would fluctuate if one were to repeat the experiment many times. The *coverage probability* refers to the fraction of intervals in such an ensemble that contain the true parameter value. Confidence intervals are constructed so

as to have a coverage probability greater than or equal to a given *confidence level*, regardless of the true parameter's value. It is important to note that in the frequentist approach, such a probability is not meaningful for a fixed interval. In this section we discuss several techniques for producing intervals that have, at least approximately, this property of coverage.

40.4.2.1 *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 using a pre-defined rule and probability $1-\alpha$ for every value of θ, a set of values $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ such that

$$P(x_1 < x < x_2;\theta) = \int_{x_1}^{x_2} f(x;\theta)\,dx \geq 1-\alpha\ . \tag{40.73}$$

If x is discrete, the integral is replaced by the corresponding sum. In that case there may not exist a range of x values whose summed probability is exactly equal to a given value of $1-\alpha$, and one requires by convention $P(x_1 < x < x_2;\theta) \geq 1-\alpha$.

This is illustrated for continuous x in Fig. 40.3: 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 a *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.

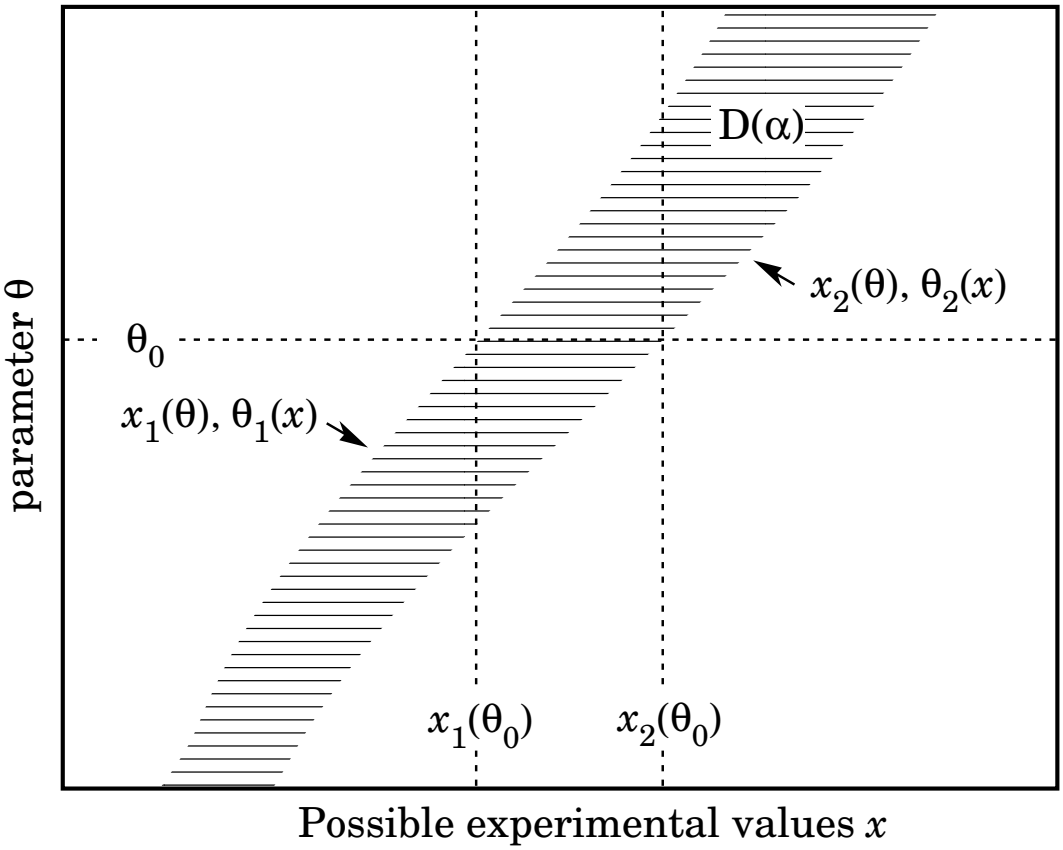

Figure 40.3: Construction of the confidence belt (see text).

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

$$1-\alpha = P(x_1(\theta) < x < x_2(\theta)) = P(\theta_2(x) < \theta < \theta_1(x)) \; . \tag{40.74}$$

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. (40.73) 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 to find x 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 one of $P(x \le x_1)$ or $P(x \ge x_2)$ can be set to α and the other 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. Because of the requirement $P(x_1 < x < x_2) \ge 1-\alpha$ when x is discrete, one obtains in that case 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. 40.3) 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 in a test of size α or less. The remaining values constitute the confidence interval at confidence level $1-\alpha$. If the critical region of the test is characterized by having a p-value $p_\theta \le \alpha$, then the endpoints of the confidence interval are found in practice by solving $p_\theta = \alpha$ for θ.

In the procedure outlined above, 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*,

$$\lambda(\theta) = \frac{f(x;\theta)}{f(x;\widehat{\theta}\,)} \; , \tag{40.75}$$

where $\widehat{\theta}$ is the value of the parameter which, out of all allowed values, maximizes $f(x;\theta)$. The confidence belt is taken to contain the values of x that give the greatest values of $\lambda(\theta)$. This results in the intervals described in Ref. [42] by Feldman and Cousins.

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 40.3.2.1. 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 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. (40.48), 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 particle physics [32] (see also Ref. [26]).

40.4.2.2 *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

$$\begin{aligned} 1-\alpha &= \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\,dx \\ &= \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) = 2\Phi\left(\frac{\delta}{\sigma}\right) - 1\,, \end{aligned} \tag{40.76}$$

where erf is the Gaussian error function, which is rewritten in the final equality using Φ, the Gaussian cumulative distribution. Fig. 40.4 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 40.1.

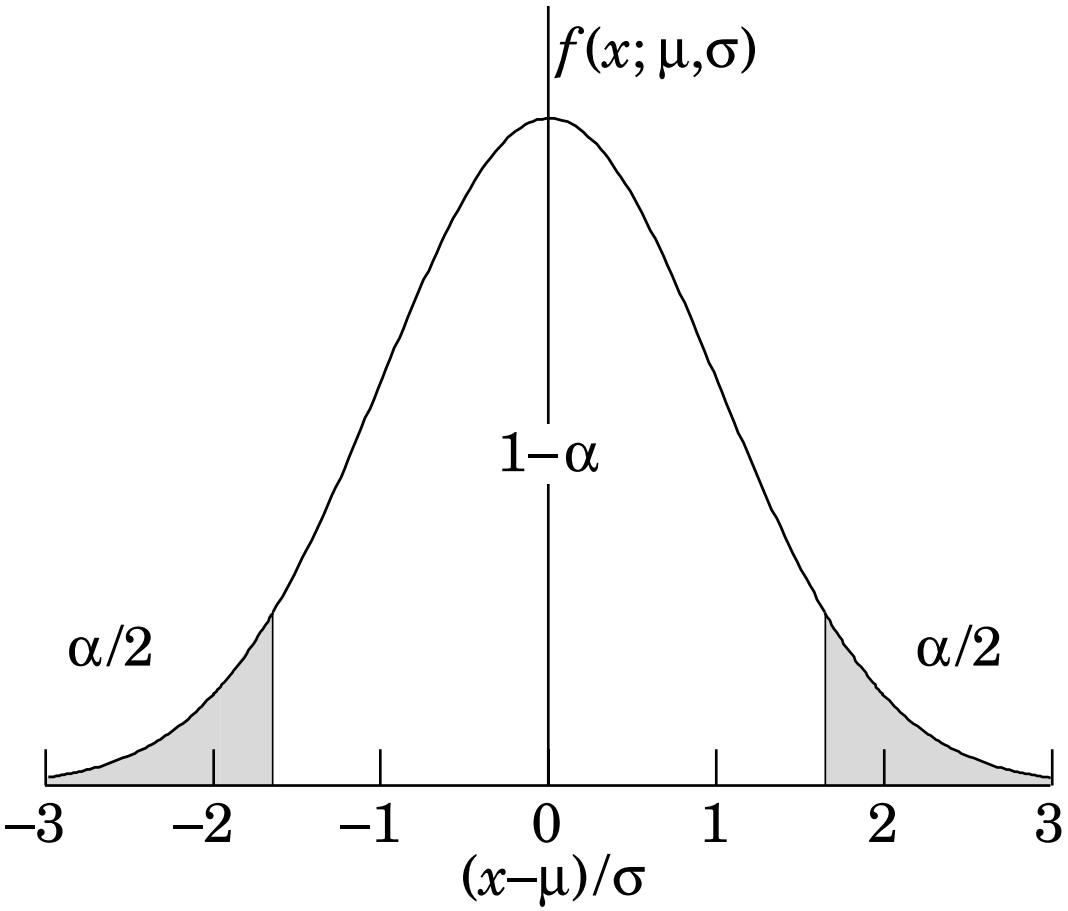

Figure 40.4: Illustration of a symmetric 90% confidence interval (unshaded) for a Gaussian-distributed measurement of a single quantity. Integrated probabilities, defined by $\alpha = 0.1$, are as shown.

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 40.1.

Table 40.1: 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

The relation (40.76) can be re-expressed using the cumulative distribution function for the χ^2 distribution as

$$\alpha = 1 - F(\chi^2; n)\ , \tag{40.77}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom. This can be seen as the $n = 1$ curve in Fig. 40.1 or obtained using standard computer libraries. For multivariate measurements of, say, M parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_M)$, construction of the confidence region requires the full covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$, which can be estimated as described in Sections 40.2.2 and 40.2.3. 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 will take on a Gaussian shape.

The standard error ellipse for the pair $(\widehat{\theta}_i, \widehat{\theta}_j)$ is shown in Fig. 40.5, 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

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2}\ , \tag{40.78}$$

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.

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 σ_{inner} as shown on the figure, where $\sigma_{\mathrm{inner}} = (1 - \rho_{ij}^2)^{1/2}\sigma_i$. 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

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\mathrm{max}} - \Delta \ln L\ , \tag{40.79}$$

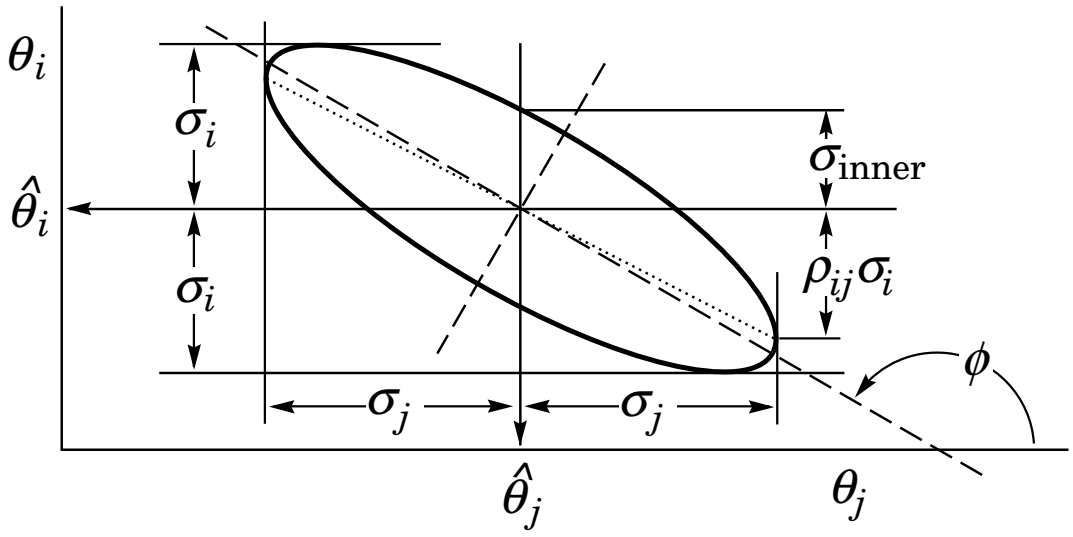

Figure 40.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In the case shown the correlation is negative.

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

$$\chi^2(\boldsymbol{\theta}) \leq \chi^2_{\text{min}} + \Delta\chi^2 \; . \tag{40.80}$$

Values of $\Delta\chi^2$ or $2\Delta \ln L$ are given in Table 40.2 for several values of the coverage probability $1-\alpha$ and number of fitted parameters M. For Gaussian distributed data, these are related by $\Delta\chi^2 = 2\Delta \ln L = F^{-1}_{\chi^2_M}(1-\alpha)$, where $F^{-1}_{\chi^2_M}$ is the chi-square quantile (inverse of the cumulative distribution) for M degrees of freedom.

Table 40.2: 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

For non-Gaussian data samples, the probability for the regions determined by Equations (40.79) or (40.80) 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. Nevertheless, they can still have a coverage probability only weakly dependent on the true parameter, and approximately as given in Table 40.2. 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,

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln L_i(\boldsymbol{\theta}) \ . \tag{40.81}$$

This can then be used to determine an approximate confidence interval or region with Eq. (40.79), just as with a single experiment.

40.4.2.3 *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 and selection efficiency ε in a fixed integrated luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\mu = \sigma\varepsilon\mathcal{L}$. 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, limits on the mean value μ can be found from the Neyman procedure as discussed in Section 40.4.2.1 with n used directly as the statistic x . The upper and lower limits are found to be

$$\mu_{\rm lo} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha_{\rm lo}; 2n) \ , \tag{40.82a}$$

$$\mu_{\rm up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{\rm up}; 2(n+1)) \ , \tag{40.82b}$$

where confidence levels of $1-\alpha_{\rm lo}$ and $1-\alpha_{\rm up}$ refer separately to the corresponding intervals $\mu \geq \mu_{\rm lo}$ and $\mu \leq \mu_{\rm up}$, and $F^{-1}_{\chi^2}$ is the quantile of the χ^2 distribution (inverse of the cumulative distribution). For central confidence intervals at confidence level $1-\alpha$, set $\alpha_{\rm lo} = \alpha_{\rm up} = \alpha/2$.

Table 40.3: 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

It happens that the upper limit from Eq. (40.82b) 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 40.3. 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

$$p_{\rm lo} = \frac{nF_F^{-1}[\alpha_{\rm lo}; 2n, 2(N-n+1)]}{N-n+1 \, + \, nF_F^{-1}[\alpha_{\rm lo}; 2n, 2(N-n+1)]} \, , \tag{40.83a}$$

$$p_{\rm up} = \frac{(n+1)F_F^{-1}[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]}{(N-n) \, + \, (n+1)F_F^{-1}[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]} \, . \tag{40.83b}$$

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

40.4.2.4 *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 all μ. Those values not rejected in a test of size α, *i.e.*, for which one does not find $p_\mu \leq \alpha$, 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.*, in a search for a new particle 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 μ 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, as discussed in Sec. 40.5.

One way to mitigate the problem of excluding models to which one is not sensitive is the $\rm CL_s$ method, where the measure used to test a parameter is increased for decreasing sensitivity [43, 44]. The procedure is based on a statistic called $\rm CL_s$, which is defined as

$$\rm CL_s = \frac{p_\mu}{1-p_b} \, , \tag{40.84}$$

where $p_{\rm b}$ is the p-value of the background-only hypothesis. In the usual formulation of the method, both p_μ and $p_{\rm b}$ are defined using a single test statistic, and the definition of $\rm CL_s$ above assumes this statistic is continuous; more details can be found in Refs. [43, 44].

A point in a model's parameter space is regarded as excluded if one finds $\rm CL_s \leq \alpha$. As the denominator in Eq. (40.84) is always less than or equal to unity, the exclusion criterion based on $\rm CL_s$ is more stringent than the usual requirement $p_\mu \leq \alpha$. In this sense the $\rm CL_s$ 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 $1 - p_b$, and thus the condition $CL_s \leq \alpha$ is effectively prevented from being satisfied. In this way the exclusion of parameters in the case of low sensitivity is suppressed.

The 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 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 this 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 in Section 40.4.2.1. The critical region for the test of a given value of μ contains data values characteristic of both higher and lower rates. As a result, for a given observed rate one can in general obtain a two-sided interval. If, however, the parameter estimate $\hat{\mu}$ is sufficiently close to the lower limit of zero, then only high values of μ are rejected, and the lower edge of the confidence interval is at zero. Note, however, that the coverage property of $1 - \alpha$ pertains to the entire interval, not to the probability for the upper edge μ_{up} to be greater than the true value μ. For parameter estimates increasingly far away from the boundary, *i.e.*, for increasing signal significance, the point $\mu = 0$ is excluded and the interval has nonzero upper and lower edges.

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. (40.73). 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 [42], 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 Ref. [42], 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*).

The intervals according to this method for the mean of Poisson variable in the absence of background are given in Table 40.4. (Note that α in Ref. [42] is defined following Neyman [41] 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 $[\mu_1, \mu_2]$. In Table 40.3 for the one-sided upper limit, however, $1 - \alpha$ refers to the probability to have $\mu_{up} \geq \mu$ (or $\mu_{lo} \leq \mu$ 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

Table 40.4: 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

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 $s = 0$ 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).

Another possibility is to construct a Bayesian interval as described in Section 40.4.1. 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. 40.2.6.

In particle physics a widely used prior for the mean μ of a Poisson distributed measurement has been the uniform distribution 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 values is zero. The procedure above can be more appropriately regarded as a way 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 uniform priors undercover, however, for some values of the parameter, although to an extent that in practical cases may not be too severe [2, 40].

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.

40.5 Experimental sensitivity

In this section we describe methods for characterizing the sensitivity of a search for a new physics signal. As discussed in Sec. 40.3, an experimental analysis can often be formulated as a test of hypothetical model parameters. Therefore we may quantify the sensitivity by giving the results

that we expect from such a test under specific assumptions about the signal process.

Here to be concrete we will consider a parameter μ proportional to the rate of a signal process, although the concepts described in this section may be easily generalized to other parameters. One may wish to establish discovery of the signal process by testing and rejecting the hypothesis that $\mu = 0$, and in addition one often wants to test nonzero values of μ to construct a confidence interval (*e.g.*, limits) as described in Sec. 40.4. In the frequentist framework, the result of each tested value of μ is the p-value p_μ or equivalently the significance $Z_\mu = \Phi^{-1}(1 - p_\mu)$, where as usual Φ is the standard Gaussian cumulative distribution and its inverse Φ^{-1} is the standard Gaussian quantile.

Prior to carrying out the experiment, one generally wants to quantify what significance Z_μ is expected under given assumptions for the presence or absence of the signal process. Specifically, for the significance of a test of $\mu = 0$ (the discovery significance) one usually quotes the Z_0 one would expect if the signal is present at a given nominal rate, which we can define in general to correspond to $\mu = 1$. For limits, one often gives the expected limit under assumption of the background-only ($\mu = 0$) model. These quantities are used to optimize the analysis and to quantify the experimental sensitivity, that is, to characterize how likely it is to make a discovery if the signal is present, and to say what values of μ one may be able to exclude if the signal is in fact absent.

First we clarify the notion of *expected significance*. Because the significance Z_μ is a function of the data, it is itself a random quantity characterized by a certain sampling distribution. This distribution depends on the assumed value of μ, which is not necessarily the same as the hypothesized value of μ being tested. We may therefore consider the distribution $f(Z_\mu|\mu')$, *i.e.*, the distribution of Z_μ that would be obtained by considering data samples generated under assumption of μ'. In a similar way one can talk about the sampling distribution of an upper limit for μ, $f(\mu_{\rm up}|\mu')$.

One can identify the expected significance or limit with either the mean or median of these distributions, but the median may be preferred since it is invariant under monotonic transformations. For example, the monotonic relation between p-value and significance, $p = 1 - \Phi(Z)$, then gives $\mathrm{med}[p_\mu|\mu'] = 1 - \Phi(\mathrm{med}[Z_\mu|\mu'])$, whereas the corresponding relation does not hold in general for the mean.

In some cases one may be able to write down approximate formulae for the distributions of Z_μ and for limits, but more generally they must be determined from Monte Carlo calculations. In many cases of interest, the significance Z_μ and the limits on μ will have approximate Gaussian distributions.

As an example, consider a Poisson counting experiment, where the result consists of an observed number n of events, modeled as a Poisson distributed variable with a mean of $\mu s + b$. Here s and b, the expected numbers of events from signal and background processes, are taken to be known. If we are interested in discovering the signal process we test and try to reject the hypothesis $\mu = 0$. To characterize the experimental sensitivity, we want to give the discovery significance expected under the assumption of $\mu = 1$.

In the limit where its mean value is large, the Poisson variable n can be approximated as an almost continuous Gaussian variable with mean $\mu s + b$ and standard deviation $\sigma = \sqrt{\mu s + b}$. In the usual case where a physical signal model corresponds to $\mu > 0$, the p-value of $\mu = 0$ is the probability to find n greater than or equal to the value observed,

$$p_0 = \Phi\left(\frac{n-b}{\sqrt{b}}\right) , \tag{40.85}$$

and the corresponding significance is $Z_0 = \Phi^{-1}(1 - p_0) = (n - b)/\sqrt{b}$. The median (here equal to the mean) of n assuming $\mu = 1$ is $s + b$, and therefore the median discovery significance is

$$\mathrm{med}[Z_0|\mu = 1] = \frac{s}{\sqrt{b}} . \tag{40.86}$$

The figure of merit "$s/\sqrt{b}$" has been widely used in particle physics as a measure of expected discovery significance. A better approximation for the Poisson counting experiment, however, may be obtained by testing $\mu = 0$ using the likelihood ratio (40.49) $\lambda(0) = L(0)/L(\hat{\mu})$, where

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \tag{40.87}$$

is the likelihood function, and $\hat{\mu} = (n - b)/s$ is the estimator of μ. In this example there are no nuisance parameters, as s and b are taken to be known. For the case where the relevant signal models correspond to positive μ, one may test the $\mu = 0$ hypothesis with the statistic $q_0 = -2 \ln \lambda(0)$ when $\hat{\mu} > 0$, *i.e.*, an excess is observed, and $q_0 = 0$ otherwise. One can show (see, *e.g.*, Ref. [25]) that in the large-sample limit, the discovery significance is then $Z_0 = \sqrt{q_0}$, for which one finds

$$Z_0 = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \tag{40.88}$$

for $n > b$ and $Z_0 = 0$ otherwise. To approximate the expected discovery significance assuming $\mu = 1$, one may simply replace n with the expected value $E[n|\mu = 1] = s + b$ (the so-called "Asimov data set"), giving

$$\text{med}[Z_0|\mu = 1] = \sqrt{2 \left((s + b) \ln \left(1 + \frac{s}{b} \right) - s \right)} \, . \tag{40.89}$$

This has been shown in Ref. [25] to provide a good approximation to the median discovery significance for values of s of several and for b well below unity. The right-hand side of Eq. (40.89) reduces to $s/\sqrt{b}$ in the limit $s \ll b$.

Beyond the simple Poisson counting experiment, in general one may test values of a parameter μ with more complicated functions of the measured data to obtain a p-value p_μ, and from this one can quote the equivalent significance Z_μ or find, *e.g.*, an upper limit μ_{up}. In this case as well one may quantify the experimental sensitivity by giving the significance Z_μ expected if the data are generated with a different value of the parameter μ'. In some problems, finding the sampling distribution of the significance or limits may be possible using large-sample formulae as described, *e.g.*, in Ref. [25]. In other cases a Monte Carlo study may be needed. Using whatever method of calculation is most appropriate, one usually quotes the expected (mean or, preferably, median) discovery significance or exclusion limit as the primary measure of experimental sensitivity.

Even if the true signal is present at its nominal rate, the actual discovery significance Z_0 obtained from the real data is subject to statistical fluctuations and will not in general be equal to its expected value. In an analogous way, the observed limit will differ from the expected limit even if the signal is absent. Upon observing such a difference one would like to know how large this is compared to expected statistical fluctuations. Therefore, in addition to the observed significance and limits it is useful to communicate not only their expected values but also a measure of the width of their distributions.

As the distributions of significance and limits are often well approximated by a Gaussian, one may indicate the intervals corresponding to plus-or-minus one and/or two standard deviations. If the distributions are significantly non-Gaussian, one may use instead the quantiles that give the same probability content, *i.e.*, $[0.1587, 0.8413]$ for $\pm 1\sigma$, $[0.02275, 0.97725]$ for $\pm 2\sigma$. 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 procedures described above pertain to frequentist hypothesis tests and limits. Bayesian limits, just like those found from a frequentist procedure, are functions of the data and one may

therefore find, usually with approximations or Monte Carlo studies, their sampling distribution and corresponding mean (or, preferably, median) and standard deviation.

When trying to establish discovery of a signal process, the Bayesian approach may employ a Bayes factor as described in Sec. 40.3.4. In the case of the Poisson counting experiment with the likelihood from Eq. (40.87), the log of the Bayes factor that compares $\mu = 1$ to $\mu = 0$ is $\ln B_{10} = \ln(L(1)/L(0)) = n \ln(1 + s/b) - s$. That is, the expectation value, assuming $\mu = 1$, of $\ln B_{10}$ for this problem is

$$E[\ln B_{10}|\mu = 1] = (s + b) \ln \left(1 + \frac{s}{b}\right) - s \, . \tag{40.90}$$

Comparing this to Eq. (40.89), one finds $\text{med}[Z_0|1] = \sqrt{2E[\ln B_{10}|1]}$. Thus for this particular problem the frequentist median discovery significance can be related to the corresponding Bayes factor in a simple way.

In some analyses, the goal may not be to establish discovery of a signal process but rather to measure, as accurately as possible, the signal rate. If we consider again the Poisson counting experiment described by the likelihood function of Eq. (40.87), the maximum-likelihood estimator $\hat{\mu} = (n - b)/s$ has a variance, assuming $\mu = 1$, of

$$V[\hat{\mu}] = V\left[\frac{n - b}{s}\right] = \frac{1}{s^2} V[n] = \frac{s + b}{s^2} \, , \tag{40.91}$$

so that the standard deviation of $\hat{\mu}$ is $\sigma_{\hat{\mu}} = \sqrt{s + b}/s$. One may therefore use $s/\sqrt{s + b}$ as a figure of merit to be maximized in order to obtain the best measurement accuracy of a rate parameter. The quantity $s/\sqrt{s + b}$ is also the expected significance with which one rejects s assuming the signal is absent, and thus can be used to optimize the expected upper limit on s.

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. James, *Statistical methods in experimental physics* (2006), ISBN 9789812567956.

[5] P. R. Rider, Journal of the American Statistical Association **55**, 289, 148 (1960).

[6] L. Lyons, *Statistics for Nuclear and Particle Physicists* (1986), ISBN 9780521379342.

[7] R. J. Barlow, Nucl. Instrum. Meth. **A297**, 496 (1990).

[8] G. Cowan, *Statistical data analysis* (1998), ISBN 9780198501565.

[9] S. Baker and R. D. Cousins, Nucl. Instrum. Meth. **221**, 437 (1984).

[10] S. S. Wilks, Annals Math. Statist. **9**, 1, 60 (1938).

[11] O. Behnke *et al.*, editors, *Data analysis in high energy physics*, Wiley-VCH, Weinheim, Germany (2013), ISBN 9783527410583, 9783527653447, 9783527653430, URL http://www.wiley-vch.de/publish/dt/books/ISBN3-527-41058-9.

[12] S. Schmitt, EPJ Web of Conferences **137**, 11008 (2017), ISSN 2100-014X, URL http://dx.doi.org/10.1051/epjconf/201713711008.

[13] L. Brenner *et al.*, Int. J. Mod. Phys. A **35**, 24, 2050145 (2020), [arXiv:1910.14654].

[14] A. O'Hagan and J.J. Forster, *Bayesian Inference*, (2nd edition, volume 2B of *Kendall's Advanced Theory of Statistics*, Arnold, London, 2004).

[15] D. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, (Oxford University Press, 2006).

[16] P.C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, (Cambridge University Press, 2005).

[17] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, (Wiley, 2000).

[18] Robert E. Kass and Larry Wasserman, J. Am. Stat. Assoc. **91**, 1343 (1996).

[19] 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.

[20] L. Demortier, S. Jain and H. B. Prosper, Phys. Rev. **D82**, 034002 (2010), [arXiv:1002.1111].

[21] K. Cranmer, J. Pavez and G. Louppe (2015), [arXiv:1506.02169].

[22] 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.

[23] W. A. Rolke, A. M. Lopez and J. Conrad, Nucl. Instrum. Meth. **A551**, 493 (2005), [arXiv:physics/0403059].

[24] Links to the *Proceedings of the PHYSTAT* conference series (Durham 2002, Stanford 2003, Oxford 2005, and Geneva 2007, 2011) can be found at `https://espace.cern.ch/phystat`.

[25] G. Cowan *et al.*, Eur. Phys. J. **C71**, 1554 (2011), [arXiv:1007.1727]; G. Cowan *et al.*, Eur. Phys. J. **C73**, 1434 (2013).

[26] L. Demortier, *P-Values and Nuisance Parameters, Proceedings of PHYSTAT 2007*, CERN-2008-001, p. 23.

[27] O. Barndorff-Nielsen, Biometrika **67**, 2, 293 (1980), ISSN 00063444, URL `http://www.jstor.org/stable/2335474`.

[28] M. S. Bartlett, Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences **160**, 901, 268 (1937), URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1937.0109`.

[29] D. N. Lawley, Biometrika **43**, 3-4, 295 (1956), ISSN 0006-3444, URL `https://doi.org/10.1093/biomet/43.3-4.295`.

[30] A. R. Brazzale, A. C. Davison and N. Reid, *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge University Press, Cambridge (2007), URL `http://infoscience.epfl.ch/record/104219`.

[31] G. Cordeiro and F. Cribari-Neto, *An Introduction to Bartlett Correction and Bias Reduction*, SpringerBriefs in Statistics, Springer Berlin Heidelberg (2014), ISBN 9783642552557, URL `https://books.google.co.uk/books?id=MJyKAwAAQBAJ`.

[32] K. Cranmer, in "Statistical Problems in Particle Physics, Astrophysics and Cosmology (PHYSTAT 05): Proceedings, Oxford, UK, September 12-15, 2005," 112–123 (2005), [arXiv:physics/0511028].

[33] R. D. Cousins, J. T. Linnemann and J. Tucker, Nucl. Instrum. Meth. A **595**, 2, 480 (2008), URL `https://doi.org/10.1016%2Fj.nima.2008.07.086`.

[34] E. Gross and O. Vitells, Eur. Phys. J. **C70**, 525 (2010), [arXiv:1005.1891].

[35] R. B. Davies, Biometrika **74**, 33 (1987).

[36] R. D. Cousins, Lectures on Statistics in Theory: Prelude to Statistics in Practice (2023), [arXiv:1807.05996].

[37] J. Skilling, *Nested Sampling*, *AIP Conference Proceedings*, **735**, 395–405 (2004).

[38] F. Feroz, M. P. Hobson and M. Bridges, Mon. Not. Roy. Astron. Soc. **398**, 1601 (2009), [arXiv:0809.3437].

[39] R. E. Kass and A. E. Raftery, J. Am. Statist. Assoc. **90**, 430, 773 (1995).

[40] B. P. Roe and M. B. Woodroofe, Phys. Rev. **D63**, 013009 (2001), [hep-ex/0007048].

[41] J. Neyman, Phil. Trans. Roy. Soc. Lond. **A236**, 767, 333 (1937); Reprinted in *A Selection of Early Statistical Papers on J. Neyman*, (University of California Press, Berkeley, 1967).

[42] G. J. Feldman and R. D. Cousins, Phys. Rev. **D57**, 3873 (1998), [arXiv:physics/9711021]; 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.

[43] A. L. Read, in "Workshop on confidence limits, CERN, Geneva, Switzerland, 17-18 Jan 2000: Proceedings," 81–101 (2000), URL `http://weblib.cern.ch/record/451614`.

[44] T. Junk, Nucl. Instrum. Meth. **A434**, 435 (1999), [hep-ex/9902006].