# Statistics

Glen Cowan

Royal Holloway, University of London, Department of Physics, Egham, Surrey TW20 0EX, U.K., g.cowan@rhul.ac.uk

## Table of Contents

Abstract . . . . . 1
1 Introduction . . . . . 2
2 Probability . . . . . 2
3 Random variables . . . . . 3
4 Parameter estimation . . . . . 7
4.1 Estimators for mean, variance and median . . . . . 8
4.2 The method of maximum likelihood . . . . . 8
4.3 The method of least squares . . . . . 10
4.4 The Bayesian approach . . . . . 12
5 Statistical tests . . . . . 14
5.1 Hypothesis tests . . . . . 15
5.2 Significance tests . . . . . 16
5.3 Bayesian model selection . . . . . 18
6 Intervals and limits . . . . . 19
6.1 Bayesian intervals . . . . . 20
6.2 Frequentist confidence intervals . . . . . 21
7 Conclusions . . . . . 26
References . . . . . 26

## Abstract

In experimental particle physics as well as in many other fields it has become increasingly important to analyze data in a manner that extracts the maximum information and takes into account all of the known uncertainties. This article provides an overview of the most important statistical methods used to carry out this task. It begins with an overview of probability, as this forms the basis for quantifying uncertainty. The statistical methods considered include

the general framework of statistical tests and parameter estimation, including methods for constructing intervals such as upper limits. Both frequentist and Bayesian approaches are described.[1]

## 1 Introduction

This chapter gives an overview of statistical methods used in high-energy physics. In statistics, we are interested in using a 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. These differ in their interpretation of probability. A review of probability and random variables is given in Sections 2 and 3.

The most important statistical tools within the frequentist framework are parameter estimation, covered in Section 4, and statistical tests, discussed in Section 5. Frequentist confidence intervals, which are constructed so as to cover the true value of a parameter with a specified probability, are treated in Section 6.2. Bayesian methods for interval estimation are discussed in Sections 6.1. These intervals quantify the degree of belief with which a parameter lies within a stated range. Intervals are discussed for the important cases of Gaussian, binomial and Poisson distributed data.

## 2 Probability

An abstract definition of *probability* can be given by considering a set $S$, called the sample space, and possible subsets $A, B, \ldots$, the interpretation of which is left open. The probability $P$ is a real-valued function defined by the following axioms due to Kolmogorov (Kolmogorov 1933):

1. For every subset $A$ in $S$, $P(A) \geq 0$;
2. For disjoint subsets (i.e., $A \cap B = \emptyset$), $P(A \cup B) = P(A) + P(B)$;
3. $P(S) = 1$.

In addition, one defines the conditional probability $P(A|B)$ (read $P$ of $A$ given $B$) as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \,. \tag{1}$$

From this definition and using the fact that $A \cap B$ and $B \cap A$ are the same, one obtains *Bayes' theorem*,

[1] This article is based largely on the reviews of Probability and Statistics contained in the *Review of Particle Physics* by the Particle Data Group (Amsler et al. 2008).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ . \tag{2}$$

From the three axioms of probability and the definition of conditional probability, one obtains the *law of total probability*,

$$P(B) = \sum_i P(B|A_i)P(A_i) \ , \tag{3}$$

for any subset $B$ and for disjoint $A_i$ with $\cup_i A_i = S$. This can be combined with Bayes' theorem Eq. (2) to give

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \ , \tag{4}$$

where the subset $A$ could, for example, be one of the $A_i$.

The most commonly used interpretation of the elements of the sample space are outcomes of a repeatable experiment. The probability $P(A)$ is assigned a value equal to the limiting frequency of occurrence of $A$. This interpretation forms the basis of *frequentist statistics*.

The elements of the sample space can also be interpreted as *hypotheses*, i.e., statements that are either true or false, such as 'The mass of the $W$ boson lies between 80.3 and 80.5 GeV.' In the frequency interpretation, such statements are either always or never true, i.e., the corresponding probabilities would be 0 or 1. Using *subjective probability*, however, $P(A)$ is interpreted as the degree of belief that the hypothesis $A$ is true. Subjective probability is used in *Bayesian* (as opposed to frequentist) statistics. Bayes' theorem can be written

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory}) \ , \tag{5}$$

where 'theory' represents some hypothesis and 'data' is the outcome of the experiment. Here $P(\text{theory})$ is the *prior* probability for the theory, which reflects the experimenter's degree of belief before carrying out the measurement, and $P(\text{data}|\text{theory})$ is the probability to have gotten the data actually obtained, given the theory, which is also called the *likelihood*.

Bayesian statistics provides no fundamental rule for obtaining the prior probability; this is necessarily subjective and may depend on previous measurements, theoretical prejudices, etc. Once this has been specified, however, Eq. (5) tells how the probability for the theory must be modified in the light of the new data to give the *posterior* probability, $P(\text{theory}|\text{data})$. As Eq. (5) is stated as a proportionality, the probability must be normalized by summing (or integrating) over all possible hypotheses.

## 3 Random variables

A *random variable* is a numerical characteristic assigned to an element of the sample space. In the frequency interpretation of probability, it corresponds

to an outcome of a repeatable experiment. Let $x$ be a possible outcome of an observation. If $x$ can take on any value from a continuous range, we write $f(x;\theta)dx$ as the probability that the measurement's outcome lies between $x$ and $x + dx$. The function $f(x;\theta)$ is called the *probability density function* (p.d.f.), which may depend on one or more parameters $\theta$. If $x$ can take on only discrete values (e.g., the non-negative integers), then $f(x;\theta)$ is itself a probability. The p.d.f. is always normalized to unit area (unit sum, if discrete). Both $x$ and $\theta$ may have multiple components and are then often written as vectors.

The *cumulative distribution function* $F(x)$ is the probability to for the random variable to be observed less than $x$:

$$F(x) = \int_{-\infty}^{x} f(x')\,dx' \ . \tag{6}$$

Here and below, if $x$ is discrete-valued, the integral is replaced by a sum. The endpoint $x$ is expressly included in the integral or sum.

Any function of random variables is itself a random variable, with (in general) a different p.d.f. The *expectation value* of any function $u(x)$ is

$$E[u(x)] = \int_{-\infty}^{\infty} u(x)\,f(x)\,dx \ , \tag{7}$$

assuming the integral is finite. If $u(x)$ and $v(x)$, are any two functions of $x$, then $E[u+v] = E[u]+E[v]$. For constant values $c$ and $k$ one finds $E[cu+k] = cE[u] + k$.

The $n^{th}$ moment of a random variable is

$$\alpha_n \equiv E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\,dx \ , \tag{8}$$

and the $n^{th}$ central moment of $x$ (or moment about the mean, $\alpha_1$) is

$$m_n \equiv E[(x-\alpha_1)^n] = \int_{-\infty}^{\infty} (x-\alpha_1)^n f(x)\,dx \ . \tag{9}$$

The most commonly used moments are the mean $\mu$ and variance $\sigma^2$:

$$\mu \equiv \alpha_1 \ , \tag{10}$$

$$\sigma^2 \equiv V[x] \equiv m_2 = \alpha_2 - \mu^2 \ . \tag{11}$$

The mean is the location of the "centre of mass" of the p.d.f., and the variance is a measure of the square of its width. Note that $V[cx + k] = c^2V[x]$. It is often convenient to use the *standard deviation* of $x$, $\sigma$, defined as the square root of the variance.

Besides the mean, another useful indicator of the "middle" of the probability distribution is the *median*, $x_{\rm med}$, defined by $F(x_{\rm med}) = 1/2$, i.e., half

the probability lies above and half lies below $x_{\rm med}$. (More rigorously, $x_{\rm med}$ is a median if $P(x \geq x_{\rm med}) \geq 1/2$ and $P(x \leq x_{\rm med}) \geq 1/2$. If only one value exists, it is called '*the* median.')

Let $x$ and $y$ be two random variables with a *joint* p.d.f. $f(x,y)$. The *marginal* p.d.f. of $x$ (the distribution of $x$ with $y$ unobserved) is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \ , \tag{12}$$

and similarly for the marginal p.d.f. $f_2(y)$. The *conditional* p.d.f. of $y$ given fixed $x$ (with $f_1(x) \neq 0$) is defined by $f_3(y|x) = f(x,y)/f_1(x)$, and similarly $f_4(x|y) = f(x,y)/f_2(y)$. From these, we immediately obtain Bayes' theorem (see Eqs. (2) and (4)),

$$f_4(x|y) = \frac{f_3(y|x)f_1(x)}{f_2(y)} = \frac{f_3(y|x)f_1(x)}{\int f_3(y|x')f_1(x')\,dx'} \ . \tag{13}$$

The mean of $x$ is

$$\mu_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,f(x,y)\,dx\,dy = \int_{-\infty}^{\infty} x\,f_1(x)\,dx \ , \tag{14}$$

and similarly for $y$. The *covariance* of $x$ and $y$ is

$$\mathrm{cov}[x,y] = E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x\mu_y \ . \tag{15}$$

A dimensionless measure of the covariance of $x$ and $y$ is given by the *correlation coefficient*,

$$\rho_{xy} = \mathrm{cov}[x,y]/\sigma_x\sigma_y \ , \tag{16}$$

where $\sigma_x$ and $\sigma_y$ are the standard deviations of $x$ and $y$. It can be shown that $-1 \leq \rho_{xy} \leq 1$.

Two random variables $x$ and $y$ are *independent* if and only if

$$f(x,y) = f_1(x)f_2(y) \ . \tag{17}$$

If $x$ and $y$ are independent, then $\rho_{xy} = 0$; the converse is not necessarily true. If $x$ and $y$ are independent, then for any functions $u(x)$ and $v(y)$ one has $E[u(x)v(y)] = E[u(x)]E[v(y)]$, and also one finds $V[x+y] = V[x] + V[y]$. If $x$ and $y$ are not independent, $V[x+y] = V[x] + V[y] + 2\mathrm{cov}[x,y]$, and $E[uv]$ does not necessarily factorize.

Consider a set of $n$ continuous random variables $\mathbf{x} = (x_1, \ldots, x_n)$ with joint p.d.f. $f(\mathbf{x})$, and a set of $n$ new variables $\mathbf{y} = (y_1, \ldots, y_n)$, related to $\mathbf{x}$ by means of a function $\mathbf{y}(\mathbf{x})$ that is one-to-one, i.e., the inverse $\mathbf{x}(\mathbf{y})$ exists. The joint p.d.f. for $\mathbf{y}$ is given by

$$g(\mathbf{y}) = f(\mathbf{x}(\mathbf{y}))|J| \ \ , \tag{18}$$

where $|J|$ is the absolute value of the determinant of the square matrix $J_{ij} = \partial x_i / \partial y_j$ (the Jacobian determinant). If the transformation from $\mathbf{x}$ to $\mathbf{y}$ is not one-to-one, the $\mathbf{x}$-space must be broken in to regions where the function $\mathbf{y}(\mathbf{x})$ can be inverted, and the contributions to $g(\mathbf{y})$ from each region summed.

Several probability functions and p.d.f.s along with their properites are given in Table 1.

**Table 1.** Probability distributions, their mean values and variances.

| Distribution | | Mean | Variance |
|---|---|---|---|
| Binomial | $f(r;N,p) = \frac{N!}{r!(N-r)!}\, p^r (1-p)^{N-r}$ | $Np$ | $Np(1-p)$ |
| Poisson | $f(n;\nu) = \frac{\nu^n e^{-\nu}}{n!}$ | $\nu$ | $\nu$ |
| Uniform | $f(x;a,b) = \begin{cases} 1/(b-a) & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| Gaussian | $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\, \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$ | $\mu$ | $\sigma^2$ |
| Multivariate Gaussian | $f(\mathbf{x};\boldsymbol{\mu},V) = \frac{1}{(2\pi)^{n/2}\sqrt{\lvert V\rvert}} \times \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T V^{-1} (\mathbf{x}-\boldsymbol{\mu})\right]$ | $\boldsymbol{\mu}$ | $V_{ij}$ |
| Exponential | $f(x;\mu) = \frac{1}{\mu} e^{-x/\mu}$ | $\mu$ | $\mu^2$ |
| Chi-square | $f(z;n) = \frac{z^{n/2-1} e^{-z/2}}{2^{n/2}\Gamma(n/2)}$ | $n$ | $2n$ |
| Student's $t$ | $f(t;n) = \frac{1}{\sqrt{n\pi}}\, \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)} \times \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$ | $0$ $(n > 1)$ | $n/(n-2)$ $(n > 2)$ |
| Gamma | $f(x;\lambda,k) = \frac{x^{k-1}\lambda^k e^{-\lambda x}}{\Gamma(k)}$ | $k/\lambda$ | $k/\lambda^2$ |
| Beta | $f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ |

## 4 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 4.2 and 4.3. The Bayesian approach is outlined in Sec. 4.4.

An *estimator* $\widehat{\theta}$ (written with a hat) is a function of the data whose value, the *estimate*, is intended as a meaningful guess for the value of the parameter $\theta$. There is no fundamental rule dictating how an estimator must be constructed. One tries, therefore, to choose that estimator which has the best properties. The most important of these are (a) *consistency*, (b) *bias*, (c) *efficiency*, and (d) *robustness*.

(a) An estimator is said to be *consistent* if the estimate $\widehat{\theta}$ converges to the true value $\theta$ 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 $\theta$, then $\widehat{\theta}^{\,2}$ is not in general an unbiased estimator for $\theta^2$.

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

$$\sigma^2_{\min} = -\left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right] , \tag{19}$$

where is $L(\theta)$ is the *likelihood function* (see below). The *mean-squared error*,

$$\mathrm{MSE} = E[(\widehat{\theta} - \theta)^2] = V[\widehat{\theta}] + b^2 , \tag{20}$$

is a measure of an estimator's quality which combines the uncertainties due to bias and variance.

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

Simultaneously optimizing for all the measures of estimator quality described above can lead to conflicting requirements. For example, there is in general a trade-off between bias and variance. For some common estimators, the properties above are known exactly. More generally, it is possible to evaluate them by Monte Carlo simulation. Note that they will often depend on the unknown $\theta$.

### 4.1 Estimators for mean, variance and median

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

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{21}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 \tag{22}$$

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

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right) , \tag{23}$$

where $m_4$ is the 4th central moment of $x$. For Gaussian distributed $x_i$, this becomes $2\sigma^4/(N-1)$ for any $N \geq 2$, and for large $N$, the standard deviation of $\widehat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2N}$. Again, if the $x_i$ are Gaussian, $\widehat{\mu}$ is an efficient estimator for $\mu$, and the estimators $\widehat{\mu}$ and $\widehat{\sigma^2}$ are uncorrelated. Otherwise the arithmetic mean (21) is not necessarily the most efficient estimator; this is discussed further in Sec. 8.7 of Ref. (James 2007).

### 4.2 The method of maximum likelihood

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

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i; \boldsymbol{\theta}) \, . \tag{24}$$

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

Note that the likelihood function is *not* a p.d.f. for the parameters $\boldsymbol{\theta}$; in frequentist statistics this is not defined. In Bayesian statistics, one can obtain from the likelihood the posterior p.d.f. for $\boldsymbol{\theta}$, but this requires multiplying by a prior p.d.f. (see Sec. 6.1).

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

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \,, \qquad i = 1, \ldots, n \,. \tag{25}$$

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

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

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

Under requirements usually satisfied in practical analyses and for a sufficiently large data sample, the inverse $V^{-1}$ of the covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$ for a set of ML estimators can be estimated by using

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\boldsymbol{\theta}}} \,. \tag{26}$$

In the large sample limit (or in a linear model with Gaussian errors), $L$ has a Gaussian form and $\ln L$ is (hyper)parabolic. In this case, it can be seen that a numerically equivalent way of determining $s$-standard-deviation errors is from the contour given by the $\boldsymbol{\theta}'$ such that

$$\ln L(\boldsymbol{\theta}') = \ln L_{\max} - s^2/2 \,, \tag{27}$$

where $\ln L_{\max}$ is the value of $\ln L$ at the solution point (compare with Eq. (65)). The extreme limits of this contour on the $\theta_i$ axis give an approximate $s$-standard-deviation confidence interval for $\theta_i$ (see Section 6.2).

In the case where the size $n$ of the data sample $x_1, \ldots, x_n$ is small, the unbinned maximum likelihood method, i.e., use of equation (24), is preferred since binning can only result in a loss of information, and hence larger statistical errors for the parameter estimates. The sample size $n$ can be regarded as fixed, or the user can choose to treat it as a Poisson-distributed variable; this latter option is sometimes called "extended maximum likelihood" (see, e.g., (Lyons 1986; Barlow 1990; Cowan 1998)).

If the sample is large, it can be convenient to bin the values in a histogram, so that one obtains a vector of data $\mathbf{n} = (n_1, \ldots, n_N)$ with expectation values

$\boldsymbol{\nu} = E[\mathbf{n}]$ and probabilities $f(\mathbf{n};\boldsymbol{\nu})$. Then one may maximize the likelihood function based on the contents of the bins (so $i$ labels bins). This is equivalent to maximizing the likelihood ratio $\lambda(\boldsymbol{\theta}) = f(\mathbf{n};\boldsymbol{\nu}(\boldsymbol{\theta}))/f(\mathbf{n};\mathbf{n})$, or to minimizing the equivalent quantity $-2\ln\lambda(\boldsymbol{\theta})$. For independent Poisson distributed $n_i$ this is (Cousins and Baker 1984)

$$-2\ln\lambda(\boldsymbol{\theta}) = 2\sum_{i=1}^{N}\left[\nu_i(\boldsymbol{\theta}) - n_i + n_i\ln\frac{n_i}{\nu_i(\boldsymbol{\theta})}\right] , \tag{28}$$

where for bins with $n_i = 0$, the last term in (28) is zero. The expression (28) without the terms $\nu_i - n_i$ also gives $-2\ln\lambda(\boldsymbol{\theta})$ for multinomially distributed $n_i$, i.e., when the total number of entries is regarded as fixed. In the limit of zero bin width, maximizing (28) is equivalent to maximizing the unbinned likelihood function (24).

A benefit of binning is that it allows for a goodness-of-fit test (see Sec. 5.2). According to Wilks' theorem, for sufficiently large $\nu_i$ and providing certain regularity conditions are met, the minimum of $-2\ln\lambda$ as defined by (28) follows a $\chi^2$ distribution (see, e.g., Ref. (Stuart et al. 1991)). If there are $N$ bins and $m$ fitted parameters, then the number of degrees of freedom for the $\chi^2$ distribution is $N - m$ if the data are treated as Poisson-distributed, and $N - m - 1$ if the $n_i$ are multinomially distributed.

Suppose the $n_i$ are Poisson-distributed and the overall normalization $\nu_{\rm tot} = \sum_i \nu_i$ is taken as an adjustable parameter, so that $\nu_i = \nu_{\rm tot} p_i(\boldsymbol{\theta})$, where the probability to be in the $i$th bin, $p_i(\boldsymbol{\theta})$, does not depend on $\nu_{\rm tot}$. Then by minimizing (28), one obtains that the area under the fitted function is equal to the sum of the histogram contents, i.e., $\sum_i \nu_i = \sum_i n_i$. This is not the case for parameter estimation methods based on a least-squares procedure with traditional weights (see, e.g., Ref. (Cowan 1998)).

### 4.3 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 $\sigma_i^2$. The goal is to construct estimators for the unknown parameters $\boldsymbol{\theta}$. The likelihood function contains the sum of squares

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N}\frac{(y_i - \mu(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} . \tag{29}$$

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

The minimum of Equation (29) 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

$$\chi^2(\boldsymbol{\theta}) = (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T V^{-1} (\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \; , \tag{30}$$

where $\mathbf{y} = (y_1, \ldots, y_N)$ is the vector of measurements, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the corresponding vector of predicted values (understood as a column vector in (30)), and the superscript $T$ denotes transposed (i.e., row) vector.

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

$$\mu(x_i; \boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) \; . \tag{31}$$

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

$$\widehat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \mathbf{y} \equiv D\mathbf{y} \; . \tag{32}$$

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

$$U = DVD^T = (H^T V^{-1} H)^{-1} \; , \tag{33}$$

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

$$(U^{-1})_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \sum_{k,l=1}^{N} h_i(x_k) (V^{-1})_{kl} h_j(x_l) \; . \tag{34}$$

The LS estimators can also be found from the expression

$$\widehat{\boldsymbol{\theta}} = U\mathbf{g} \; , \tag{35}$$

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

$$g_i = \sum_{j,k=1}^{N} y_j h_i(x_k) (V^{-1})_{jk} \; . \tag{36}$$

For the case of uncorrelated $y_i$, for example, one can use (35) with

$$(U^{-1})_{ij} = \sum_{k=1}^{N} \frac{h_i(x_k)h_j(x_k)}{\sigma_k^2} \ , \tag{37}$$

$$g_i = \sum_{k=1}^{N} \frac{y_k h_i(x_k)}{\sigma_k^2} \ . \tag{38}$$

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

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\widehat{\boldsymbol{\theta}}) + 1 = \chi^2_{\rm min} + 1 \tag{39}$$

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

In constructing the quantity $\chi^2(\boldsymbol{\theta})$, one requires the variances or, in the case of correlated measurements, the covariance matrix. Often these quantities are not known *a priori* and must be estimated from the data; an important example is where the measured value $y_i$ represents a counted number of events in the bin of a histogram. If, for example, $y_i$ represents a Poisson variable, for which the variance is equal to the mean, then one can either estimate the variance from the predicted value, $\mu(x_i;\boldsymbol{\theta})$, or from the observed number itself, $y_i$. In the first option, the variances become functions of the fitted parameters, which may lead to calculational difficulties. The second option can be undefined if $y_i$ is zero, and in both cases for small $y_i$, the variance will be poorly estimated. In either case, one should constrain the normalization of the fitted curve to the correct value, i.e., one should determine the area under the fitted curve directly from the number of entries in the histogram (see (Cowan 1998), Section 7.4). A further alternative is to use the method of maximum likelihood; for binned data this can be done by minimizing Eq. (28)

As the minimum value of the $\chi^2$ 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 5.2.

### 4.4 The Bayesian approach

In the frequentist methods discussed above, probability is associated only with data, not with the value of a parameter. This is no longer the case in Bayesian statistics, however, which we introduce in this section. Bayesian methods are considered further in Sec. 6.1 for interval estimation and in Sec. 5.3 for model selection. For general introductions to Bayesian statistics see, e.g., Refs. (O'Hagan and Forster 2004; Sivia and Skilling 2006; Gregory 2005; Bernardo and Smith 2000).

Suppose the outcome of an experiment is characterized by a vector of data $\mathbf{x}$, 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. $p(\boldsymbol{\theta}|\mathbf{x})$, which gives the degree of belief for $\boldsymbol{\theta}$ to take on values in a certain region given the data $\mathbf{x}$. It is obtained by using Bayes' theorem,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'} \,, \tag{40}$$

where $L(\mathbf{x}|\boldsymbol{\theta})$ 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 (40) serves to normalize the posterior p.d.f. to unity.

As it can be difficult to report the full posterior p.d.f. $p(\boldsymbol{\theta}|\mathbf{x})$, one would usually summarize it with statistics such as the mean (or median), and covariance matrix. In addition one may construct intervals with a given probability content, as is discussed in Sec. 6.1 on Bayesian interval estimation.

Bayesian statistics supplies no unique rule for determining the prior $\pi(\boldsymbol{\theta})$; in a subjective Bayesian analysis this reflects the experimenter's 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) that the posterior is proportional to the likelihood, and therefore the mode (peak position) of the posterior is equal to the ML estimator. The posterior mode, however, will change in general upon a transformation of parameter. A summary statistic other than the mode may be used as the Bayesian estimator, such as the median, which is invariant under parameter transformation. But this will not in general coincide with the ML estimator.

The difficult and subjective nature of encoding personal knowledge into priors has led to what is called *objective Bayesian statistics*, where prior probabilities are based not on an actual degree of belief but rather derived from formal rules. These give, for example, priors which are invariant under a transformation of parameters or which result in a maximum gain in information for a given set of measurements. For an extensive review see, e.g., Ref. (Kass and Wasserman 1996).

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

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))} \,, \tag{41}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln L(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} L(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \qquad (42)$$

is the *Fisher information matrix*. One can show that the Jeffreys prior leads to inference that is invariant under a transformation of parameters.

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

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

$$p(\boldsymbol{\theta}|\mathbf{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{\nu}|\mathbf{x})\, d\boldsymbol{\nu}\ . \qquad (43)$$

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) is often used for computing integrals of this type.

## 5 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. 5.1, provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. In *significance tests*, covered in Sec. 5.2, one gives the probability to obtain a level of incompatibility with a certain hypothesis that is greater than or equal to the level observed with the actual data. In the Bayesian approach, the corresponding procedure is referred to as model selection, which is based fundamentally on the probabilities of competing hypotheses. In Sec. 5.3 we describe a related construct called the Bayes factor, which can be used to quantify the degree to which the data prefer one or another hypothesis.

### 5.1 Hypothesis tests

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

A *statistical test* is a rule that states for which values of $\mathbf{x}$ a given hypothesis (often called the null hypothesis, $H_0$) should be rejected in favour of its alternative $H_1$. This is done by defining a region of $\mathbf{x}$-space called the critical region; if the outcome of the experiment lands in this region, $H_0$ is rejected, otherwise it is accepted.

Rejecting $H_0$ if it is true is called an error of the first kind. The probability for this to occur is called the *size* or *significance level* of the test, $\alpha$, which is chosen to be equal to some pre-specified value. It can also happen that $H_0$ is false and the true hypothesis is the alternative, $H_1$. If $H_0$ is accepted in such a case, this is called an error of the second kind, which will have some probability $\beta$. The quantity $1 - \beta$ is called the *power* of the test relative to $H_1$.

In high-energy physics, the components of $\mathbf{x}$ might represent the measured properties of candidate events, and the acceptance region is defined by the cuts that one imposes in order to select events of a certain desired type. Here $H_0$ could represent the background hypothesis and the alternative $H_1$ could represent the sought after signal.

Often rather than using the full set of quantities $\mathbf{x}$, it is convenient to define a *test statistic*, $t$, which can be a single number, or in any case a vector with fewer components than $\mathbf{x}$. Each hypothesis for the distribution of $\mathbf{x}$ will determine a distribution for $t$, and the acceptance region in $\mathbf{x}$-space will correspond to a specific range of values of $t$.

Often one tries to construct a test to maximize power for a given significance level, i.e., to maximize the signal efficiency for a given significance level. The *Neyman–Pearson lemma* states that this is done by defining the critical region for the test of the background hypothesis $H_0$ (i.e., the acceptance region for signal, $H_1$) such that, for $\mathbf{x}$ in that region, the ratio of p.d.f.s for the hypotheses $H_1$ and $H_0$,

$$\lambda(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} \,, \tag{44}$$

is greater than a given constant, the value of which is chosen to give the desired signal efficiency. Here $H_0$ and $H_1$ must be simple hypotheses, i.e., they should not contain undetermined parameters. The lemma is equivalent to the statement that Eq. (44) represents the test statistic with which one may obtain the highest signal efficiency for a given purity for the selected sample. It can be difficult in practice, however, to determine $\lambda(\mathbf{x})$, since this requires knowledge of the joint p.d.f.s $f(\mathbf{x}|H_0)$ and $f(\mathbf{x}|H_1)$.

In the usual case where the likelihood ratio (44) cannot be used explicitly, there exist a variety of other multivariate classifiers that effectively separate different types of events. Methods often used in HEP include *neural networks* or *Fisher discriminants*. Recently, further classification methods from machine-learning have been applied in HEP analyses; these include *probability density estimation (PDE)* techniques, *kernel-based PDE* (*KDE* or *Parzen window*), *support vector machines*, and *decision trees*. Techniques such as "boosting" and "bagging" can be applied to combine a number of classifiers into a stronger one with greater stability with respect to fluctuations in the training data. Descriptions of these methods can be found in (Hastie et al. 2009; Webb 2002; Kuncheva 2004) and Proceedings of the PHYSTAT conference series (PHYSTAT Conference Series). Software for HEP includes the `TMVA` (Höcker et al. 2007) and `StatPatternRecognition` (Narsky 2005) packages.

### 5.2 Significance tests

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

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

$$p = \int_{t_{\rm obs}}^{\infty} g(t|H_0)\, dt \;, \tag{45}$$

where $t_{\rm obs}$ is the value of the statistic obtained in the actual experiment. The $p$-value should not be confused with the size (significance level) of a test, or the confidence level of a confidence interval (Sec. 6), both of which are pre-specified constants.

The $p$-value is a function of the data, and is therefore itself a random variable. If the hypothesis used to compute the $p$-value is true, then for continuous data, $p$ will be uniformly distributed between zero and one. Note that the $p$-value is not the probability for the hypothesis; in frequentist statistics, this is not defined. Rather, the $p$-value is the probability, under the assumption of a hypothesis $H_0$, of obtaining data at least as incompatible with $H_0$ as the data actually observed.

When searching for a new phenomenon, one tries to reject the hypothesis $H_0$ that the data are consistent with known, e.g., Standard Model processes. If the $p$-value of $H_0$ is sufficiently low, then one is willing to accept that some alternative hypothesis is true. Often one converts the $p$-value into an equivalent significance $Z$, defined so that a $Z$ standard deviation upward fluctuation of a Gaussian random variable would have an upper tail area equal to $p$, i.e.,

$$Z = \Phi^{-1}(1 - p) \, . \tag{46}$$

Here $\Phi$ is the cumulative distribution of the Standard Gaussian, and $\Phi^{-1}$ is its inverse (quantile) function. Often in HEP, the level of significance where an effect is said to qualify as a discovery is $Z = 5$, i.e., a $5\sigma$ effect, corresponding to a $p$-value of $2.87 \times 10^{-7}$. One's actual degree of belief that a new process is present, however, will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data, one's confidence in the model that led to the observed $p$-value, and possible corrections for multiple observations out of which one focuses on the smallest $p$-value obtained (the "look-elsewhere effect"). For a review of how to incorporate systematic uncertainties into $p$-values see, e.g., (Demortier 2007).

When estimating parameters using the method of least squares, one obtains the minimum value of the quantity $\chi^2$ (29). This statistic can be used to test the *goodness-of-fit*, i.e., the test provides a measure of the significance of a discrepancy between the data and the hypothesized functional form used in the fit. It may also happen that no parameters are estimated from the data, but that one simply wants to compare a histogram, e.g., a vector of Poisson distributed numbers $\mathbf{n} = (n_1, \ldots, n_N)$, with a hypothesis for their expectation values $\nu_i = E[n_i]$. As the distribution is Poisson with variances $\sigma_i^2 = \nu_i$, the $\chi^2$ (29) becomes *Pearson's* $\chi^2$ *statistic*,

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \, . \tag{47}$$

If the hypothesis $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ is correct, and if the expected values $\nu_i$ in (47) are sufficiently large (in practice, this will be a good approximation if all $\nu_i > 5$), then the $\chi^2$ statistic will follow the $\chi^2$ p.d.f. with the number of degrees of freedom equal to the number of measurements $N$ minus the number of fitted parameters. The minimized $\chi^2$ from Eq. (29) also has this property if the measurements $y_i$ are Gaussian.

Alternatively, one may fit parameters and evaluate goodness-of-fit by minimizing $-2 \ln \lambda$ from (28). One finds that the distribution of this statistic approaches the asymptotic limit faster than does Pearson's $\chi^2$, and thus computing the $p$-value with the $\chi^2$ p.d.f. will in general be better justified (see (Cousins and Baker 1984) and references therein).

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

$$p = \int_{\chi^2}^{\infty} f(z; n_\mathrm{d})\, dz \;, \tag{48}$$

where $f(z; n_\mathrm{d})$ is the $\chi^2$ p.d.f. and $n_\mathrm{d}$ is the appropriate number of degrees of freedom. If the conditions for using the $\chi^2$ p.d.f. do not hold, the statistic can still be defined as before, but its p.d.f. must be determined by other means in order to obtain the $p$-value, e.g., using a Monte Carlo calculation.

Since the mean of the $\chi^2$ distribution is equal to $n_\mathrm{d}$, one expects in a "reasonable" experiment to obtain $\chi^2 \approx n_\mathrm{d}$. Hence the quantity $\chi^2/n_\mathrm{d}$ is sometimes reported. Since the p.d.f. of $\chi^2/n_\mathrm{d}$ depends on $n_\mathrm{d}$, however, one must report $n_\mathrm{d}$ as well if one wishes to determine the $p$-value.

### 5.3 Bayesian model selection

In Bayesian statistics, all of one's knowledge about a model is contained in its posterior probability, which one obtains using Bayes' theorem (40). Thus one could reject a hypothesis $H$ if its posterior probability $P(H|\mathbf{x})$ is sufficiently small. The difficulty here is that $P(H|\mathbf{x})$ 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

$$\pi(H_i, \boldsymbol{\theta}_i) = P(H_i)\pi(\boldsymbol{\theta}_i|H_i) \;, \tag{49}$$

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,

$$P(H_i|\mathbf{x}) = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i) P(H_i) \pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{P(\mathbf{x})} \;, \tag{50}$$

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

$$\frac{P(H_i|\mathbf{x})}{P(H_j|\mathbf{x})} = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int L(\mathbf{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \frac{P(H_i)}{P(H_j)} \;. \tag{51}$$

The *Bayes factor* is defined as

$$B_{ij} = \frac{\int L(\mathbf{x}|\boldsymbol{\theta}_i, H_i)\pi(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int L(\mathbf{x}|\boldsymbol{\theta}_j, H_j)\pi(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j} \;. \tag{52}$$

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

Although the Bayes factor is by construction independent of the overall prior probabilities $P(H_i)$ and $P(H_j)$, it does require priors for all internal parameters of a model, i.e., one needs the functions $\pi(\boldsymbol{\theta}_i|H_i)$ and $\pi(\boldsymbol{\theta}_j|H_j)$. In a Bayesian analysis where one is only interested in the posterior p.d.f. of a parameter, it may be acceptable to take an unnormalizable function for the prior (an improper prior) as long as the product of likelihood and prior can be normalized. But improper priors are only defined up to an arbitrary multiplicative constant, which does not cancel in the ratio (52). 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 (52) are of the form

$$m = \int L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}\ , \tag{53}$$

which is called the *marginal likelihood* (or in some fields called the *evidence*). A review of Bayes factors including a discussion of computational issues is Ref. (Kass and Raftery 1995).

## 6 Intervals and limits

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

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

### 6.1 Bayesian intervals

As described in Sec. 4.4, a Bayesian posterior probability may be used to determine regions that will have a given probability of containing the true value of a parameter. In the single parameter case, for example, an interval (called a Bayesian or credible interval) $[\theta_{\rm lo}, \theta_{\rm up}]$ can be determined which contains a given fraction $1-\alpha$ of the posterior probability, i.e.,

$$1-\alpha = \int_{\theta_{\rm lo}}^{\theta_{\rm up}} p(\theta|\mathbf{x})\, d\theta \ . \tag{54}$$

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

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

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0 \end{cases} \ . \tag{55}$$

This prior is regarded as providing an interval whose frequentist properties can be studied, rather than as representing a degree of belief. In the absence of a clear discovery, (e.g., if $n = 0$ or if in any case $n$ is compatible with the expected background), one usually wishes to place an upper limit on $s$ (see, however, (Feldman and Cousins 1998) on "flip-flopping" concerning frequentist coverage). Using the likelihood function for Poisson distributed $n$,

$$L(n|s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{56}$$

along with the prior (55) in (40) gives the posterior density for $s$. An upper limit $s_{\rm up}$ at confidence level (or here, rather, credibility level) $1-\alpha$ can be obtained by requiring

$$1 - \alpha = \int_{-\infty}^{s_{\rm up}} p(s|n) ds = \frac{\int_{-\infty}^{s_{\rm up}} L(n|s)\,\pi(s)\,ds}{\int_{-\infty}^{\infty} L(n|s)\,\pi(s)\,ds}\,, \tag{57}$$

where the lower limit of integration is effectively zero because of the cut-off in $\pi(s)$. By relating the integrals in Eq. (57) to incomplete gamma functions, the equation reduces to

$$\alpha = e^{-s_{\rm up}} \frac{\sum_{m=0}^{n} (s_{\rm up} + b)^m / m!}{\sum_{m=0}^{n} b^m / m!}\,. \tag{58}$$

This must be solved numerically for the limit $s_{\rm up}$. It so happens that for the case of $b = 0$, the upper limits obtained in this way coincide numerically with the values of the frequentist upper limits discussed in Section 6.2. The frequentist properties of confidence intervals for the Poisson mean obtained in this way are discussed in Refs. (Cousins 1995) and (Roe and Woodroofe 2001).

As in any Bayesian analysis, it is important to show how the result would change if one uses different prior probabilities. For example, one could consider the Jeffreys prior as described in Sec. 4.4. For this problem one finds the Jeffreys prior $\pi(s) \propto 1/\sqrt{s+b}$ for $s \geq 0$ and zero otherwise. As with the constant prior, one would not regard this as representing one's prior beliefs about $s$, both because it is improper and also as it depends on $b$. Rather it is used with Bayes' theorem to produce an interval whose frequentist properties can be studied.

### 6.2 Frequentist confidence intervals

The frequentist approach to interval estimation is based on the the concept of a confidence interval. These are constructed so as to contain the true value of the parameter with a minimum specified probability, called the confidence level. To construct a confidence interval, consider a test (see Sec. 5) of the hypothesis that the parameter's true value is $\theta$ (assume one constructs a test for all physical values of $\theta$). One then excludes all values of $\theta$ where the hypothesis would be rejected at a significance level less than $\alpha$. The remaining values constitute the confidence interval at confidence level $\mathrm{CL} = 1 - \alpha$.

In this procedure, one is still free to choose the test to be used. One possibility is use a test statistic based on the *likelihood ratio*,

$$\lambda = \frac{f(x;\theta)}{f(x;\widehat{\theta}\,)}\,, \tag{59}$$

where $\widehat{\theta}$ is the value of the parameter which, out of all allowed values, maximizes $f(x;\theta)$. This results in the intervals described in (Feldman and Cousins 1998) by Feldman and Cousins.

**Profile likelihood and treatment of nuisance parameters**

As mentioned in Section 6.1, one may have a model containing parameters that must be determined from data, but which are not of any interest in the final result (nuisance parameters). Suppose the likelihood $L(\boldsymbol{\theta}, \boldsymbol{\nu})$ depends on parameters of interest $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\nu}$. The nuisance parameters can be effectively removed from the problem by constructing the *profile likelihood*, defined by

$$L_{\rm p}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\nu}}}(\boldsymbol{\theta})) \; , \tag{60}$$

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

$$\lambda_p(\boldsymbol{\theta}) = \frac{L_{\rm p}(\boldsymbol{\theta})}{L(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\nu}})} \; , \tag{61}$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\nu}}$ are the ML estimators. The ratio $\lambda_{\rm p}$ can be used in place of the likelihood ratio (59) for inference about $\boldsymbol{\theta}$. The resulting intervals for the parameters of interest are not guaranteed to have the exact coverage probability for all values of the nuisance parameters, but in cases of practical interest the approximation is found to be very good.

**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 $\sigma$,

$$1 - \alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2} \, dx = {\rm erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right) \tag{62}$$

is the probability that the measured value $x$ will fall within $\pm\delta$ of the true value $\mu$. From the symmetry of the Gaussian with respect to $x$ and $\mu$, this is also the probability for the interval $x \pm \delta$ to include $\mu$. Figure 1 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 $\sigma$ is known. Values of $\alpha$ for other frequently used choices of $\delta$ are given in Table tab:stat-statone.

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

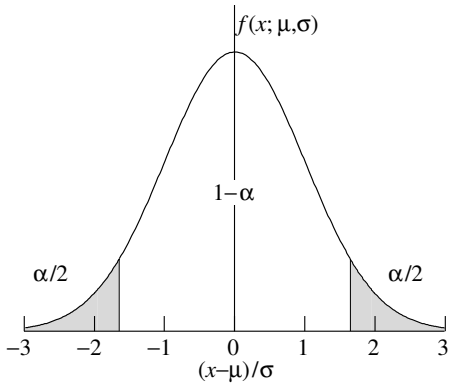


**Fig. 1.** Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by $\alpha$, are as shown.

**Table 2.** Area of the tails $\alpha$ outside $\pm\delta$ from the mean of a Gaussian distribution.

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

The relation (62) can be re-expressed using the cumulative distribution function for the $\chi^2$ distribution as

$$\alpha = 1 - F(\chi^2; n) \,, \tag{63}$$

for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom.

For the case of a $n$ parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, one requires the full covariance matrix $V_{ij} = \mathrm{cov}[\widehat{\theta}_i, \widehat{\theta}_j]$, which can be estimated as described in Sections 4.2 and 4.3. 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 centred about the true (unknown) values $\boldsymbol{\theta}$, and furthermore, the likelihood function itself takes on a Gaussian shape.

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

$$\tan 2\phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_j^2 - \sigma_i^2} \ , \tag{64}$$

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

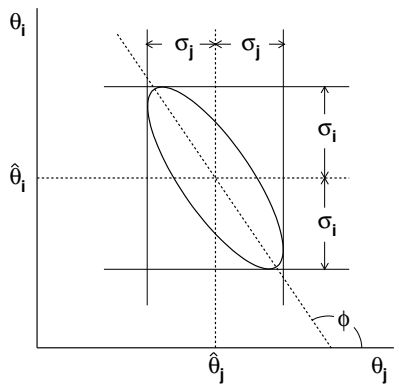


**Fig. 2.** Standard error ellipse for the estimators $\hat{\theta}_i$ and $\hat{\theta}_j$. In this case the correlation is negative.

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

$$\ln L(\boldsymbol{\theta}) \geq \ln L_{\max} - \Delta \ln L \ , \tag{65}$$

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

$$\chi^2(\boldsymbol{\theta}) \leq \chi^2_{\min} + \Delta\chi^2 \ . \tag{66}$$

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

For finite data samples, the probability for the regions determined by equations (65) or (66) to cover the true value of $\boldsymbol{\theta}$ will depend on $\boldsymbol{\theta}$, so these are not exact confidence regions according to our previous definition. Nevertheless, they can still have a coverage probability only weakly dependent on the true parameter, and approximately as given in Table tab:stat-stattwo. 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.

**Table 3.** $\Delta\chi^2$ 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 |
| 90.0 | 2.71 | 4.61 | 6.25 |
| 95.0 | 3.84 | 5.99 | 7.82 |
| 95.45 | 4.00 | 6.18 | 8.03 |
| 99.0 | 6.63 | 9.21 | 11.34 |
| 99.73 | 9.00 | 11.83 | 14.16 |

### 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 $\sigma$, say, in a fixed integrated luminosity $\mathcal{L}$, then it follows a Poisson distribution with mean $\nu = \sigma\mathcal{L}$. If, on the other hand, one has selected a larger sample of $N$ events and found $n$ of them to have a particular property, then $n$ follows a binomial distribution where the parameter $p$ gives the probability for the event to possess the property in question. This is appropriate, e.g., for estimates of branching ratios or selection efficiencies based on a given total number of events.

For the case of Poisson distributed $n$, the upper and lower limits on the mean value $\nu$ can be found from

$$\nu_{\rm lo} = \tfrac{1}{2} F^{-1}_{\chi^2}(\alpha_{\rm lo}; 2n) \,, \tag{67}$$

$$\nu_{\rm up} = \tfrac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{\rm up}; 2(n+1)) \,, \tag{68}$$

where the upper and lower limits are at confidence levels of $1-\alpha_{\rm lo}$ and $1-\alpha_{\rm up}$, respectively, and $F^{-1}_{\chi^2}$ is the *quantile* of the $\chi^2$ distribution (inverse of the cumulative distribution). For central confidence intervals at confidence level $1-\alpha$, set $\alpha_{\rm lo} = \alpha_{\rm up} = \alpha/2$.

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

$$p_{\rm lo} = \frac{nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]}{N-n+1 \,+\, nF^{-1}_F[\alpha_{\rm lo}; 2n, 2(N-n+1)]} \,, \tag{69}$$

$$p_{\rm up} = \frac{(n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]}{(N-n) \,+\, (n+1)F^{-1}_F[1-\alpha_{\rm up}; 2(n+1), 2(N-n)]} \,. \tag{70}$$

Here $F^{-1}_F$ is the quantile of the $F$ distribution (also called the Fisher–Snedecor distribution; see (James 2007)).

## 7 Conclusions

Given the high cost and complexity of experiments in Particle Physics it has become increasingly important to use data analysis methods that extract the maximum information from the data in a way that takes into account all of the known uncertainties in the measurement. Here the key is to construct a probabilistic model which is sufficiently accurate to be regarded as correct, and this can be achieved with a model containing a sufficient number of adjustable parameters. The accuracy achieved by including additional parameters must be balanced against the price, however, of reducing one's sensitivity to the parameters of interest, such as those which may point to a potential discovery.

The two primary schools of statistical inference — frequentist and Bayesian — provide different but related approaches to this task. In the fortunate case where the information from the data overwhelms any prior knowledge, the two approaches appear to coalesce, although the interpretation of the results remains distinct. One can of course use both approaches; if they point to the same conclusion, then this can only increase one's confidence. If they do not, then one will have to find out why, and this can also lead to important realisations, such as unexpected sensitivity to prior information or to specific model assumptions. In either case the result of an experiment should be presented along with sufficient information so that it can be incorporated into future analyses.

## Acknowledgements

The author is indebted to the Particle Data Group for permission to adapt the material from the *Review of Particle Physics* (Amsler et al. 2008) for this article, as well as to Dr Tilo Stroh and Professor Claus Grupen for their support and assistance.

## References

[1] C. Amsler et al. (Particle Data Group), Physics Letters B667, 1 (2008).
[2] A.N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, (Springer, Berlin 1933); *Foundations of the Theory of Probability*, 2nd Ed., (Chelsea, New York 1956).
[3] R.D. Cousins, Am. J. Phys. **63**, 398 (1995).
[4] 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.
[5] F.E. James, *Statistical Methods in Experimental Physics*, 2nd ed., World Scientific, Singapore, 2007.

[6] L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, New York, 1986.
[7] R. Barlow, Nucl. Inst. Meth. A297, 496 (1990).
[8] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.
[9] For a review, see S. Baker and R. Cousins, Nucl. Inst. Meth. 221, 437 (1984).
[10] A. O'Hagan and J.J. Forster, *Bayesian Inference*, (2nd edition, volume 2B of *Kendall's Advanced Theory of Statistics*, Arnold, London, 2004).
[11] Devinderjit Sivia and John Skilling, *Data Analysis: A Bayesian Tutorial*, (Oxford University Press, 2006).
[12] P.C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, (Cambridge University Press, 2005)
[13] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, (Wiley, 2000).
[14] Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996)
[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd edition, Springer, New York, 2009.
[16] A. Webb, *Statistical Pattern Recognition*, 2nd ed., (Wiley, New York, 2002).
[17] L.I. Kuncheva, *Combining Pattern Classifiers*, (Wiley, New York, 2004).
[18] Links to the Proceedings of the PHYSTAT conference series (Durham 2002, Stanford 2003, Oxford 2005, and Geneva 2007) can be found at `phystat.org`.
[19] A. Höcker et al., *TMVA Users Guide*, `physics/0703039` (2007); software available from `tmva.sf.net`.
[20] I. Narsky, *StatPatternRecognition: A C++ Package for Statistical Analysis of High Energy Physics Data*, `physics/0507143`(2005); software avail. from `sourceforge.net/projects/statpatrec`.
[21] L. Demortier, *P-Values and Nuisance Parameters*, Proceedings of PHYSTAT 2007, CERN-2008-001, p. 23.
[22] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D63**, 13009 (2001)
[23] Robert E. Kass and Adrian E. Raftery, *Bayes Factors*, J. Am. Stat. Assoc., Vol. 90, No. 430, pp. 773-795 (1995)
[24] G.J. Feldman and R.D. Cousins, Phys. Rev. D 57, 3873 (1998).