Draft - February 2001

The Maximum-Likelihood Method

The maximum-likelihood principle
Information and variance
Properties of the maximum-likelihood method
Application of the maximum-likelihood method
Gaussian distributions
Binned distributions
Glossary

Volker Blobel

II. Institut für Experimentalphysik
Universität Hamburg
2001

List of mathematical symbols

x, $y\ldots$	random variable (one-dimensional)
$\boldsymbol{x}$, $\boldsymbol{y}\ldots$	random vector (n-dimensional)
μ	expectation value, mean value
$\langle x \rangle$	mean value
μ_x	expectation value, mean value of random variable x
σ	standard deviation (or cross section)
σ_x	standard deviation of random variable x
σ_{xx}	variance = square of standard deviation of random variable x, $\sigma_{xx} = \sigma_x^2$
σ_{xy}	covariance between random variables x and y
$\boldsymbol{V}$	covariance matrix
ρ	correlation coefficient
ρ_{xy}	correlation coefficient between random variables x and y, $\rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y)$
$N(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$N(0,1)$	Normalized Gaussian distribution with mean 0 and variance 1
$U(0,1)$	Uniform distribution between 0 and 1

Contents

1 The maximum-likelihood principle **1**
1.1 Estimation of a single parameter . . . 1
1.2 The variance of parameter estimates . . . 3
1.3 The case of several parameters . . . 5

2 Information and variance **6**
2.1 The information inequality . . . 6
2.2 Experimental design . . . 8
2.3 The case of several variables . . . 9

3 Properties of the maximum-likelihood method **10**
3.1 Consistency . . . 10
3.2 Asymptotic normality . . . 11
3.3 Invariance . . . 12

4 Application of the maximum-likelihood method **14**
4.1 Limited acceptance . . . 14

5 Gaussian distributions **16**
5.1 Weighted mean of Gaussian data . . . 16
5.2 Mean and variance of a Gaussian . . . 16

6 Binned distributions **18**
6.1 The problem . . . 18
6.2 Fitting using finite Monte Carlo samples . . . 21

Glossary **23**

1 The maximum-likelihood principle

1.1 Estimation of a single parameter

Assume that a measurement is performed in the space of the random variable x. The distribution of the measured values x is assumed to be known to follow the (normalized) *probability density* $p(x|a)$

$$p(x|a) \geq 0 \qquad \text{with} \qquad \int_\Omega p(x|a)\,\mathrm{d}x = 1 \tag{1}$$

in the x-space, which depends on a single parameter a. The normalization integral is over the complete x-space Ω. The expression $p(x|a)$ represents a family of density functions, depending on the parameter a. A standard problem of data analysis is the estimation of parameters from measured values, called *point estimation.*

> From a given set of n measured values $x_1, \ldots, x_i, \ldots, x_n$ the optimal value of the parameter a has to be estimated. The distribution of the random values x_i is known to follow the probability density $p(x|a)$.

The *maximum-likelihood method* starts from the *joint* probability distribution of the n measured values $x_1, \ldots, x_i, \ldots, x_n$, which for *independent* measurements is given by the product of the individual densities $p(x|a)$, which is

$$\mathcal{L}(a) = p(x_1|a) \cdot p(x_2|a) \cdots p(x_n|a) = \prod_{i=1}^{n} p(x_i|a) \; . \tag{2}$$

The function $\mathcal{L}(a)$, for a given set $\{x_i\}$ of measurements considered as a function of the parameter a, is called the *likelihood function* and is a *function* (not a density function) of the parameter a.
According to the **Principle of Maximum Likelihood** (ML) the estimate $\widehat{a}$ for the parameters a is the value, which *maximizes* the likelihood function $\mathcal{L}(x|a)$. In regular cases the maximum is determined by setting the derivative of the likelihood function with respect to the parameter a to zero: $\mathrm{d}\mathcal{L}/\mathrm{d}a = 0$. The condition locates maxima as well as minima and one has to avoid a root of the equation which actually corresponds to a minimum. In certain cases the likelihood function does not have zero slope at its maximum, and the maximum has to be located by other means. An example is the case, where the maximum of the function $\mathcal{L}(a)$ is at the boundary of the physical allowed region of the parameter a.

For technical and also for theoretical reasons it is easier to work with the logarithm (a monotonically increasing function of its argument) of the likelihood function $\mathcal{L}(\boldsymbol{a})$, or with the *negative* logarithm. In the following the *negative* log-likelihood function[1] is considered,

$$\boxed{F(a) = -\sum_{i=1}^{n} \ln p(x_i|a)} \tag{3}$$

and the maximum likelihood estimate $\widehat{a}$ is the value that *minimizes* this function.
In regular cases it is sufficient to solve the *Likelihood equation*

$$\frac{\mathrm{d}F(a)}{\mathrm{d}a} = 0 \tag{4}$$

for the case of a single parameter a. The derivative of the function $F(a)$ with repect to a parameter is usually to a good approximation *linear* in the neighbourhood of the minimum at $\widehat{a}$ with a positive slope, as required for the estimate $\widehat{a}$ to be at a minimum.

> If an efficient estimate $\widehat{a}$ of a exists, the likelihood equation will have a unique solution equal to $\widehat{a}$.

As seen from the definition of the negative log-likelihood function in equation (3), the fact that the normalization does *not* depend on the value of the parameter a is essential, while the actual value of the integral does not matter for the position of the minimum or the error analysis, to be discussed later. Thus

[1] Sometimes a factor of 2 is included in the definition of the negative log-likehood function; this factor makes it similar to the χ^2-expression of the method of least squares in certain applications.

normalization factors not dependent on the parameters do not influence the position of the minimum and can be dropped.

The maximum likelihood method has some optimal properties and provides an efficient estimate of a parameter. The application of the Maximum Likelihood method requires the knowledge of the true distribution in the sample space, apart from the parameters values a, which are determined by this method. The detailed knowledge of the distribution is the source of the power of the method. Maximum likelihood estimate are not robust: assumption of an incorrect distribution function will introduce a systematic shift or *bias* in the estimated parameter value.

The interpretation of the *value* of the log-likelihood function is often not easy, although it contains the information about the goodness-of-fit. Results of several experiments can be combined by adding the log-likelihood functions.

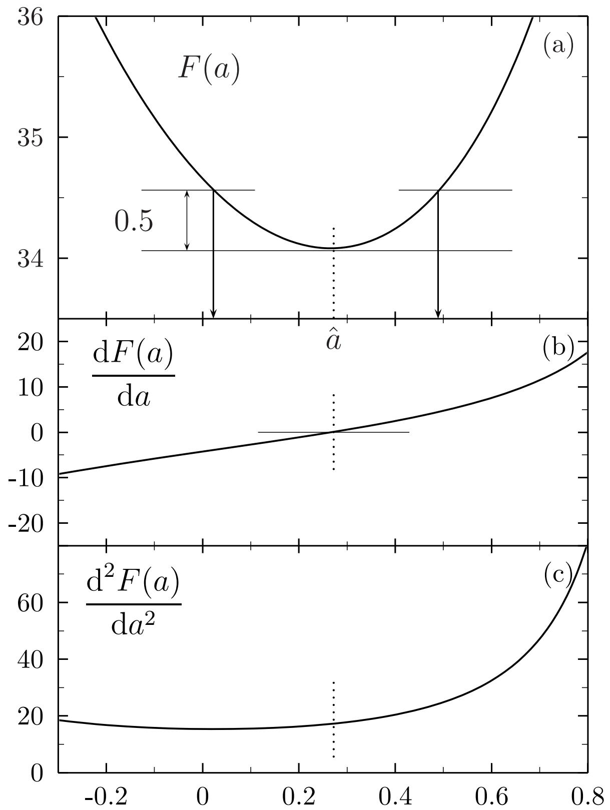

Figure 1: The negative log-likelihood function $F(a)$ for the angular distribution (a) as a function of the parameter a. The best estimate $\widehat{a}$ is given by the minimum F_{min} of the function. The parameter values a with $F(a) = F_{\text{min}}+0.5$ define the 1σ limits (arrows). The first derivative $\mathrm{d}F/\mathrm{d}a$ in (b) is approximately linear. the second derivative $\mathrm{d}^2F/\mathrm{d}a^2$ in (c) is approximately constant in the region of ± 1 standard deviation, but outside this region there are clear deviations.

Example 1 Decay angular distribution of an elementary particle

The decay angular distribution of certain elementary particles is given by the probability density

$$p(x|a) = \frac{1}{2}\,(1 + ax) \qquad \text{with} \quad x \equiv \cos\vartheta\ . \tag{5}$$

Assume that n values x_i have been measured for a sample of decays and should be used to determine an estimate $\widehat{a}$ of the decay parameter a. The probability density above is normalized correctly for the physical region $-1 \leq x \leq +1$, and can thus directly be used for the definition of the function to be minimized:

$$F(a) = -\sum_{i=1}^{n} \ln \frac{1}{2}\,(1 + ax_i)\ . \tag{6}$$

In this case the minimum can be read off simply from a graph of the function of the parameter a, since only one parameter a is involved. The figure 1 shows an example of the shape of the function as a function of the parameter a for a sample of measured values, generated by Monte Carlo methods for the true value of the parameter $a = 0.25$. The probability function $p(x|a)$ defines here the physical region of the parameter a as $-1 \le a \le +1$ (outside this region the probability density will become negative for certain values of the random variable x. In addition the first and the second derivatives with respect to the parameter a are shown. As can be seen from the figure the shape is approximately parabolic (second derivative approximately constant).

Example 2 Parameters of an uniform distribution

The density of the uniform distribution of a random variable x,

$$f(x) = \frac{1}{a_2 - a_1} \qquad\qquad a_1 < x < a_2$$

depends on the two parameters a_1 and a_2, which are the limits of the distribution. In order to estimate the values of these two parameters from a sample of observed values $x_i, i = 1, 2, \ldots n$, the negative log-likelihood function

$$F(a_1, a_2) = -\ln \left[\prod_{i=1}^{n} \frac{1}{a_2 - a_1} \right] = n \cdot \ln(a_2 - a_1)$$

is formed, which in this special case does not depend on the observed values. If the derivatives of this expression with respect to a_1 and a_2 are put to zero, at least one of the parameters must be infinite. However in this case the function $F(a_1, a_2)$ does not have a zero slope at the minimum. It is easy to see, that the function $F(a_1, a_2)$ has the minimum value for the shortest interval $a_1 \ldots a_2$. Thus the Maximum Likelihood estimates for a_1 and a_2 are the smallest and largest values of the observations:

$$\widehat{a}_1 = \min_{i=1}^{n}(x_i) \qquad\qquad \widehat{a}_2 = \max_{i=1}^{n}(x_i)$$

Apparently no use is made of the intervening observations.

1.2 The variance of parameter estimates

The ML principles gives an estimate $\widehat{a} = \widehat{a}(x_1, \ldots, x_i, \ldots, x_n)$. In addition to the estimate a measure of the uncertainty is required, which is usually the standard deviation σ_a of the estimate, the square root of the variance. The variance

$$V[\widehat{a}] = E\left[\widehat{a}^2\right] - \left(E[\widehat{a}]\right)^2$$

is a function of the measured data x_i and standard methods of error propagation can be used to determine the variance in a given case.

If the experiment were repeated many times under the same conditions (same number of observations n), then one would expect that the distribution of the estimate would have the standard deviation σ_a. For a Gaussian distribution of the estimates $\pm$ the standard deviation has a fixed meaning of a 68.3 % confidence interval. In the large sample limit the ML estimates are in fact distributed according to a Gaussian distribution.

Asymptotically, that is in the limit of an infinite number n of observations ($n \to \infty$) the likelihood estimates follow a Gaussian density with respect to the parameters a. This is shown in the *Central Limit theorem.*

In this limit the negative log-likelihood function around the minimum is approximated by a quadratic function

$$F(a) = F_{\min} + \frac{1}{2} \left. \frac{\mathrm{d}^2 F}{\mathrm{d}a^2} \right|_{a=\widehat{a}} (a - \widehat{a})^2 + \ldots \tag{7}$$

and the variance of the parameter is given by the inverse of the expectation value of the second derivative of the function $F(a)$:

$$\sigma^2(\widehat{a}) = \left(E\left[\frac{\mathrm{d}^2 F}{\mathrm{d}a^2} \right] \right)^{-1} \tag{8}$$

and in the limit ($n \to \infty$) the matrix of second derivatives becomes a constant, which can be evaluated at $a = \widehat{a}$. The standard deviation is approximately given as the square root of the inverse of the second derivative of F with respect to the parameter a, with the formula[2]

$$\text{Magic formula I} \qquad \boxed{\sigma(\widehat{a}) \approx \left(\left. \frac{\mathrm{d}^2 F}{\mathrm{d}a^2} \right|_{a=\widehat{a}} \right)^{-1/2}} \tag{9}$$

A numerically equivalent method to determine the deviation corresponding to k standard deviations is given in this approximation by the equation

$$F(a \pm k \cdot \sigma) = F_{\min} + \frac{1}{2} k^2 \tag{10}$$

which for example allows to determine the (one) standard deviations from the values a, where the function $F(a) = F_{\min} + 1/2$; this definition appears to be better in practice and allows to determine *asymmetric* deviations.

Example 3 Estimate for the parameter of the Poisson distribution

An estimate $\widehat{\mu}$ for the parameter μ of the Poisson distribution has to be determined from a sample of n single measurements $x_i,\ i = 1, 2, \ldots n$ (the x_i are integer numbers). From the Poisson probability, to observe a count of x_i, if the mean value is μ:

$$P_\mu(x_i) = \frac{\mu^{x_i}}{x_i!} e^{-\mu} \ ,$$

the negative log-likelihood function

$$F(\mu) = -\sum_{i=1}^{n} \ln P_\mu(x_i) = -\sum_{i=1}^{n} \left[x_i \ln \mu - \mu - \ln x_i! \right] = -\ln \mu \cdot \sum_{i=1}^{n} x_i + n\mu + \text{const}$$

follows (the terms $\ln x_i!$ do not depend on the parameter μ and are thus a constant for $F(\mu)$). The first and second derivative of the function $F(\mu)$ with respect to μ are:

$$\frac{\mathrm{d}F}{\mathrm{d}\mu} = n - \frac{1}{\mu} \sum_{i=1}^{n} x_i \qquad\qquad \frac{\mathrm{d}^2 F}{\mathrm{d}\mu^2} = \frac{1}{\mu^2} \sum_{i=1}^{n} x_i$$

The Maximum Likelihood estimate $\widehat{\mu}$ is determined by the requirement that the first derivative is zero (minimum of the Likelihood function); the variance of the estimate is the inverse of the second derivative at the estimated position of the minimum:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad\qquad V(\widehat{\mu}) = \frac{\widehat{\mu}}{n}$$

Thus the estimate of the parameter of the Poisson distribution is the arithmetic mean of the single measurements, as expected.

Example 4 Binomial distribution

If the probability for a certain event A is equal to p, the probability to observe in n experiments exactly k of the events A is given by the Binomial distribution

$$P_n(k) = \binom{n}{k} p^k \, (1-p)^{n-k} \qquad\qquad k = 0, 1, \ldots n \tag{11}$$

[2]This formula is sometimes called *Magic Formula I.*

with mean value $E[k] = pn$ and variance $V[k] = np(1-p)$. In a single event there is a random variable x with possible values $x = 1$ (event A with probability p) and $x = 0$ (event $\overline{A}$ with probability$(1-p)$). For a single event (or measurement) x_i the probability for the random variable x is given by the point binomial distribution:

$$P_p(x) = p^x(1-p)^{(1-x)} \qquad x = 1,\, 0 \qquad 0 \le p \le 1\,.$$

The probability p has to be estimated from a sample $x_i,\ i = 1, 2, \ldots n$ of observed values; thus the sample is a series of 0's and 1's. The negative log-likelihood function

$$F(p) = -\ln p \sum_{i=1}^{n} x_i - \ln(1-p) \sum_{i=1}^{n} (1-x_i) = -k \cdot \ln p - (n-k) \cdot \ln(1-p)$$

with $k = \sum_i x_i$. First and second derivate are

$$\frac{\mathrm{d}F}{\mathrm{d}p} = -\frac{k}{p} + \frac{n-k}{1-p} \qquad \frac{\mathrm{d}^2F}{\mathrm{d}p^2} = \frac{k}{p^2} + \frac{n-k}{(1-p)^2}\;.$$

The result for the estimate $\widehat{p}$ from $\mathrm{d}F/\mathrm{d}p = 0$ is the obvious estimator for the parameter:

$$\widehat{p} = \frac{r}{n} \qquad V[\widehat{p}] = \sigma^2(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n} = \frac{k\,(n-k)}{n^3}$$

and the variance for $\widehat{p}$ follows from the variance of k. The same value for the variance follows from the inverse of the second derivative with $p = \widehat{p}$:

$$\frac{\mathrm{d}^2F}{\mathrm{d}p^2} = \frac{n}{k^2} + \frac{n-k}{(1-k/n)^2} = \frac{n^3}{k(n-k)}\;.$$

1.3 The case of several parameters

If the probability density function depends on several parameters $a_1, \ldots, a_j, \ldots, a_m$ the m parameters can be combined to a vector $\boldsymbol{a}$ and the negative log-likelihood function is then a function of the vector $\boldsymbol{a}$. The measured value can also be vectors, written as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$.

$$\boxed{F(\boldsymbol{a}) = -\sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\boldsymbol{a})} \tag{12}$$

In order to determine the minimum a system of m equations has to be solved, the *likelihood equation(s)*

$$\nabla F = 0 \qquad \frac{\partial F(\boldsymbol{a})}{\partial a_j} = 0 \quad j = 1, 2, \ldots, m \tag{13}$$

The covariance matrix $\boldsymbol{V}$ of the maximum-likelihood estimate is estimated by the inverse of the Hessian $\boldsymbol{H}$

$$\boldsymbol{V}(\widehat{\boldsymbol{a}}) = \boldsymbol{H}^{-1} \tag{14}$$

where the elements of the matrix $\boldsymbol{H}$ are defined by the expectation value of the second derivatives of the function $F(\boldsymbol{a})$ with respect to the parameters:

$$H_{jk} = E\left[\frac{\partial^2 F}{\partial a_j \partial a_k}\right]\;. \tag{15}$$

Again, in the asymptotic case the likelihood function $\mathcal{L}(\boldsymbol{a})$ becomes a Gaussian and the function $F(\boldsymbol{a})$ becomes a multidimensional quadratic function, with constant second derivatives.
In practice the elements of the matrix $\boldsymbol{H}$ are often approximated by the second derivatives at the minimum:

$$H_{jk} = \left.\frac{\partial^2 F}{\partial a_j \partial a_k}\right|_{\boldsymbol{a}=\widehat{\boldsymbol{a}}}\;. \tag{16}$$

Around the minimum the negative log-likelihood function is approximated by the expansion

$$F(\boldsymbol{a}) = F_{\min} + \frac{1}{2}\,(\boldsymbol{a} - \widehat{\boldsymbol{a}})^T\, H\,(\boldsymbol{a} - \widehat{\boldsymbol{a}}) + \ldots \quad . \tag{17}$$

2 Information and variance

In this chapter the amount of *information* about parameters given by a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from a distribution $p(x|a)$ is expressed and related to the variance of the estimated parameter value. The discussion is first restricted to the case of a single parameter a, and later generalized to the case of several parameters.

2.1 The information inequality

Basis of the Maximum Likelihood method is the likelihood function $\mathcal{L}(a)$, which is formed from the experimental data and the knowledge of the distribution of the experimental data. The likelihood function $\mathcal{L}(a)$ can be considered as

- a *function* of the parameter a, or as
- the joint *density* of the observed values of the random variable x.

In the case of a given sample of n *independent* measurements x from the density $p(x|a)$ the likelihood function $\mathcal{L}(a)$ is simply the *product* of the n density values:

$$\mathcal{L}(a) = \prod_{i=1}^{n} p(x_i|a) \; . \tag{18}$$

In more complicated cases of not-independent measurements it will have a different form, but the two statements above are generally valid. The concept of the likelihood function $\mathcal{L}(a)$ and of the information is useful not only for applications of the maximum likelihood method, but in general for estimation methods.

According to the maximum likelihod principle the likelihood estimate $\widehat{a}$ is determined by the maximum of $\mathcal{L}(a)$, considered as a function of a. The likelihood estimate $\widehat{a}$ and also any other statistical estimate is not a function of the parameter a, but of the data $\{\boldsymbol{x}\}$, that is: $\widehat{a} = \widehat{a}(x_1, \ldots, x_i, \ldots, x_n)$ or short: $\widehat{a} = \widehat{a}(x)$. The statistical properties of the estimate can be derived from the expression $\widehat{a} = \widehat{a}(x)$ in each special case. A basic quantity characterizing the precision of an estimate $\widehat{a}$ is the variance of the estimate defined by the expectation value

$$V[\widehat{a}] = \int_{\Omega} (\widehat{a} - a)^2 \, \mathcal{L} \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \; . \tag{19}$$

Another important quantity is the bias $B(\widehat{a})$ defined by

$$B_n(\widehat{a}) = E[\widehat{a}] - a_{\text{true}} \; ; \tag{20}$$

the subscript n in $B_n(\widehat{a})$ indicates that the bias depends on the sample size n. Likelihood estimates may have a bias $B_n(\widehat{a}) \neq 0$ for finite values of n, i.e. a systematic shift of the estimate $\widehat{a}$ compared to the true value of the parameter.

In the information inequality, to be discussed in this section, the variance of the estimator is related to a quantity I called *Information*, which is defined as a measure of the parameter information given by a data sample of size n. The information I is defined by the expectation value

$$\text{Information} \qquad \boxed{I(a) = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2\right] = \int_{\Omega} \left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2 \mathcal{L} \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n} \, , \tag{21}$$

where $\mathcal{L}$ is the joint density of the n observed values of the random variable x. The information I is related to the variance of an estimator, which may be derived using the maximum likelihood principle or by other methods.

Information inequality: The *inverse* of the information $I_n(a)$, or short I, is the lower limit of the variance of the parameter estimate $\widehat{a}$:

$$\text{Information inequality} \qquad \boxed{V[\widehat{a}] \geq \frac{1}{I}} \, . \tag{22}$$

for an unbiased estimate $\widehat{a}$. The inequality is also called Rao-Cramér-Frechet inequality, and is valid in this form for any unbiased estimate $\widehat{a} = \widehat{a}(x)$.

Proof of the information inequality : The proof ist based on the two following equations:

$$E[\widehat{a}] = \int_\Omega \widehat{a}\, \mathcal{L}(a)\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \quad = \quad a \tag{23}$$

$$\int_\Omega \mathcal{L}(a)\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \quad = \quad 1\,. \tag{24}$$

The first expression (23) is valid for an unbiased estimate $\widehat{a} = \widehat{a}(x)$, and the second expression (24) about the normalization of $\mathcal{L}(a)$ with respect to $x_1, \ldots, x_i, \ldots, x_n$ follows from the normalization of the density function $p(x|a)$.

In the proof sufficient regularity has to be assumed to allow differentiation with respect to a under the integral sign. This assumption may be wrong if the density function $p(x,a)$ does not fall off rapidly enough to ensure existence of the integral of the differentiated integrand; in practice this assumption should not cause problems. The assumption requires in addition that the effective range (where $p(x,a)$ is non-zero) of the x-integration does not depend on the value of the parameter a.

The first step in the proof is to take the derivative of the expression (23) with respect to the parameter a:

$$\frac{\partial}{\partial a} \int_\Omega \widehat{a}\, \mathcal{L}(a)\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1\,. \tag{25}$$

Assuming sufficient regularity as discussed above the derivative is written in the form (note that $\widehat{a}$ is not a function of a)[3]

$$\int_\Omega \widehat{a}(x) \frac{\partial \mathcal{L}(a)}{\partial a}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = \int_\Omega \widehat{a}(x) \frac{\partial \ln \mathcal{L}(a)}{\partial a}\, \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1\,. \tag{26}$$

Now the second equation (24) expressing the normalization is differentiated with repect to the parameter a:

$$\int_\Omega \frac{\partial \mathcal{L}}{\partial a}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = \int_\Omega \frac{\partial \ln \mathcal{L}}{\partial a}\, \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)\right] = 0 \tag{27}$$

Multiplying equation (27) by a und subtracting it from the equation (26) the expression

$$\int_\Omega (\widehat{a} - a) \left(\frac{\partial \ln \mathcal{L}}{\partial a}\right) \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1 \tag{28}$$

is obtained. Now the → Schwarz inequality

$$\int f^2\, \mathrm{d}x \cdot \int g^2\, \mathrm{d}x \geq \left(\int f\, g\, \mathrm{d}x\right)^2 \tag{29}$$

is invoked with $f \equiv (\widehat{a} - a)\sqrt{\mathcal{L}}$ and $g \equiv (\partial \ln \mathcal{L}/\partial a)\sqrt{\mathcal{L}}$ to give the inequality

$$\underbrace{\left(\int_\Omega (\widehat{a} - a)^2\, \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n\right)}_{V[\widehat{a}]} \underbrace{\left(\int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2 \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n\right)}_{I(a)} \geq 1\,. \tag{30}$$

The first factor is identical to the definition of the variance of the estimate $\widehat{a}$, and the second factor is the information as defined by equation (21), and thus the inequality (30) is identical to the information equality (22).

---

[3]Here and in other equations use is made of the logarithmic derivative with $\partial f/\partial a = f \cdot \partial \ln f/\partial a$. This allows to interprete integrals as expectation values.

The information $I(a)$ can be rewritten in terms of the second derivative of the likelihood function. Differentiating equation (27) with respect to the parameter a yields

$$\int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a} \frac{\partial \mathcal{L}}{\partial a} + \frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \mathcal{L} \right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 0 \, , \tag{31}$$

which can be written in terms of expectation values:

$$I(a) = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \right] \tag{32}$$

and both expectation values are identical to the information $I(a)$, defined in equation (21).

Biased estimates. Estimates may be biased and in this case the information inequality has to be modified to

$$V\left[\widehat{a}\right] \geq \frac{\left(1 + \partial B/\partial a\right)^2}{I} \, . \tag{33}$$

Efficiency of an estimator. The actual variance of an estimator $\widehat{a}$ can be compared to the lower limit given by the information inequality. For an unbiased estimator $\hat{a}$ one can define the *efficiency* $\mathrm{eff}(\widehat{a})$ by the ratio of the mininal to the actual variance:

$$\mathrm{eff}(\widehat{a}) = \frac{I^{-1}}{V[\widehat{a}]} \qquad\qquad 0 \leq \mathrm{eff}(\widehat{a}) \leq 1 \tag{34}$$

The actual efficiency of a maximum likelihood estimator depends on the specific problem, for example on the density function $p(x, a)$ and on the choice of the parameter (the efficiency for the parameter a and for the transformed parameter $b = b(a)$ may be different).

2.2 Experimental design

Suppose an experiment is planned which allows the measurement of a set of parameters. The information $\boldsymbol{I}$ as defined before can be used to design the experiment in order to get a maximum of information about the parameters, i.e. to get the most precise measurement of the parameters. The information can be used in different ways:

- In a Monte Carlo simulation many samples $\{\boldsymbol{x}\}$ of a certain size n can be generated and used for a Maximum Likelihood determination of the parameters. Both the parameter values and their covariance matrix in the simulations can be studied as well as the distribution of the parameter values in the simulations.
- The formulae for the information can be directly evaluated.

In the latter case the information is directly related to the assumed density $p(\boldsymbol{x}|\boldsymbol{a})$.
If the set $\{x_1, \ldots, x_i, \ldots, x_n\}$ constitutes an independent sample from a distribution, the likelihood function $\mathcal{L}(a)$ is of the form of equation (18) and the the information provided by the set of n elements is n times the information provided by a single element.

$$I_n(a) = n \cdot E\left[\left(\frac{\partial \ln p(x,a)}{\partial a} \right)^2 \right] = n \cdot \int_\Omega \left(\frac{\partial \ln p(x,a)}{\partial a} \right)^2 p(x,a) \; \mathrm{d}x \tag{35}$$

$$= -n \cdot E\left[\frac{\partial^2 \ln p(x,a)}{\partial a^2} \right] \tag{36}$$

Example 5 Decay angular distribution of an elementary particle II.

The decay angular distribution is given by

$$p(x|a) = \frac{1}{2} \left(1 + ax\right) \qquad\qquad \text{with} \quad x \equiv \cos\vartheta \tag{37}$$

The value of the parameter a is expected to be around $a = -1/3$. The question is, how many decays have to be observed to determine the parameter a with a relative precision of 1 %.

The information is calculated using equation (36).

$$I_1 = \int_{-1}^{+1} \frac{1}{p(x,a)} \left(\frac{\partial p}{\partial a}\right)^2 \mathrm{d}x = \begin{cases} \frac{1}{2a^3}\left[\ln\frac{1+a}{1-a} - 2a\right] & \text{if} \quad a \neq 0 \\ \frac{1}{3} & \text{if} \quad a = 0 \end{cases} \tag{38}$$

The standard deviation depends on the value of the parameter. For $a = -1/3$ the value is $\sigma = \sqrt{2.797/n}$ and it follows that more than 2.52×10^5 decays have to be measured in order to obtain a standard deviation $\sigma < 0.01\,a$.

2.3 The case of several variables

In the case of m variables $a_1, \ldots, a_j, \ldots, a_m$ an information matrix $\boldsymbol{I}$ is defined with matrix elements in analogy to the one-dimensional case (21)

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right] \ . \tag{39}$$

The minimal variance of an estimate, expressed in terms of the elements of the covariance matrix of the parameters $a_1, \ldots, a_j, \ldots, a_m$, is then, according to a generalization of the information inequality, given by the inverse of the information matrix $\boldsymbol{I}$:

$$\boldsymbol{V} = \boldsymbol{I}^{-1} \tag{40}$$

Example 6 Weighing four object with a balance.

The direct approach of weighing four objects $a_j, j = 1, \ldots 4$ is to weigh each object separately in four independent measurements. Alternatively one can try to improve the precision of four measurements by weighing suitable combinations of the objects instead. It is assumed that the measurements follow a normal distribution law and that the standard deviation σ in each measurement of the balance is always the same. In a combination the measured quantities are $x_i, i = 1, \ldots 4$, with

$$x_i = \sum_{j=1}^{4} A_{ij} a_j \tag{41}$$

with $A_{ij} = -1$, 0 or $+1$, and the likelihood function, assuming normal distributions, is

$$\mathcal{L}(\boldsymbol{a}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^4 \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{4}\left(x_i - \sum_{j=1}^{4} A_{ij} a_j\right)^2\right] \tag{42}$$

The elements of the information matrix are given by

$$I_{jk} = E\left[-\frac{\partial^2 \mathcal{L}}{\partial a_j \partial a_k}\right] = \frac{1}{\sigma^2}\sum_{i=1}^{4} A_{ij} A_{ik} \qquad \text{or short} \quad \boldsymbol{I} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A} \tag{43}$$

Different combinations can now be compared. For the direct measurement matrix $\boldsymbol{A}$ is the unit matrix ($A_{ij} = \delta_{ij}$) and $\boldsymbol{I} = 1/\sigma^2\,\mathbf{1}$ and $V = \sigma^2\,\mathbf{1}$ follow, i.e. the objects a_j are measured with standard deviation σ.
Another possibility is the linearly independent combination given by the matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{I} = \frac{4}{\sigma^2}\mathbf{1} \qquad \boldsymbol{A}(\widehat{a}) = \frac{\sigma^2}{4}\mathbf{1} \tag{44}$$

Obviously this method is preferable since the objects a_j are now measured with standard deviation $\sigma/2$. In both cases the measured values are uncorrelated.

3 Properties of the maximum-likelihood method

The properties of the maximum-likelihood method are discussed for simplicity for the case of one variable parameter a. At the end of the chapter general formulas for the case of n parameters are given. The following statements about the properties of the maximum-likelihood method assume, that the true distribution function $p(x|a)$ is known and is used for the evaluation of the likelihood function.
The maximum-likelihood estimate $\widehat{a}$ for the parameter has the following properties, which are discussed in more detail in the following chapters:

Consistency: The estimate $\widehat{a}$ of the MLM is asymptotically $(n \to \infty)$ consistent, i.e.

$$\lim_{n\to\infty} \widehat{a} = a_{\text{true}} \qquad \text{or} \quad \lim_{n\to\infty} (\widehat{a} - a_{\text{true}}) = 0 \tag{45}$$

Normality: The estimate $\widehat{a}$ is, under very general conditions, asymptotical normally distributed with minimal variance $V(\widehat{a})$:

$$\lim_{n\to\infty} V[\widehat{a}] = \left\{ E\left[\frac{\partial \ln p}{\partial a}\right]^2 \right\}^{-1} \quad . \tag{46}$$

Asymptically the likelihood equation becomes a function, which is *linear* in the parameter a.

Invariance: The estimate $\widehat{b}$ of a function $b = b(a)$ is given by $\widehat{b} = b(\widehat{a})$, or in other words: The maximum likelihood solution is invariant under change of parameter. For *finite* values of n there may a bias $\propto 1/n$, which is different for different functions of the parameter (see Example 5).

Efficiency The efficiency of an estimator is defined in relation to the *information* I; if efficient estimators exist for a given problem the maximum leikelihood method will find them.

In the discussion use is made of the law of large numbers and of the central limit theorem (see Glossary).

3.1 Consistency

The estimate $\widehat{a}$ is obtained from the solution of the equation

$$\frac{\partial F}{\partial a} = \sum_{i=1}^{n} \frac{\partial \ln p(x_i; a)}{\partial a} = 0 \,, \tag{47}$$

or by finding the root of the function $Q(a)$, i.e.

$$Q(a) = 0 \qquad \text{with} \quad Q(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln p(x_i; a)}{\partial a} \tag{48}$$

The function $Q(a)$ depends on the measured values x_i, which are random variables. The law of large numbers states, that for each function $y(x)$ of a random variable x with probability density $q(x)$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} y(x) = E[y(x)] = \langle y \rangle = \int y(x) q(x) \, \mathrm{d}x \tag{49}$$

is valid; the mean of a function of a random variable converges in the limit $n \to \infty$ to the expectation value. It follows that

$$\lim_{n\to\infty} Q(a_{\text{true}}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln p(x_i|a)}{\partial a} = E\left[\frac{\partial \ln p}{\partial a}\right] \tag{50}$$

The latter expectation value of $\partial \ln p/\partial a = 1/p \, \partial p/\partial a$ vanishes:

$$E\left[\frac{\partial \ln p}{\partial a}\right] = \int_\Omega \frac{1}{p} \frac{\partial p}{\partial a} p \, \mathrm{d}x = \frac{\partial}{\partial a} \int_\Omega p \, \mathrm{d}x = 0 \tag{51}$$

since the integral is a constant. It follows $Q(a_{\text{true}}) \to 0$ for $n \to \infty$. Because the estimate $\widehat{a}$ is obtained as solution of the equation $Q(a) = 0$, the limit $\widehat{a} \to a_{\text{true}}$ for $n \to \infty$ follows.

3.2 Asymptotic normality

The function $Q(a)$ defined in equation (48) is expanded around the true value a_{true}:

$$Q(a) = Q(a_{\text{true}}) + (a - a_{\text{true}}) \left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} + \dots \tag{52}$$

Asymptotically, quadratic and higher terms can be neglected, and the expression

$$\widehat{a} - a_{\text{true}} = \frac{Q(a_{\text{true}})}{(\partial Q/\partial a)_{a=a_{\text{true}}}} \tag{53}$$

becomes valid for the estimate $\widehat{a}$, defined by $Q(a) = 0$. As shown below the denominator of this expression approaches a constant and thus the variance of $\widehat{a}$ will be given by the variance of the nominator $Q(a_{\text{true}})$.

The denominator is the derivative of the function $Q(a)$ with respect to the parameter a:

$$\frac{\partial Q}{\partial a} = \frac{1}{n} \frac{\partial}{\partial a} \sum_{i=1}^{n} \frac{\partial \ln p(x_i;a)}{\partial a} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln p(x_i;a)}{\partial a^2} \tag{54}$$

Applying the law of large numbers to the above mean value the result

$$\lim_{n \to \infty} \left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} = E\left[\frac{\partial^2 \ln p}{\partial a^2} \right] \tag{55}$$

follows. The calculation is as follows:

$$\begin{aligned} \frac{\partial^2 \ln p}{\partial a^2} &= \frac{\partial}{\partial a}\left(\frac{\partial \ln p}{\partial a} \right) = \frac{\partial}{\partial a}\left(\frac{1}{p}\frac{\partial p}{\partial a} \right) = \frac{1}{p}\frac{\partial^2 p}{\partial a^2} - \left(\frac{\partial \ln p}{\partial a} \right)^2 \\ E\left[\frac{\partial^2 \ln p}{\partial a^2} \right] &= E\left[\frac{1}{p}\frac{\partial^2 p}{\partial a^2} \right] - E\left[\left(\frac{\partial \ln p}{\partial a} \right)^2 \right] \end{aligned}$$

The first term on the right-hand side vanishes:

$$E\left[\frac{1}{p}\frac{\partial^2 p}{\partial a^2} \right] = \int_\Omega \frac{1}{p}\frac{\partial^2 p}{\partial a^2} p \,\mathrm{d}x = \frac{\partial^2}{\partial a^2} \int_\Omega p \,\mathrm{d}x = 0 \tag{56}$$

because of the normalization of the density p. Thus the second term on the right-hand side and the expectation value of the slope of the function Q is a *negative* constant:

$$E\left[\left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} \right] = E\left[\frac{\partial^2 \ln p}{\partial a^2} \right] = -E\left[\left(\frac{\partial \ln p}{\partial a} \right)^2 \right] . \tag{57}$$

The central limit theorem states, that for each function $y(x)$ of a random variable x with density $q(x)$ the sample mean

$$\frac{1}{n} \sum_{i=1}^{n} y(x_i) \tag{58}$$

is asymptotically normal distributed with mean value and variance

$$E\left[y(x)\right] = \int_\Omega y(x) q(x) \,\mathrm{d}x = \langle y \rangle \qquad \frac{1}{n} V\left[y(x)\right] = \frac{1}{n} \int_\Omega \left(y(x) - \langle y \rangle\right)^2 q(x) \,\mathrm{d}x . \tag{59}$$

Thus asymptotically the variance of the nominator $Q(a_{\text{true}})$ (definition in equation (48)) becomes

$$\lim_{n \to \infty} V\left[Q(a_{\text{true}})\right] = \frac{1}{n} V\left[\frac{\partial \ln p}{\partial a} \right] = \frac{1}{n} E\left[\left(\frac{\partial \ln p}{\partial a} \right)^2 \right] = \frac{1}{n} E\left[\left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} \right] . \tag{60}$$

according to equation (57). The variance of the estimate $\widehat{a}$ is obtained according to the law of error propagation by

$$V(\widehat{a}) = \left\{ E\left[\left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} \right] \right\}^{-1} \cdot V\left[Q(a_{\text{true}})\right] \cdot \left\{ E\left[\left(\frac{\partial Q}{\partial a} \right)_{a=a_{\text{true}}} \right] \right\}^{-1} \tag{61}$$

with the result:

$$V(\widehat{a}) = \frac{1}{n}\left\{E\left[\left(\frac{\partial \ln p}{\partial a}\right)^2\right]\right\}^{-1} = \left\{E\left[n\left(\frac{\partial \ln p}{\partial a}\right)^2\right]\right\}^{-1} = \left\{E\left[n\left(\frac{\partial^2 \ln p}{\partial a^2}\right)\right]\right\}^{-1} \tag{62}$$

This is the asymptotic variance of the estimate; the *normality* of the estimate follows from the central limit theorem. The formula shows, that the variance of the estimate $\widehat{a}$ is determined by the property of the density $p(x|a)$ and includes a factor $1/n$ from the number of observations. Thus the asymptotic variance can be calculated before a measurement is done. Especially the number of observations required for a certain accuracy can be determined before the actual measurement.

If the measurement is done the empirical negative log-likelihood function $F(a)$ can be calculated from the data, and the variance of the estimate is usually calculated, as given by the last expression of equation (62), from the second derivative of this function:

$$V(\widehat{a}) = \left[\left(\frac{\partial^2 F}{\partial a^2}\right)_{a=\widehat{a}}\right]^{-1} \tag{63}$$

In the asymptotic limit the second derivative of F is a constant, all higher derivatives vanish.

3.3 Invariance

The expression $p(x|a)dx$ is a probability density in the sample space. Maximum-likelihood estimators possess a property called *invariance*. Let $\widehat{a}$ be the maximum-likelihood estimator of the parameter a in the density $p(x|a)$. If $b(a)$ is a unique function of a (single-valued inverse), then the maximum-likelihood estimator of b is $b(\widehat{a})$. This property corresponds to the fact, that the likelihood function is *not* a density; if it would be a density, it would get an additional Jacobi factor in a transformation, which could affect the position of the maximum of the likelihood function $\mathcal{L}(a)$.

For finite values of n the maximum likelihood estimate may have a bias proportional to $1/n$. A possible bias has to be determined in each case from the behaviour of the estimate $\widehat{a} = \widehat{a}(x_1, \ldots, x_i, \ldots, x_n)$, and the bias will be different for different functions of the parameters (see example 5).

Example 7 Parameter of an exponential distribution

The decay times of unstable particles follow an exponential probability density distribution

$$p(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$$

with the mean lifetime τ as the only parameters. The negative log-likelihood function for given data $t_1, \ldots, t_i, \ldots, t_n$ is

$$F(\tau) = -\sum_{i=1}^{n} \ln p(t_i;\tau) = \sum_{i=1}^{n}\left(\frac{t_i}{\tau} - \ln\frac{1}{\tau}\right) .$$

Minimizing the negative log-likelihood function with respect to τ give the maximum likelihood estimator $\widehat{\tau}$

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i ,$$

which is equal to the sample mean. This is an *unbiased* estimate for the parameter τ:

$$\begin{aligned}
E\left[\widehat{\tau}(t_1,\ldots,t_i,\ldots,t_n)\right] &= \int\cdots\int \widehat{\tau}(t_1,\ldots,t_i,\ldots,t_n)\, p(t_1,\tau)\cdots p(t_n,\tau)\,\mathrm{d}t_1\cdots\mathrm{d}t_n \\
&= \int\cdots\int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)\frac{1}{\tau}e^{-t_1/\tau}\ldots\frac{1}{\tau}e^{-t_n/\tau}\,\mathrm{d}t_1\cdots\mathrm{d}t_n \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\int\frac{t_i}{\tau}e^{-t_i/\tau}\,\mathrm{d}t_i\prod_{j\neq i}\frac{1}{\tau}e^{-t_j/\tau}\,\mathrm{d}t_j\right) = \frac{1}{n}\sum_{i=1}^{n}\tau = \tau .
\end{aligned}$$

Instead of the parameter τ one could also derive an estimate for the decay constant $\lambda = 1/\tau$. From the invariance property of maximum likelihood estimates the estimator $\widehat{\lambda} = 1/\widehat{\tau}$, or

$$\widehat{\lambda}(t_1, \ldots, t_i, \ldots, t_n) = \frac{n}{\sum_{i=1}^{n} t_i}$$

follows with the expectation value

$$E\left[\widehat{\lambda}\right] = \frac{n}{n-1}\lambda = \frac{n}{n-1}\frac{1}{\tau} \, .$$

Thus the estimator $\widehat{\lambda} = 1/\widehat{\tau}$ has a bias for finite n; the bias vanishes in the asympotic limit.

4 Application of the maximum-likelihood method

In practice the application of the maximum likelihood method may be complicated due to imperfections of the measurement. The most common problems are:

- limited acceptance,
- finite resolution
- non-negligible background contribution.

In principle there seem to be two possibilities, to take these conditions into account: one could *either* try to *correct* the data *or* one could modify the theoretical distribution according to the properties of the measurement. The latter method will turn out to be the correct choice for the application of the Maximum-Likelihood method. Often it requires to use an individual p.d.f. for each entry x_i in the set of measurements, which may require a lot of computations. However, the possibility to take into account all the details of the measurement process is just the advantage of the Maximum-Likelihood method.

4.1 Limited acceptance

The limited acceptance is described by an acceptance function $A(x)$, which gives the acceptance (probability of observation) at the value x of the argument ($0 \leq A(x) \leq 1$). A limited acceptance is sometimes taken into account by the assignment of a weight $w_i = 1/A(x_i)$ to the data element x_i. If for example $A(x_i) = 0.5$, the measurement x_i would get a weight $w_i = 2$. This method may be acceptable for a histogram of the measured distribution; the weighted histogram is then acceptance-corrected.

In the weighting method one could replace in the likelihood function $f(x_i, \boldsymbol{a})$ by $f(x_i, \boldsymbol{a})^{w_i}$ with the result $F(\mathbf{a}) = -\sum_{i=1}^{n} w_i \ln f(x_i|\mathbf{a})$. Although this procedure seems to be straightforward and also the parameter values from such a modification could come out reasonable, the resulting parameter errors are wrong; the assigment of weights (different from 1) destroys the statistical behaviour of the sample. There are especially problems with large weights (acceptance $A(x_i) \ll 1$). These problems are also visible in a weighted histogram.

The correct way is to modify the expectation and to replace $p(x_i|\boldsymbol{a})$ by the properly normalized probability density $\mathcal{N}(\boldsymbol{a})^{-1} \cdot A(x_i) \cdot p(x_i, \boldsymbol{a})$, with the normalization factor $\mathcal{N}$ defined by

$$\mathcal{N} = \int_{\Omega} A(x) p(x, \boldsymbol{a}) \, \mathrm{d}x \; . \tag{64}$$

In general the normalizing factor will depend on the actual parameter value.

Example 8 Determination of the mean lifetime of an elementary particle from the decay length

Decays of the K^0_S are observed in a detector, which allows to measure the decay length ℓ in the range $\ell_1 = 1$ cm $\ldots \ell_2 = 10$ cm. The acceptance outside this range vanishes and this fact has to be taken into account in a Maximum Likelihood fit for the determination of the mean life time.

According to the exponential decay law the decay time distribution is proportional to $e^{-t/\tau}$, where τ is the mean lifetime. For a measurement, which is sensitive only in the time region $t_1 \ldots t_2$, the p.d.f is correctly normalized by the condition $\int_{t_1}^{t_2} f(t) \, \mathrm{d}t = 1$, resulting in the expression

$$f(t) \, \mathrm{d}t = \frac{e^{-t/\tau}}{\tau \left(e^{-t_1/\tau} - e^{-t_2/\tau} \right)} \, \mathrm{d}t \; , \tag{65}$$

valid in the rest system of the particle (the mean decay time τ is defined in the rest system of the particle). This p.d.f. has to be considered as a function of the decay length ℓ, which is actually measured in the range $\ell_1 \ldots \ell_2$. The relation between the decay time t (in the rest system) and the decay length is given by $\ell = \beta c \gamma t$ with $\beta = v/c$ (v = particle velocity, c = velocity of light), due to the time dilatation for a fast moving particle by the factor $\gamma = 1/\sqrt{1 - \beta^2}$. Using the mass m (in GeV/c^2) and the momentum p (in GeV/c) of the particle, the decay length can be written as $\ell = ctp/m$, and the probability density becomes

$$f(\ell) \, \mathrm{d}\ell = \left(\frac{m}{c\tau \, p} \right) \frac{e^{-\ell \, m/(c\tau \, p)}}{\left(e^{-\ell_1 \, m/(c\tau \, p)} - e^{-\ell_2 \, m/(c\tau \, p)} \right)} \, \mathrm{d}\ell \; , \tag{66}$$

Figure 2 shows the negative log-likehood function (a) versus $c\tau$, the mean life time, multiplied by the velocity of light, in cm. and (b) versus the inverse $1/(c\tau)$, from a sample of 50 values. The minimum is near the expected value of $c\tau = 2.68$ cm.

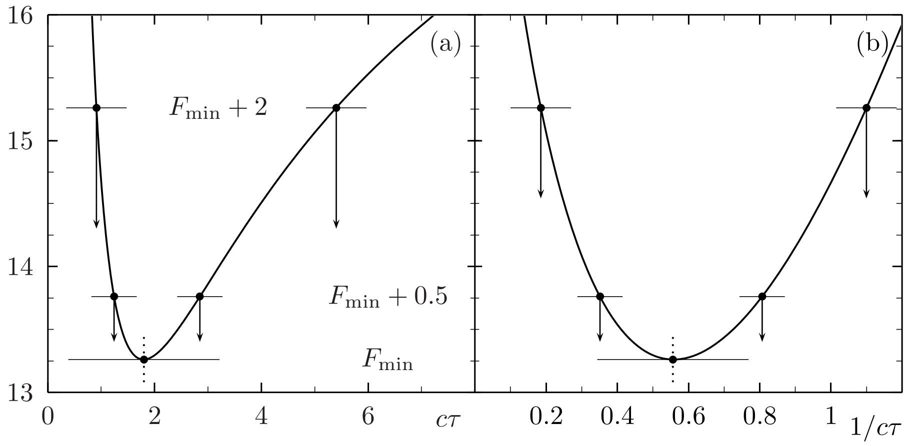

Figure 2: The negative log-likelihood function, shown as a function of the life time $c\tau$ (a) and as function of the inverse life time $1/c\tau$ (b).

5 Gaussian distributions

5.1 Weighted mean of Gaussian data

One-dimensional variables. A series of n measurements $x_i,\ i = 1, 2, \ldots n$ has been made, in which each measurement x_i is known to be distributed according to the Gaussian distribution with *known* standard deviations σ_i, different for each measurement. The probability density distribution is given by

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\dfrac{(x_i-\mu)^2}{2\sigma_i^2}} .$$

In order to determine the maximum likelihood estimate for the mean value of the distribution, the negative log-likelihood function

$$F(\mu) = -\sum_{j=1}^{n} \ln f(x_i) = \sum_{j=1}^{n} \left[\frac{(x_i-\mu)^2}{2\sigma_i^2} + \ln\sqrt{2\pi} + \ln\sigma_i \right] = \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma_i^2} + \text{const}$$

is used. The first and second derivative of $F(\mu)$ w.r.t. the parameter μ are:

$$\frac{\mathrm{d}F}{\mathrm{d}\mu} = -\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma_i^2}\right) = \mu\cdot\sum_{i=1}^{n}\left(\frac{1}{\sigma_i^2}\right) - \sum_{i=1}^{n}\left(\frac{x_i}{\sigma_i^2}\right) \qquad \frac{\mathrm{d}^2F}{\mathrm{d}\mu^2} = \sum_{i=1}^{n}\left(\frac{1}{\sigma_i^2}\right)$$

These expressions show, that the inverse variance can be assumed to be a weight factor $w_i = 1/\sigma_i^2$, which is used in the following equations instead of the inverse variance.

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_j x_i}{\sum_{i=1}^{n} w_i} \qquad V[\widehat{\mu}] = \frac{1}{\sum_{i=1}^{n} w_i}$$

Thus the estimate of the mean is the *weighted* mean value of the single measurements, with weight equal to the inverse variance.

5.2 Mean and variance of a Gaussian

Next the case where both parameters of the Gaussian are estimated from a sample is considered. The parameter a_1 is the mean and the parameter a_2 is the variance of the Gaussian.

$$\mathcal{L}(a_1, a_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi a_2}} e^{-\dfrac{(x_i-a_1)^2}{2a_2}}$$

The negative log-likelihood function is

$$F(a_1, a_2) = -\ln\mathcal{L}(a_1, a_2) = \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-a_1)^2}{a_2} + \frac{n}{2}\ln a_2 + \frac{n}{2}\ln(2\pi)$$

From the first derivatives with respect to the parameters the parameter estimates are obtained:

$$\frac{\partial F}{\partial a_1} = \sum_{i=1}^{n}\frac{1}{a_2}(x_i-a_1) \qquad \frac{\partial F}{\partial a_2} = -\frac{1}{2a_2^2}\sum_{i=1}^{n}(x_i-a_1)^2 + \frac{n}{2a_2}$$

$$\widehat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \widehat{a}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{a}_1)^2$$

The second derivatives are:

$$\frac{\partial^2 F}{\partial a_1^2} = \frac{n}{a_2} \qquad \frac{\partial^2 F}{\partial a_1\partial a_2} = \frac{1}{a_2^2}\sum_{i=1}^{n}(x_i-a_1) \qquad \frac{\partial^2 F}{\partial a_2^2} = \frac{1}{a_2^3}\sum_{i=1}^{n}(x_i-a_1)^2 - \frac{n}{2a_2^2}$$

The estimates $\widehat{a}_1$ and $\widehat{a}_2$ are inserted for a_1 and a_2; the mixed derivatives vanishes and the derivative $\partial^2 F/\partial a_2^2$ can be simplified to $2n/\widehat{a}_2$. Thus the Hessian $\boldsymbol{H}$ and the estimate for the covariance matrix $\boldsymbol{V}$ are:

$$\boldsymbol{H} = \begin{pmatrix} n\widehat{a}_2 & 0 \\ 0 & n/(2\widehat{a}_2) \end{pmatrix} \qquad \boldsymbol{V} = \begin{pmatrix} \widehat{a}_2/n & 0 \\ 0 & 2\widehat{a}_2/n \end{pmatrix}$$

The two estimates $\widehat{a}_1$ and $\widehat{a}_2$ are uncorrelated, since the off-diagonal element of matrices $\boldsymbol{H}$ and $\boldsymbol{V}$ vanishes.

The estimate $\widehat{a}_1$ of parameter a_1 is identical to the sample mean $\overline{x}$; the sample mean is an unbiased estimate of the mean μ of the distribution (→ Sample mean and variance). The expectation value for the second parameter, the variance, is however

$$E\left[\widehat{a}_2\right] = \frac{n-1}{n}\,\sigma^2$$

and thus differs from the population variance. By applying the factor $n/(n-1)$ the estimate can be corrected[4] to

$$\widehat{a}_2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{a}_1\right)^2$$

which is consistent and free of bias, and is equal to the sample variance s^2 (→ Sample mean and variance).

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{a}_1\right)^2 \left(1 \pm \sqrt{\frac{2}{n}}\right)$$

For the standard deviation $\widehat{\sigma} = s$ is a consistent estimator; although s^2 is an unbiased estimate of σ^2, the estimate $\widehat{\sigma} = s$ is not unbiased for finite n.

$$V\left[\widehat{\sigma^2}\right] = \left(\frac{\mathrm{d}\sigma^2}{\mathrm{d}\sigma}\right)^2 V\left[\widehat{\sigma}\right] = 4\sigma^2 V\left[\widehat{\sigma}\right]$$

from which

$$V\left[\widehat{\sigma}\right] = \frac{\sigma^2}{2n}$$

follows. The error of $\widehat{\sigma}$ is sometimes called the error of the error:

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{a}_1\right)^2}\left(1 \pm \frac{1}{\sqrt{2n}}\right) .$$

[4]This factor $n/(n-1)$ is known as Bessel's correction.

6 Binned distributions

6.1 The problem

As the result of measurements there is a sample of real data, each member of which consists of a set of values $\{\boldsymbol{x}\}$. The data are binned, dividing the one- or multi-dimensional space of the $\{\boldsymbol{x}\}$ into n bins. This subdivision gives a set of numbers $\{d_1, d_2, \ldots, d_n\}$, where d_i is the number of events in the real data that fall into bin i. The values d_i are integers 0, 1, ...

$$d_i = \text{number of events in the real data that fall into bin } i \tag{67}$$

$$N_D = \sum_{i=1}^{n} d_i = \text{total number in the data sample} \tag{68}$$

The real data arise from a number of sources (or physical processes) and the aim is to determine the proportions P_j of the different sources in the data from the statistical data d_i and from models for the sources. One has to distinguish the case, where analytic forms are available for the distribution of the sources and where no analytic forms are available.

Models with analytic forms. Analytic forms available for the distribution of the sources can be used to calculate, by integration over the bins, numbers a_{ji} proportional to the expected number of events from source j in bin i; these values have no statistical errors.

$$P_j = \text{proportion of source } j \text{ in the data} \tag{69}$$

$$N_j = \sum_{i=1}^{n} a_{ji} \tag{70}$$

$$a_{ji}/N_j = \text{fraction of events from source } j \text{ expected in bin } i \tag{71}$$

The predicted number of events in bin i is, for the proportions P_j, given by the sum

$$f_i = N_D \sum_{j=1}^{m} P_j \, \frac{a_{ji}}{N_j} \,, \tag{72}$$

where the proportions should sum to unity. It may be convenient to use instead the equivalent form

$$f_i = \sum_{j=1}^{m} p_j \, a_{ji} \qquad \text{with} \quad p_j = N_D \frac{P_j}{N_j} \tag{73}$$

with strength factors p_j.

A standard method to estimate the factors p_j is the χ^2-minimization, where the p_j are adjusted to minimize

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(d_i - f_i)^2}{d_i}} \,, \tag{74}$$

assuming that the distribution for d_i is Gaussian with a standard deviation of $\sqrt{d_i}$. The distribution for d_i in reality is Poisson, but the Gaussian is assumed to be a good approximation to the Poisson distribution for large numbers, say $d_i > 10$. This χ^2 method corresponds to a linear least square fit and can be solved directly without iterations and without the need for initial values of the factors p_j. In the approximation of the Poisson distribution by a Gaussian, values d_i fluctuating down will get a smaller absolute standard deviation and thus a larger weight, compared to values d_i fluctuating up. For smaller values d_i the result will be biased and the fit is distorted, but the parameter values obtained may still be useful as starting values for a nonlinear fit

A better approximation of the Poisson distribution is obtained by replacing d_i in the denominator by f_i; however then the minimization requires iterations and initial values.

Often some or many of the data d_i are small or even zero; this is especially the case for multidimensional data, where even a modest subdivision of the axes will give a large total number n of bins. The χ^2 minimization will no longer be appropriate.

Instead the correct statistical distribution, the Poisson distribution, has to be used. If the mean value in a bin is f, then the observed values $d = 0, 1, \ldots$ follow the Poisson distribution

$$P_f(d) = e^{-f} \frac{f^d}{d!} \qquad \text{with normalization} \quad \sum_{d=0}^{\infty} P_f(d) = 1 \; . \tag{75}$$

The variance (or square of the standard deviation) is equal to the mean value f. Estimates of the factors are found by maximizing the total likelihood

$$\mathcal{L}(\boldsymbol{p}) = \prod_{i=1}^{n} P_{f_i}(d_i) = \prod_{i=1}^{n} e^{-f_i} \frac{f^{d_i}}{d_i!} \tag{76}$$

with respect to the parameters $\boldsymbol{p}$ or equivalently, by minimizing the negative logarithm of the likelihood $F(\boldsymbol{p}) = -\ln \mathcal{L}(\boldsymbol{p})$:

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} (f_i - d_i \cdot \ln f_i)} \tag{77}$$

omitting constant factors in the product like $1/d_i!$. This expression correctly accounts for small numbers of data events d_i in a bin or even zero data events in some bins. The method is called *binned maximum likelihood* fit.

Constant factors in the Likelihood product can be omitted or added, because they do no influence the position of the minimum. There is some advantage in redefining the function F to be minimized by adding some constants as

$$\boxed{F(\boldsymbol{p}) = \sum_{i=1}^{n} g_i \qquad \text{with} \quad g_i = \begin{cases} (f_i - d_i) - d_i \cdot \ln(f_i/d_i) & \text{if } d_i > 0 \\ f_i & \text{if } d_i = 0 \end{cases}} \; . \tag{78}$$

The position of the minimum and the shape of F is not changed by his modification, but now the minimum value of F can be interpreted, as shown below. Considering one term of the sum only and assuming $d_i > 0$ the approximation

$$\begin{aligned} (f_i - d_i) - d_i \cdot \ln \frac{f_i}{d_i} &= (f_i - d_i) - d_i \cdot \ln\left(1 + \frac{f_i - d_i}{d_i}\right) \\ &\approx (f_i - d_i) - d_i \left(\frac{(f_i - d_i)}{d_i} - \frac{1}{2}\frac{(f_i - d_i)^2}{d_i^2}\right) = \frac{1}{2}\frac{(f_i - d_i)^2}{d_i} \end{aligned} \tag{79}$$

based on the expansion $\ln(1 + \varepsilon) \approx \varepsilon - \varepsilon^2/2$ can be done, which shows, that the value F is (within this approximation) identical to $\chi^2/2$. Thus, as in least squares methods, the value of the function can be interpreted in terms of goodness-of-fit χ^2.

$$g_j = \sum_i \left(1 - \frac{d_i}{f_i}\right) a_{ji} \tag{80}$$

$$H_{jk} = \sum_i \frac{d_i}{f_i^2} a_{ji} a_{ki} \tag{81}$$

Extended maximum-likelihood method. As mentioned above the Poisson version of the negative log-likelihood function is valid even if the measured bin contents are small or even zero. Thus the binning can be made very narrow. It is interesting to consider the limiting case of a very small bin width Δx and a large number of bins, such that the bins contents d_i are either zero or one. In this case the negative log-likelihood function becomes

$$F(\boldsymbol{p}) = -\sum_{i=1}^{N} \ln f(x_i|a)\Delta x + \sum_{j=1}^{\text{bins}} f(x_j|a)\Delta x = -\sum_{i=1}^{N} \ln f(x_i|a)\Delta x + \int_\Omega f(x_j|a)\mathrm{d}x \tag{82}$$

Here the integral (last term) is the equivalent to the number of cases N_{exp} expected from then function $f(x|a)$.

A comparison with equation (77) shows, that here the term with the integral is added to the previous equation, which was valid for *normalized* functions $f(x|a)$. The integral term represent the number of entries (sample size), expected for the given function. The use of the equation represent the so-called *extended* maximum-likelihood method; it works with unnormalized functions and introduces, by the variable integral over the function, which represents the expected number of entries, one additional parameter.

Models with no analytic forms. Often no analytic forms are available for the distributions of the sources. Instead there are samples of data, generated according to the model of the source by Monte Carlo simulation, (and perhaps measured background samples). These data samples can be binned in the same way as the real data, giving a set of integer numbers $\{a_{j1}, a_{j2}, \ldots a_{jn}\}$ for source j. Now both the real data d_i and the data a_{ji} are of statistical nature with integer values 0, 1, ...

$$a_{ji} = \text{number of Monte Carlo events from source } j \text{ in bin } i \tag{83}$$

$$N_j = \sum_{i=1}^{n} a_{ji} = \text{total number in the MC sample for source } j \tag{84}$$

$$\tag{85}$$

The Monte Carlo samples are finite, leading to statistical fluctuations in the numbers a_{ji}. If the Monte Carlo samples are much larger than the data sample, one may ignore these fluctuations and use

$$f_i = \sum_{j=1}^{m} p_j \, a_{ji} \tag{86}$$

as before; the fluctuations in the a_{ji} are damped by the factor N_D/N_j. Usually it is assumed that a factor of 10 in the size of the Monte Carlo samples compared to the date sample should be large enough. However often the statistical fluctuations in the a_{ij} of the Monte Carlo sample can not be ignored, and one has to consider them together with the statistical fluctuations of the data d_i. A method to treat the problem within the maximum-likelihood method has been developed by R. Barlow and the following derivation follows closely the paper by Barlow. In this method there is for each source, in each bin, some (unknown) expected number of events A_{ji}

$$f_i = \sum_{j=1}^{m} p_j A_{ji} \; . \tag{87}$$

From each A_{ji} the corresponding a_{ji} is generated by a distribution which can be taken as Poisson. The total Likelihood is the combined probability of the observed $\{d_i\}$ and of the observed $\{A_{ji}\}$ and one has have to minimize

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n} [f_i - d_i \cdot \ln f_i] + \sum_{i=1}^{n} \sum_{j=1}^{m} [A_{ji} - a_{ji} \cdot \ln A_{ji}] \tag{88}$$

or, more complete, showing directly the dependence on the strength factors p_j and the expected numbers of events A_{ji}:

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n} \left[\sum_{j=1}^{m} p_j A_{ji} - d_i \cdot \ln \sum_{j=1}^{m} p_j A_{ji} \right] + \sum_{i=1}^{n} \sum_{j=1}^{m} [A_{ji} - a_{ji} \cdot \ln A_{ji}] \; . \tag{89}$$

Estimates for the p_j and the A_{ji} are found by minimising the function F. The interest is of course in the determination of the p_j, while the final values of the A_{ji} are not required. This is a minimisation problem in $m + m \times n$ unknowns, which is a rather large number, compared to the n observed data values $\{d_1, d_2, \ldots, d_n\}$. Conditions for the minimum are obtained by the fact, that the derivatives of F with respect to the strength factors p_j and the event numbers A_{ji} have to be zero:

$$\sum_{i=1}^{n} \left[A_{ji} - \frac{d_i A_{ji}}{f_i} \right] = 0 \qquad j = 1, 2, \ldots m \tag{90}$$

$$1 - \frac{a_{ji}}{A_{ji}} + p_j - \frac{d_i \, p_j}{f_i} = 0 \qquad j = 1, 2, \ldots m \qquad i = 1, 2, \ldots n \tag{91}$$

Thus one has to solve a system of $m + m \times n$ simultaneous, nonlinear equations for the $m + m \times n$ unknowns.

In the original paper by BARLOW the equations (91) are rewritten in the form

$$1 - \frac{d_i}{f_i} = \frac{1}{p_j}\left(\frac{a_{ji}}{A_{ji}} - 1\right) \qquad j = 1, 2, \ldots m \qquad i = 1, 2, \ldots n \tag{92}$$

The left hand side depends on index i only, so it can be written as

$$t_i = 1 - \frac{d_i}{f_i} \tag{93}$$

and one can express A_{ji} in the form

$$A_{ji} = \frac{a_{ji}}{1 + p_j\ t_i}\ , \tag{94}$$

which simplifies the problem. For a given set of p_j the $m \times n$ unknowns A_{ji} are given by the n unknowns t_i, which are defined above. If $d_i = 0$ then $t_i = 1$, and if not, then

$$\frac{d_i}{1 - t_i} = f_i = \sum_{j=1}^{m} p_j A_{ji} = \sum_{j=1}^{m} p_j \frac{a_{ji}}{1 + p_j\ t_i} \tag{95}$$

The equations are *not coupled* and the values t_i are determined independently for each bin i. This is a great simplification: instead of $m \times n$ unknowns A_{ji} the are only n unknowns t_i. Since, for a given bin i, all the m values A_{ji} are combined to one value f_i, which has to be compared to one measured value d_i, it is clear that this reduction of the number of parameters per bin from m to 1 has to be possible.

The following treatment is slightly different from BARLOWs treatment. The t_i are considered as the essential parameters, not the A_{ji}, which are not used in the following, but eliminated using the relation (94). The function (88) to be minimised is written in the form

$$F = -\ln \mathcal{L} = \sum_{i=1}^{n}\left[\left(\sum_{j=1}^{m}\frac{p_j a_{ji}}{1 + p_j t_i}\right) - d_i \cdot \ln\left(\sum_{j=1}^{m}\frac{p_j a_{ji}}{1 + p_j t_i}\right)\right] + \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{a_{ji}}{1 + p_j t_i} - a_{ji} \cdot \ln\frac{a_{ji}}{1 + p_j t_i}\right] \tag{96}$$

with the m unknowns p_j and the n unknowns t_i.

6.2 Fitting using finite Monte Carlo samples

$$F_1 = \sum_{i=1}^{n}\left[d_i \cdot \ln f_i - f_i\right] = \sum_{i=1}^{n}\left[d_i \cdot \ln\left(\sum_{j=1}^{m}\frac{p_j a_{ji}}{1 + p_j t_i}\right) - \sum_{j=1}^{m}\frac{p_j a_{ji}}{1 + p_j t_i}\right]$$

$$F_2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{m}\left(a_{ji} \cdot \ln\left(\frac{a_{ji}}{1 + p_j t_i}\right) - \frac{a_{ji}}{1 + p_j t_i}\right)\right]$$

$$f_i \equiv S_1 = \sum_{j=1}^{m}\frac{p_j a_{ji}}{1 + p_j t_i} \qquad S_2 = \sum_{j=1}^{m}\frac{p_j^2 a_{ji}}{(1 + p_j t_i)^2}$$

$$S_3 = \sum_{j=1}^{m}\frac{p_j^3 a_{ji}}{(1 + p_j t_i)^3} \qquad S_4 = \sum_{j=1}^{m}\frac{p_j^2 a_{ji}}{(1 + p_j t_i)^3}$$

$$S_5 = \sum_{j=1}^{m}\frac{p_j a_{ji}}{(1 + p_j t_i)^2}$$

$$\frac{\partial F}{\partial t_i} = S_2\left(1 - \frac{d_i}{S_1}\right) - S_1 + S_5$$

$$\frac{\partial^2 F}{\partial t_i^2} = 2S_3\left(-1+\frac{d_i}{S_1}\right) - d_i\left(\frac{S_2}{S_1}\right)^2 - 2S_4 + S_2$$

$$\frac{\partial F_2}{\partial p_j} = -\frac{a_{ji}p_j t_i^2}{\left(1+p_j t_i\right)^2}$$

$$\frac{\partial^2 F_2}{\partial p_j \partial t_i} = -\frac{2a_{ji}p_j t_i}{\left(1+p_j t_i\right)^3}$$

$$\frac{\partial^2 F_2}{\partial p_j^2} = a_{ji}t_i^2\left(\frac{p_j t_i - 1}{\left(1+p_j t_i\right)^3}\right)$$

$$\frac{\partial F_1}{\partial p_j} = \frac{a_{ji}}{\left(1+p_j t_i\right)^2}\left[\frac{d_i}{S_1} - 1\right]$$

$$\frac{\partial^2 F_1}{\partial p_j \partial p_k} = -\frac{d_i}{S_1^2}\frac{a_{ji}}{\left(1+p_j t_i\right)^2}\cdot\frac{a_{ki}}{\left(1+p_k t_i\right)^2}$$

$$\frac{\partial^2 F_1}{\partial p_j^2} = \frac{2a_{ji}t_i}{\left(1+p_j t_i\right)^3}\cdot\left(1-\frac{d_i}{S_1}\right) - \frac{d_i}{S_1^2}\left(\frac{a_{ji}}{\left(1+p_j t_i\right)^2}\right)^2$$

$$\frac{\partial^2 F_1}{\partial p_j \partial t_i} = \frac{2a_{ji}p_j}{\left(1+p_j t_i\right)^3}\left(1-\frac{d_i}{S_1}\right) + \frac{d_i a_{ji}}{S_1^2\left(1+p_j t_i\right)^2}\cdot S_2$$

Glossary

(a) Probability and probability distributions

Bayesian probability. In Bayesian or subjective probability the elements of the sample space correspond to hypotheses. $\mathcal{H}$ is a hypothesis or proposition, which is either true or not true, and the probability $P(\mathcal{H})$ is interpreted as a measure of the *degree of belief*, that $\mathcal{H}$ is true. Bayesian probability is the basis of → Bayesian statistics (see also → Probability).

Bayesian statistics. Basis of the Bayesian statistics is the → Bayesian probability or subjective probability. In the Bayesian statistics Bayes' theorem takes the form of a proportionality

$$P(\mathcal{H}|\text{data}) \propto P(\text{data}|\mathcal{H}) \cdot P(\mathcal{H}) ,$$

where $P(\mathcal{H})$ represents the prior probability of hypothesis $\mathcal{H}$ being true, and $P(\text{data}|\mathcal{H})$ is the likelihood of observing the data under the assumption of the hypothesis $\mathcal{H}$. The posterior probability of the hypothesis $\mathcal{H}$ after having measured the data is $P(\mathcal{H}|\text{data})$.

There is no fundamental rule how to assign the prior probabilities to a hypothesis. Once an assignment has been made, the Bayesian statistics gives the change of the degree of belief after having measured the data (see also → Classical statistics, page 25).

Bayes theorem. Bayes' theorem[5] relates the → conditional probabilities $P(\mathcal{A}|\mathcal{B})$ and $P(\mathcal{B}|\mathcal{A})$. The theorem is derived from the basic probability formula

$$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}|\mathcal{A}) = P(\mathcal{B}) \cdot P(\mathcal{A}|\mathcal{B})$$

and has the following form:

$$\text{Bayes' theorem} \qquad P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A}) \cdot P(\mathcal{A})}{P(\mathcal{B})}$$

For a sample space with disjoint subsets $\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_n$, i.e. $P(\mathcal{A}_i) \cdot P(\mathcal{A}_j) = 0$ for $i \neq j$, the total probability for $\mathcal{B}$ is

$$P(\mathcal{B}) = \sum_i P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)$$

and the Bayes theorem is written in the form

$$P(\mathcal{A}_i|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)}{\sum_i P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)} .$$

Beta distribution. The Beta distribution depends on the positive integers n and m and is nonzero over the interval $0 < x < 1$ of the random variable x:

$$f(x) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} x^{m-1} (1-x)^{n-1} .$$

Bias. The difference between the expectation value of an estimator $\widehat{a}$ (→ estimator, page 26) of a physical parameter a and the true value a_{true} is called bias:

$$B_n(\widehat{a}) = E[\widehat{a}] - a_{\text{true}} .$$

The bias depends on the size n of the sample, used in the estimator. The estimator is called *unbiased* if $E[\widehat{a}] = a_{\text{true}}$ for *all* n. The estimator is called *consistent*, if the bias $B_n(\widehat{a})$ vanishes asymptotically (for $n \to \infty$).

[5] Thomas Bayes, British reverend, 1702 – 1761

Binomial distribution. Consider an experiment which yields the event A ("success") with probability p (and the event $\overline{A}$ ("failure") with probability $q = 1 - p$). The Binomial distribution gives the probability of k successes (or events) in n trials

$$\boxed{P(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0,\, 1,\, 2 \,\ldots\, n} \,.$$

The probability that the event accurs in the first k trials is $p^k(1-p)^{n-k}$; this ordering is only one out of $\binom{n}{r} = n!/\,(k!(n-k)!)$ ways with the same number k of events in n trials. The normalization follows from the Binomial theorem (→ Binomial theorem, page 35). The Binomial distribution is determined by the two parameters n and p, the random variable r has the mean value np and a variance of $\sigma^2 = np(1-p)$.

The maximum likelihood estimator of the parameter p and its variance is, for a sample of n events with k successes, given by

$$\widehat{p} = \frac{k}{n} \qquad\qquad \sigma^2(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n}$$

The latter estimate is biased for finite n; an unbiased estimate of the variance is $\sigma^2(\widehat{p}) = \widehat{p}(1 - \widehat{p})/(n-1)$.

Breit Wigner distribution. The distribution, also called Lorentz distribution, is the Fourier transform of the → exponential distribution with the relation $\tau\,\Gamma = 1$ between the mean value τ of the exponential distributin and the width parameter Γ of the Breit Wigner distribution

$$\boxed{p(x;\, x_0, \Gamma) = \frac{1}{2\pi} \frac{\Gamma}{(x-x_0)^2 + \Gamma^2/4}} \,.$$

The parameter x_0 is the mode of the distribution and Γ is the full width at half maximum (FWHM) (see also → Cauchy distribution)

Cauchy distribution. The Cauchy distribution is symmetric around $x = 0$ and is defined by

$$\boxed{p(x) = \frac{1}{\pi} \frac{1}{1+x^2}} \,.$$

Mean, variance and higher moments of the Cauchy distribution are undefined, because the integrals diverge; the Cauchy distribution is the example of a density with long tails. Mode and the median are zero, the full width at half maximum (FWHM) is two. A more general form of the Cauchy distribution is the → Breit Wigner distribution, page 24.

If a variable φ has an uniform distribution between $-\pi/2$ and $+\pi/2$, then the variable $x = \tan\varphi$ follows the Cauchy distribution. If y and z follow (independent) standard Gaussian distributions $N(0, 1)$, then the ratio $x = y/z$ follows the Cauchy distribution.

χ^2 distribution. If $x_1, \ldots, x_i, \ldots, x_n$ are independently distributed according to the standardized Gaussian distribution $N(0, 1)$, then

$$u = \sum_{i=1}^{n} x_i^2$$

follows a χ^2 distribution with n degrees of freedom. The density is given by

$$f_n\,(u)\;\mathrm{d}u = \frac{\frac{1}{2}\left(\frac{u}{2}\right)^{n/2-1} e^{-u/2}}{\Gamma(n/2)}\,\mathrm{d}u\;.$$

Expectation value and variance are:

$$E[u] = n \qquad\qquad V[u] = 2n$$

The χ^2 distribution approaches the Gaussian distribution with mean n and variance $2n$ in the limit $n \to \infty$. The transformed variable

$$t = \sqrt{2u} - \sqrt{2n-1}$$

approaches the standard Gaussian distribution faster, and this variable can be used in comparing χ^2-distributed values with different values of n.

If $x_1, \ldots, x_i, \ldots, x_n$ are *independently* Gaussian distributed with means μ_i and variances σ_i^2, then the quantity

$$u = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

follows a χ^2 distribution with n degrees of freedom. The generalization to correlated Gaussian variables with a covariance matrix $\boldsymbol{V}$ is written in matrix notation as quadratic form

$$u = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \; .$$

Classical statistics. (→ Bayes statistics, page 23)

Basic tasks of classical statistics are to estimate probabilities, given a finite set of measured data, and to test wether probabilities, predicted by a certain model or theory, are compatible with the measured data. It is based on the interpretation of probability as a relative frequency.

Conditional distribution. Given a joint probability function $p(x_1, x_2)$ of random variables x_1, x_2, the conditional probability density function of x_2 is the probability density function of x_2 for a given fixed value of x_1. The variable x_1 is treated as a constant. The conditional probability density function for the random varibale x_2, given x_1, written in the form $q(x_2|x_1)$, is defined as

$$q(x_2|x_1) = \frac{p(x_1, x_2)}{\int p(x_1, x_2)\, \mathrm{d}x_2}$$

The density is renormalized such that its area is unity when integrated over x_2. The generalization to the case of a joint probability function of more variables is straightforward.

Conditional probability. Consider a sample space with subsets $\mathcal{A}$ and $\mathcal{B}$. The conditional probability of $\mathcal{B}$, given $\mathcal{A}$, is defined as

$$P(\mathcal{B}|\mathcal{A}) = \frac{P(\mathcal{A} \,\mathrm{and}\, \mathcal{B})}{P(\mathcal{A})} \; .$$

Confidence interval. The statistical error of an estimate $\widehat{a}$ can be expressed by a confidence interval for a given → confidence level with probabilities α and β. Suppose that the estimate $\widehat{a}$ has been obtained from data by an estimator $\widehat{a}(x_1, \ldots, x_i, \ldots, x_n$, and that the probability density function $g(\widehat{a}; a)$, depending on the true value a of the parameter, is known. The limits of the confidence interval are called a_{lower} and a_{upper} and can be determined by solving equations

$$\alpha = \int_{\widehat{a}}^{+\infty} g(\widehat{a}; a_{\mathrm{lower}})\, \mathrm{d}\widehat{a} \qquad\qquad \beta = \int_{-\infty}^{\widehat{a}} g(\widehat{a}; a_{\mathrm{upper}})\, \mathrm{d}\widehat{a}$$

for a_{lower} and a_{upper} (one of the limits is the observed value $\widehat{a}$). The confidence interval is called *central* for $\alpha = \beta = 1/2\,\gamma$.

One-sided confidence intervals represent limits. The value a_{lower} is a lower limit on the parameter a such that $a_{\mathrm{lower}} \leq a$ with probability $1 - \alpha$, and the value a_{upper} is an upper limit on the parameter a such that $a \leq a_{\mathrm{upper}}$ with probability $1 - \beta$.

Confidence level. The confidence level is a probability associated with the comparison of observed values with a given probability density function. The confidence level corresponds to the integral over the probability density function from the observed value to (minus or plus) infinity (→ Confidence interval, page 25, and → Goodness-of-fit, page 30).

Correlation coefficient. The correlation coefficient ρ_{jk} between two (correlated) random variables x_j and x_k is defined by

$$\rho_{jk} = \frac{V_{jk}}{\sqrt{V_{jj}V_{kk}}} = \frac{\text{cov}\,(x_j, x_k)}{\sigma_j \sigma_k} \qquad -1 \leq \rho_{jk} \leq +1$$

If $|\rho_{jk}| = 1$, the variables x_j and x_k are linearly dependent; in this case the covariance matrix is singular (→ Variance, page 34, → Covariance, page 26, and → Covariance matrix, page 26).

Covariance. The Covariance $\text{cov}(x_j, x_k)$ between two random variables x_j and x_k is equal to the element V_{jk} of the covariance matrix, and can be written in terms of expectation values:

$$\text{cov}(x_j, x_k) = E\left[(x_j - \langle x_j \rangle)(x_k - \langle x_k \rangle)\right]$$

For independent variables (→ Independence, page 30) the covariance vanishes, but the converse is not true.

Covariance matrix. For a n-dimensional random variable $\boldsymbol{x}$ the symmetric $n \times n$ covariance matrix is defined by

$$\boldsymbol{V}[\boldsymbol{x}] = \int_\Omega (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)\, p(\boldsymbol{x})\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n$$

where $p(\boldsymbol{x})$ is the probability density of the random variable $\boldsymbol{x}$, with elements

$$\begin{aligned} V_{jj} = \sigma_j^2 &= \int_\Omega (x_j - \langle x_j \rangle)^2\, p(\boldsymbol{x})\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \\ V_{jk} = \sigma_{jk} &= \int_\Omega (x_j - \langle x_j \rangle)(x_k - \langle x_k \rangle)\, p(\boldsymbol{x})\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n\ . \end{aligned}$$

The elements σ_j^2 are called variances (→ Variance, page 34), and the elements σ_{jk} are called covariances (→ Covariance, page 26); the covariance can be expressed by $\sigma_{jk} = \rho_{jk}\sigma_j\sigma_k$ in terms of the correlation coefficient ρ_{jk} (→ correlation coefficient, page 26) with the range $-1 < \rho_{jk} < +1$.

Error propagation. Given a random varibale $\boldsymbol{x}$ with covariance matrix $V_{\boldsymbol{x}}$. For a variable $\boldsymbol{y}$, related to $\boldsymbol{x}$ by a linear transformation $\boldsymbol{y} = B\boldsymbol{x}$, the covariance matrix $V_{\boldsymbol{y}}$ is given by

$$\boldsymbol{V_y} = \boldsymbol{A V_x A}^T\ .$$

In case of a nonlinera transformation ann approximation for $V_{\boldsymbol{y}}$ is calculated by the same formula whre the elements B_{ik} are the derivatives $\partial y_i / \partial x_k$, calculated at the position of mean value.

Estimator. An estimator for a parameter a is written as $\widehat{a}$, a function of the data $x_1, \ldots, x_i, \ldots, x_n$, and is itself a random variable.

An estimator $\widehat{a}(x_1, \ldots, x_i, \ldots, x_n)$ is called *consistent*, if it converges to the true value in the limit of large n (asymptotic limit). A consistent estimator is asymptotically unbiased.

The mean squared error of an estimate is

$$V[\widehat{a}] = E\left[(\widehat{a} - a_{\text{true}})^2\right] = E\left[(\widehat{a} - E[\widehat{a}])^2\right] + \left(E\left[(\widehat{a} - a_{\text{true}}]\right)\right)^2 = V[\widehat{a}] + B^2$$

where B ist the bias (→ Bias).

An estimator is called *efficient*, if its variance attains the minimum variance bound given by the → Cramer-Rao inequality.

Expectation value. The expectation value of the continuous random variable x (or mean value, population mean) with probability density function $p(x)$, and a discrete random variable with probabilities $P(x)$, respectively, is defined by

$$E[x] = \int_\Omega x\, p(x)\, \mathrm{d}x \qquad E[x] = \sum_i x_i\, P(x_i)\ .$$

The expectation value for a function $g(x)$ of a random variable is defined by

$$E[g(x)] = \int_\Omega g(x)\, p(x)\, \mathrm{d}x \qquad E[g(x)] = \sum_i g(x_i)\, P(x_i)\ .$$

Expectation values are also denoted by $\langle x \rangle$ and $\langle g(x) \rangle$, respectively.

The expectation value is a linear operator, for example

$$E[a \cdot g(x) + b \cdot h(x)] = a \cdot E[g(x)] + b \cdot E[h(x)]\ .$$

Exponential distribution. The exponential distribution of the continuous random variable x is defined by

$$p(x;\, \tau) = \frac{1}{\tau} e^{-x/\tau} \qquad 0 \le x < \infty$$

and depends on a single parameter τ, which is equal to the mean and the standard deviation of the random variable:

$$E[x] = \tau \qquad \sigma = \sqrt{V[x]} = \tau$$

An example for a random varibale with an exponential distribution is the decay time of an unstable particle, where the parameter τ is the mean decay time. The transformed parameter $\lambda = 1/\tau$ is the decay probability.

The exponential distribution can also describe attenuation phenomena, where the parameter τ has the meaning of an attenuation length.

F distribution. Let s_1^2 and s_2^2 be estimates of the two variances σ_1^2 and σ_2^2 with n_1 and n_2 degrees of freedom. In this case the ratio $n_1 s_1^2/\sigma_1^2$ is distributed as χ^2 with n_1 degrees of freedom and $n_2 s_1^2/\sigma_2^2$ is distributed as χ^2 with n_2 degrees of freedom. The random variable F, defined by

$$F = \frac{s_1^2}{\sigma_1^2} \frac{\sigma_2^2}{s_2^2}$$

follows the F distributon with (n_1, n_2) degrees of freedom. The density is given by

$$p(F;\, n_1, n_2) = \left(\frac{n_1}{n_2}\right)^{n_1/2} \cdot \frac{\Gamma\left((n_1+n_2)/2\right)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot F^{(n_1-2)/2} \left(1 + \frac{n_1}{n_2} F\right)^{-(n_1+n_2)/2}$$

Gamma distribution. The Gamma distribution is given by the expression:

$$\boxed{f(x|k) = \frac{x^{k-1} e^{-x}}{\Gamma(k)}}\ .$$

For an integer value of k with $k \ge 1$ (order k) the Gamma distribution gives the waiting time to the k-th event in a Poisson random process of unit mean ($\lambda = 1$); however the parameter k is not required to be an integer. The generalization for other values of the parameter λ is:

$$\boxed{f(x|k, \lambda) = \frac{x^{k-1} \lambda^k e^{-\lambda x}}{\Gamma(k)}}\ .$$

The parameter k affects the *shape* of the distribution, while the parameter λ is merely a scale parameter. For $\lambda = 1/2$ and $k = n/2$ the Gamma distribution becomes the χ^2-distribution.

Gaussian distribution. The Gaussian (or normal) distribution of the continuous random variable x is defined by

$$\boxed{f(x;\, \mu, \sigma^2)\, dx = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \qquad x \in (-\infty, +\infty)}$$

and has two parameters, the mean μ and the variance σ^2; it is also denoted $N(\mu, \sigma^2)$.

The distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called standard Gaussian, and denoted $N(0,1)$. The cumulative distribution of the Gaussian can be expressed by the Gaussian error function $\mathrm{erf}(x)$, defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,\mathrm{d}t$$

with $\mathrm{erf}(0) = 0$ and $\mathrm{erf}(\infty) = 1$. The cumulative Gaussian distribution is

$$\Phi(x) = \int_{-\infty}^{x} f(x; 0, 1)\,\mathrm{d}x = \frac{1}{2}\left(1 + \ \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right) .$$

Some values of the cumulative Gaussian distribution for values $x = 0 \ldots 3.49$ are shown below in a table.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Gaussian distribution, bivariate. The twodimensional of bivariate Gaussian density for two random variables x_1 and x_2 is given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

Marginal distributions. From this two-dimensional density the marginal distribution for x_1

$$p(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\,\mathrm{d}x_2 = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right\}$$

and the marginal distribution for x_2

$$p(x_2) = \int_{-\infty}^{+\infty} f(x_1, x_2)\,\mathrm{d}x_1 = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}$$

follow by integration over the other variable. The mean values of the two marginal distributions are μ_1 and μ_2, and the standard deviations are σ_1 und σ_2.

Conditional distributions. The conditional distributions depend on the correlation coefficient, defined by

$$\rho = \frac{E\left[(x_1-\mu_1)(x_2-\mu_2)\right]}{\sigma_1\sigma_2} \qquad\qquad -1 < \rho < +1$$

The conditional distribution of the random variable x_1 for a given value of the second variable x_2

$$p(x_1|x_2) = \frac{f(x_1,x_2)}{p(x_2)} = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x_1-\mu_1}{\sigma_1} - \rho\frac{x_2-\mu_2}{\sigma_2}\right]^2\right\}$$

and the conditional distribution of the random variable x_2 for a given value of the first variable x_1

$$p(x_2|x_1) = \frac{f(x_1,x_2)}{p(x_1)} = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x_2-\mu_2}{\sigma_2} - \rho\frac{x_1-\mu_1}{\sigma_1}\right]^2\right\}$$

follow. The mean values of the conditional distributions

$$\begin{aligned} E\left[x_1|x_2\right] &= \int_{-\infty}^{+\infty} x_1\, p(x_1|x_2)\, \mathrm{d}x_1 = \mu_1 + \rho\sigma_1\left(\frac{x_2-\mu_2}{\sigma_2}\right) \\ E\left[x_2|x_1\right] &= \int_{-\infty}^{+\infty} x_2\, p(x_2|x_1)\, \mathrm{d}x_1 = \mu_2 + \rho\sigma_2\left(\frac{x_1-\mu_1}{\sigma_1}\right) \end{aligned}$$

depend linearly on the correlation coefficient ρ. The variances of the conditional distributions are equal respectively to $\sigma_1^2(1-\rho^2)$ and $\sigma_2^2(1-\rho^2)$. The conditional means in the x_1, x_2 plane are straight lines with the slopes (dx_2/dx_1) of respectively $1/\rho(\sigma_2/\sigma_1)$ and $\rho(\sigma_2/\sigma_1)$. For values of $|\rho|$ close to one the two slopes are almos identical.

Lines of constant probability density. Lines of constant probability density of the two-dimensional distribution correspond to constant values of the exponent and are are ellipes. If the constant is choosen to be one, the condition is

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\frac{(x_2-\mu_2)^2}{\sigma_2^2} = 1-\rho^2$$

The probability to observe a point (x_1, x_2) inside this error ellipse is $1 - 1/\sqrt{e} = 0.3935$. To find a point for a given probability P (x_1, x_2) inside an ellipse the corresponding value of the χ^2-distribution with two degrees of freedom is $\chi_P^2 = 2\ln(1-P)$. The axes of the ellipse are inclined to the x_1-axis at the angles φ, defines by $\tan 2\varphi = \pm 2\rho\sigma_1\sigma_2/(\sigma_1^2 - \sigma_2^2)$. The relation

$$\rho = \pm(2-\varepsilon^2)^{-1}\sqrt{\varepsilon^4 - (1-\varepsilon^2)(\sigma_1/\sigma_2 - \sigma_2/\sigma_1)^2}$$

is valid for the excentricity ε of the ellipse.

Gaussian distribution, multivariate. The Gaussian distribution of a n-dimensional random variable $\boldsymbol{x}$ with elements $x_1, \ldots, x_i, \ldots, x_n$ is defined by

$$\boxed{f(\boldsymbol{x};\, \boldsymbol{\mu}, \boldsymbol{V}) = \frac{1}{(2\pi)^{n/2}\,|\boldsymbol{V}|^{1/2}} \exp\left\{-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^T\,\boldsymbol{V}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})\right\}}$$

and depends on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{V}$.
The n-dimensional vector $\boldsymbol{\mu}$ with elements $\mu_1, \ldots, \mu_i, \ldots, \mu_n$ is the vector of mean values, and the symmetric $n \times n$-matrix $\boldsymbol{V}$ is the covariance matrix:

$$\boldsymbol{\mu} = E\left[\boldsymbol{x}\right] \qquad\qquad \boldsymbol{V} = V\left[\boldsymbol{x}\right] = E\left[(\boldsymbol{x}-\boldsymbol{\mu})\,(\boldsymbol{x}-\boldsymbol{\mu})^T\right]$$

The n-dimensional Gaussian distribution is the model for the distribution of correlated variables, since its parameters are just the mean values, the variances and the covariances.

Global correlation coefficient. The (unsigned) largest correlation coefficient between the random variable x_j and a linear combination of all other variables is called global correlation coefficient and is given by

$$\rho_j = \sqrt{1 - \left[(\boldsymbol{V})_{jj} \cdot \left(\boldsymbol{V}^{-1}\right)_{jj}\right]^{-1}} \qquad 0 \leq \rho_j \leq 1 \; .$$

It measures the total amount of correlation between the variable x_j and all other variables. If $\rho_j = 0$, the variable x_j is uncorrelated with all other variables, and in the limit $\rho_j \to 1$, the variable x_j is completely correlated with at least one linear combination of the other variables.

Goodness-of-fit test. For a goodness-of-fit test, testing a given hypothesis, a test statistic is defined, which is a function of the data, measuring the distance between the data and a given hypothesis. The distance is converted to a probability P of obtaining data with a larger value of this test statistic than the observed value, assuming the hypothesis is true. The value of P is called the → confidence level of the fit. A small value of P indicates a low level of confidence in the hypothesis. The most common test for goodness-of-fit is the χ^2-test.

Independence. Consider a sample space with subset $\mathcal{A}$ and $\mathcal{B}$. The two subsets are called independent if

$$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}) \; .$$

Information. The amount of *information* about a parameter a given by a sample $x_1, \ldots, x_i, \ldots, x_n$ of measured values from a distribution $p(x|a)$ is given by the information I defined by the expection value

$$I(a) = E\left[\left(\frac{\partial}{\partial a} \ln \prod_{i=1}^{n} p(x_i|a)\right)^2\right] \; .$$

The inverse of the information is the minimal variance of any estimate $\widehat{a} = \widehat{a}(x_1, \ldots, x_i, \ldots, x_n)$.

Jacobian peak. A peak in a probability density can be caused by the variation of the Jacobi determinant. An example is a cross section which, as a function of tranverse momentum squared, falls off exponentially:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_T^2} = \frac{1}{2p_T} \frac{\mathrm{d}\sigma}{\mathrm{d}p_T} = b\, e^{-ap_T} \; .$$

As a function of the transverse momentum p_T the distribution is

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_T} = 2b\, p_T\, e^{-ap_T} \; ,$$

which has a peak at $p_T = 1/a$.

Marginal distribution. Given a joint probability function $p(x_1, x_2)$ of random variables x_1, x_2, the marginal probability density of the random variable x_1 is the density regardless of the value of x_2. The marginal probability density of x_1 is obtained by integrating the $p(x_1, x_2)$ over x_2:

$$p_{x_1}(x_1) = \int_{-\infty}^{+\infty} p(x_1, x_2)\, \mathrm{d}x_2$$

and corresponds to a projection of $p(x_1, x_2)$ onto the x_1-axis.

Median. The median $x_{1/2}$ of a probability density function is the → quantile of order 50 %. The probabilities for x to be observed greater of less than $x_{1/2}$ are equal.

For a random sample $x_1, \ldots, x_i, \ldots, x_n$ a practical method for the determination of the median is to sort the x-values and then to define

$$\text{median} \; = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \left(x_{n/2} + x_{(1+n/2)}\right)/2 & \text{if } n \text{ is even} \end{cases}$$

The median can be used as a measure of the location of the random variable, which is rather insensitive to long tails of the distribution.

Mode. The mode of a probability density function is the value of the random variable x at which the probability density function is a maximum. A probability density function may have local maxima.

Moment. The (central) moment of order m, or the moment of order m abut the mean $E[x] = \langle x \rangle$ of a random variable x is defined by the expectation value

$$\mu = E\left[(x - \langle x \rangle)^m\right]$$

In the case of a multidimensional random variable $\boldsymbol{x} = (x_1, x_2, \ldots)$ the moment of order $(m, n, \ldots)$ about the mean $\langle \boldsymbol{x} \rangle$ is defined by the expectation value

$$\mu = E\left[(x_1 - \langle x_1 \rangle)^m (x_2 - \langle x_2 \rangle)^n \ldots\right]$$

Multinomial distribution. The multinomial distribution is a generalization of the binomial distribution with n (more than two) classes into which an event can fall. If p_i is the probability of an individual event belonging to the i-th class, the probability of exacly k_i events belonging to class i for each i is given by the multinomial distribution

$$P(k_1, k_2, \ldots k_n) = \frac{N!}{p_1! p_2! \cdots p_n!} p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$$

with the normalization $\sum_i p_i = 1$ and with $\sum_i k_i = N$. Expectation value and variance of the k_i are $E[k_i] = Np_i$ and $V[k_i] = Np_i(1 - p_i)$; the correlation between k_i and k_j is $\rho_{ij} = -\sqrt{p_i p_j / (1 - p_i)(1 - p_j)}$.

A common example for the multinomial distribution is a histogram containing a fixed number N of independent events distributed into n bins. There is a correlation between bin contents because the sum is fixed to N. For $p_i \ll 1$ the variance $Np_i(1 - p_i)$ of k_i can be approximated by the Poisson variance Np_i. If the total number N of independent events is itself a Poisson random number, then all k_i follow individual Poisson distributions and the correlation vanishes.

Point Binomial distribution. Consider a process with two possible results in a trial, called "success" and "failure" (for example "head" or "tail", when a coin is tossed). The random variable x can take on two different values $x = 1$ and 0, associated to the two outcomes "success" and "failure". The probability can be written in the form

$$\boxed{P = p^x (1 - p)^{(1-x)} \qquad x = 0,\ 1}$$

called point binomial, which depends on one parameter p, with $0 \le p \le 1$. The probability for $x = 1$ is p, and the probability for $x = 0$ is $(1 - p)$.

Poisson distribution. The Poisson distribution gives the probability of exactly k events, if the number N of trials is very large and the probability p of the event to occur in a trial is very small, with a finite average value of $\mu = Np$. The Poisson distribution can be derived as the limiting form of the Binomial distribution (→ Binomial distribution, page 24) and has only one parameter μ, which is equal to the mean value *and* equal to the variance.

$$\boxed{P_\mu(k) = \frac{\mu^k e^{-\mu}}{k!} \qquad k = 0,\ 1,\ 2,\ \ldots}\ ;.$$

It is easy to see that the Poisson distribution is correctly normalized:

$$\sum_{k=0}^{\infty} P_\mu(k) = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{+\mu} = 1$$

Poisson distribution as limiting form of the Binomial distribution. From the Binomial distribution the probability for k successes in N trials, where p is the probability for a success in 1 trial, is given by the formula

$$P(k) = \binom{N}{k} p^k (1 - p)^{N-k} .$$

with an average (finite) number of $\mu = Np$. Expressing p by the formula μ/N the formula becomes

$$P(k) = \frac{N(N-1)\cdots(N-k+1)}{k!}\left(\frac{\mu}{N}\right)^k\left(1-\frac{\mu}{N}\right)^{N-k}$$
$$= \frac{N(N-1)\cdots(N-k+1)}{N^k}\frac{\mu^r}{k!}\left(1-\frac{\mu}{N}\right)^{N-k}$$

and by reordering the formula becomes

$$P(k) = \frac{\mu^k}{k!}\left[1\cdot\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\frac{(1-\mu/N)^N}{(1-\mu/N)^k}\right] .$$

Now the number N is increased and the probability p is decreased, keeping the average $\mu = Np$ at a fixed value. In the limit $N \to \infty$ for one factor in the parenthesis $\lim_{N\to\infty}(1-\mu/N)^N = e^{-\mu}$ and the other factors become 1, thus in this limit

$$\lim_{N\to\infty} P(k) = \frac{\mu^k}{k!}e^{-\mu}$$

the Poisson probability is obtained.

Probability. In the analysis of experimental data probability is interpreted as a limiting relative frequency. Elements of the sample space are the possible outcomes of a measurement and an elementary outcome or event $\mathcal{A}$ corresponds to the occurance of this single outcome. The probability $P(\mathcal{A})$ is the fraction of times event $\mathcal{A}$ occurs in repeated measurements. The assigned probabilities can be tested in experiments repeated under identical conditions.

The concept of probability as a relative frequency can not be applied to unique phanomena, where repeated tests are impossible. In these cases the probability expresses the *degree of belief* about the phenomena and is subjective ($\rightarrow$ Bayesian probability).

Quantile. The quantile of order α is defined as the value x_α of the random variable x such that

$$\alpha = F(x_\alpha) = \int_{-\infty}^{x_\alpha} p(x)\,\mathrm{d}x$$

with $0 \leq \alpha \leq 1$.

Random variable. A variable that takes on certain values with given probabilities is a random variable. A random variable is called continuous, if the value can take on value from a continuum (real numbers), and the distribution are described by a probability density function $p(x)$. A random variable k is called discrete, if it can take on a finite or infinite number of discrete values, which can be numbered $k_1, k_2, \ldots k_i, \ldots$; the probabilities are given by values P_i.

Sample. A sample $x_1, \ldots, x_i, \ldots, x_n$ of size n from a probability density function $p(x)$ or from a discrete distribution.

Sample mean and variance. For a sample $x_1, \ldots, x_i, \ldots, x_n$ of size n the arithmetic mean of the elements is called sample mean and denoted by $\overline{x}$; it is an estimate of the expectation value of x. The sample mean $\overline{x}$ and the sample variance s^2 of a sample of n values $x_1, \ldots, x_i, \ldots, x_n$ is defined and calculated by the formulas

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = x_{\text{est}} + \frac{1}{n}S_x \\ s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left\{S_{xx} - \frac{1}{n}S_x^2\right\} \end{aligned}$$

where S_x and S_{xx} are the sums

$$S_x = \sum_{i=1}^{n}(x_i - x_{\text{est}}) \qquad S_{xx} = \sum_{i=1}^{n}(x_i - x_{\text{est}})^2 .$$

The quantity x_{est} is a guess for the sample mean $\overline{x}$, for example $x_{\text{est}} = x_1$, introduced to avoid roundoff-errors.

The sample mean $\overline{x}$ is a consistent estimator for the population mean μ, if the variance of x exists (weak law of large numbers). Sample mean is an unbiased estimate of the population mean μ;

$$E\left[\overline{x}\right] = E\left[\frac{1}{n}\sum_i x_i\right] = \frac{1}{n}\sum_i E\left[x_i\right] = \frac{1}{n}\sum_i \mu = \mu$$

The sample variance, containing the factor $1/(n-1)$, is an unbiased estimate of the population variance σ^2.

For a sample of multidimensional random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n$ the generalization of the n sample means is straightforward. The quantities

$$\widehat{V}_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left((\boldsymbol{x}_i)_j - (\overline{\boldsymbol{x}})_j\right)\left((\boldsymbol{x}_i)_k - (\overline{\boldsymbol{x}})_k\right)\right]$$

are unbiased estimates for the covariance between elements j and k of the random variables $\boldsymbol{x}$.

Standard deviation. The standard deviation of a random variable with a probability density function $p(x)$, symbol σ, is defined as the square root of the variance:

$$\sigma = \sqrt{V\left[x\right]}$$

The actual probability content P corresponding to a deviation of k times the standard deviation from the mean of the distribution,

$$P = P\left(\langle x\rangle - k\sigma < x < \langle x\rangle + k\sigma\right)$$

depends on the specific probability density function $p(x)$. The Tschebyscheff inequality gives a (weak) limit on this probability, which is valid for any distribution (→ Tschebyscheff inequality, page 36). The following values are valid for the Gaussian density:

k	P in %
1	31.73
2	4.55
3	0.27
4	6.3×10^{-3}
5	5.7×10^{-5}
6	2.0×10^{-7}

k	P in %
0.67	50
1.28	20
1.64	10
1.96	5
2.58	1
3.29	0.1
3.89	0.01

Student's distribution. The t distribution is used for statistical tests of sample means. The distribution of the random variable t for n degrees of freedom is given by

$$p_n\left(t\right) = \frac{1}{\sqrt{n\pi}}\frac{\Gamma\left(\left(n+1\right)/2\right)}{\Gamma\left(n/2\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad -\infty < t < +\infty$$

Student's distribution is used for a test wether an observed mean $\overline{x}$ from a sample of size n is consistent with a theoretical value μ, with a s determined from $s^2 = \sum_i (x_i - \overline{x})^2/(n-1)$. The correct statistics for this test is

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

which follows a Student's distribution with $(n-1)$ degrees of freedom.

Transformation of random variable.

Single variable. Given the probability density $p(x)$ of a random variable x. If the variable x is transformed to a new random variable $y = y(x)$, the probability density $q(y)$ of y is given by

$$q(y) = p\left(x\left(y\right)\right)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$$

where $x = x(y)$ is the function inverse to $y = y(x)$. The shape of the density is changed by a transformation, unless the transformation is linear. The transformation has to be done for each branch for a transformation, that does not have a unique inverse.

Several variables. A linear transformation ist given by $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$, where the n-vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu_x}$ and covariance matrix $\boldsymbol{V_x}$ is transformed to the m-vector $\boldsymbol{y}$ by the $m \times n$ matrix $\boldsymbol{B}$. The dimensions n and m may be different; the matrix elements correspond to $B_{ik} = \partial y_i/\partial x_k$. The vector of mean values $\boldsymbol{\mu_y}$ and the covariance matrix $\boldsymbol{V_y}$ are given by

$$\langle \boldsymbol{y} \rangle = \boldsymbol{\mu_y} = \boldsymbol{B}\boldsymbol{\mu_x} \qquad \boldsymbol{V_y} = \boldsymbol{B}\boldsymbol{V_x}\boldsymbol{B}^T$$

The transformed probability density in the special case $n = m$ is given by

$$q(\boldsymbol{y}) = p\left(\boldsymbol{x}(\boldsymbol{y})\right) \det |\boldsymbol{B}^{-1}|$$

with the inverse transformation $\boldsymbol{x} = \boldsymbol{x}\left(\boldsymbol{y}\right) = \boldsymbol{B}^{-1}\boldsymbol{y}$.

In the general (nonlinear) case the transformation is defined by equations of the form

$$y_i = y_i(x_1, x_2 \ldots x_n) = y_i(\boldsymbol{x})$$

In the case $n = m$ the probility density of $\boldsymbol{y}$ is given by

$$q(\boldsymbol{y}) = p(\boldsymbol{x}) \cdot J\left(\frac{x}{y}\right) \qquad J\left(\frac{x}{y}\right) = \left|\frac{\partial x_1 \partial x_2 \ldots \partial x_n}{\partial y_1 \partial y_2 \ldots \partial y_n}\right|$$

Only an approximate transformation of the covariance matrix is possible in the general case of a nonlinear transforamtion of the n-vector $\boldsymbol{x}$ to the m-vector $\boldsymbol{y}$ by the formula (→ Error propagation)

$$\boldsymbol{V_y} = \boldsymbol{B}\boldsymbol{V_x}\boldsymbol{B}^T$$

where the elements $B_{ik} = \partial y_i/\partial x_k$ are calculated at the position of the mean $\boldsymbol{\mu_y}$.

Uniform distribution. The uniform distribution is constant over a finite range and zero outside:

$$\boxed{p(x;\, a, b) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{elsewhere .} \end{cases}}$$

The distribution with limits a and b is also denoted $U(a, b)$. Mean and variance are

$$\mu = E\left[x\right] = \frac{a+b}{2} \qquad \sigma^2 = V\left[x\right] = \frac{(b-a)^2}{12}$$

The uniform distribution $U(0, 1)$ with limits 0 and 1, denoted $U(0, 1)$, has mean 1/2 and standard deviation $\sqrt{1/12} = 0.2887$. Variables from $U(0, 1)$ are generated by standard → random number generators.

Variance. The variance of a continuous random variable x is defined by

$$V\left[x\right] = E\left[(x - \langle x \rangle)^2\right] = \int_\Omega (x - \langle x \rangle)^2\, p(x)\, \mathrm{d}x$$

where $p(x)$ is the probability density of the random variable x. It is the second moment about the mean value $\langle x \rangle = E\left[x\right]$ and equal to the square of the standard deviation σ (→ standard deviation, page 33).

$$V\left[a \cdot x\right] = a^2 \cdot V\left[x\right] \qquad V\left[x\right] = E\left[x^2\right] - \langle r \rangle^2$$

For a discrete random variable k the definition of the variance is

$$V\left[k\right] = \sum_i \left(k_i - \langle k \rangle\right)^2 P(k_i)$$

with the probability $P(k)$

For the variance of a finite sample → Sample mean and variance, page 32. A matrix called covariance matrix is defined for a multivariate random variable (→ Covariance matrix, page 26).

(b) Theorems

Binomial theorem. For p with $0 < p < 1$ and $q = 1 - p$ the following expression is called the Binomial theorem:

$$(p+q)^n = 1 = \sum_{k=0}^{n} \binom{k}{n} p^k \cdot (1-p)^{n-k}$$

Central limit theorem. *The density function of the sum $w = \sum_{i=1}^{n} x_i$ of a sample of n independent random values x_i of any distribution with mean $\bar{x}$ and variance σ^2 approaches in the limit $n \to \infty$ a Gaussian density function with mean $n \cdot \bar{x}$ and Variance $n \cdot \sigma$.*

Cramer-Rao inequality. This inequality gives a lower limit for the variance of any estimator $\widehat{a}$, which has a bias $B(a)$:

$$V(\widehat{a}) \geq \frac{(1 + (\mathrm{d}B/\mathrm{d}a))^2}{I(a)}$$

The variance of an unbiased estimate $\widehat{a}$ is bounded below by the inverse of the information $I(a)$ contained in the data sample (→ Information, page 30).

Characteristic function. For a real random variable x with density function $p(x)$ the *characteristic function* $\phi(t)$ is defined as expectation value of e^{itx}:

$$\phi(t) = E\left[e^{itx}\right] = \int_{-\infty}^{+\infty} e^{itx} \cdot p(x)\, \mathrm{d}x$$

The inverse operation is:

$$p(x) = \int_{-\infty}^{+\infty} e^{itx} \cdot \phi(t)\, \mathrm{d}t\ .$$

The values of the characteristic function and its first and second derivative at the special value $t = 0$ are easily calculated:

$$\phi(0) = 1 \qquad (97)$$

$$\frac{d\phi(0)}{dt} = \left. \int_{-\infty}^{+\infty} ixe^{itx} P(x) \mathrm{d}x \right|_{t=0} = i\,\langle x\rangle \qquad (98)$$

$$\frac{d^2\phi(0)}{dt^2} = \left. \int_{-\infty}^{+\infty} -x^2 e^{itx} P(x) \mathrm{d}x \right|_{t=0} = -E\left\{x^2\right\} = -\left(\sigma^2 + \langle x\rangle^2\right) \qquad (99)$$

For the characteristic function with respect to the mean value $\langle x\rangle = E[x]$, i.e. for the expectation value $E\left[\exp(it(x - \langle x\rangle))\right]$ the terms containing $\langle x\rangle$ are removed from the last two expressions.

The characteristic function for the Gaussian density. The characteristic function of the standard Gaussian density is defined by

$$\phi_{\mathrm{G}}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx} e^{-x^2/2\sigma^2}\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-((x-it\sigma^2)^2 - (it\sigma^2)^2)/2\sigma^2}\, \mathrm{d}x$$

$$\frac{1}{\sqrt{2\pi}} e^{-t^2\sigma^2/2} \int_{-\infty}^{+\infty} e^{-(x-it\sigma^2)^2/2\sigma^2}\, \mathrm{d}x$$

The last integral turns out to be a constant:

$$\int_{-\infty}^{+\infty} e^{-(x-it\sigma^2)^2/2\sigma^2}\, \mathrm{d}x = \sigma^2 \int_{r=-\infty-it}^{r=+\infty-it} e^{-r^2/2}\, \mathrm{d}r$$

by the transformation $r = (x - it\sigma^2)/\sigma^2$ with $\mathrm{d}x = \sigma^2\, \mathrm{d}r$; the integral vanishes on a closed path, because no singularities are included (residue theorem), or symbolically

$$\int_{-L}^{+L} e^{-r^2/2}\, \mathrm{d}r + \int_{-L}^{+L-it} e^{-r^2/2}\, \mathrm{d}r + \int_{L-it}^{-L-it} e^{-r^2/2}\, \mathrm{d}r + \int_{-L-it}^{-L} e^{-r^2/2}\, \mathrm{d}r = 0$$

The 2. and 4. integral vanish for $L \to \infty$, because the integrand vanishes. It follows:

$$\int_{-\infty-it}^{+\infty-it} e^{-r^2/2}\,\mathrm{d}r = \int_{-\infty}^{+\infty} e^{-r^2/2}\,\mathrm{d}r = \text{constant}$$

The characteristic function

$$\phi_{\mathrm{G}}(t) = N_0 \cdot e^{-t^2\sigma^2/2}$$

follows for the Gaussian density.

Gauß-Markov theorem. The Gauß-Markov theorem makes a statement about the variance of $\to$ estimators, which are *unbiased* and are *linear functions* of the data. The linear least squares estimator has the smallest possible variance among all estimators with these properties, if the data are unbiased, regardless of the distribution of the data; i.e. the distribution does not need to be Gaussian.

Law of large numbers. In N statistical experiments the number n_j of an event j is counted. The ratio $h_j = n_j/N$ is a random variable, which follows the Binomial distribution. The expectation value of the random variable h_j, $E(h_j)$, is the probability p_j for the event j. In real life the probabilities p_j of the different events are unknown; they have to be determined (estmated) from experiments. The question ist: How accurate is the estimate h_j for p_j? Because the random variable $h(j)$ follows a Binomial distribution, the variance of h_j can be calculated:

$$\sigma^2(h_j) = \sigma^2(n_j/N) = \frac{1}{N^2} \cdot \sigma^2(n_j) = \frac{1}{N^2} \cdot Np_j(1-p_j)$$

Since the product $p_j(1-p_j)$ is limited (maximum value is $1/4$), the inequality

$$\sigma^2(h_j) \leq 1/N$$

follows, known as the **Law of large numbers**.

Tschebyschew inequality. *The probability for $|x - \bar{x}| \geq k\sigma$ is less or equal to $1/k^2$ for a random variable x with any density function $p(x)$.*

The inequality gives an upper limit on the probability of exceeding a given number of standard deviations, which does not depend on the shape of the density, provided the variance is known.

Proof: The proof starts from the general definition of the variance σ^2.

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x =$$

$$\int_{-\infty}^{\bar{x}-k\sigma} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x + \int_{\bar{x}-k\sigma}^{\bar{x}+k\sigma} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x + \int_{\bar{x}+k\sigma}^{+\infty} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x$$

The integrands in all three integrals are positive. By removing the middle integral an inequality is defined:

$$\sigma^2 \geq \int_{-\infty}^{\bar{x}-k\sigma} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x + \int_{\bar{x}+k\sigma}^{+\infty} (x-\bar{x})^2 \cdot p(x)\,\mathrm{d}x$$

In the first and last integral the inequalities $x > \bar{x} + k\sigma$ and $x < \bar{x} - k\sigma$, respectively, are valid, thus

$$\sigma^2 \geq k^2\sigma^2 \left[\int_{-\infty}^{\bar{x}-k\sigma} p(x)\,\mathrm{d}x + \int_{\bar{x}+k\sigma}^{+\infty} p(x)\,\mathrm{d}x \right]$$

The expression in parentheses is the probability for $|x - \bar{x}| \geq k\sigma$, which proves the Tschebyscheff inequality.

The inequality is valid under rather general conditions and is rather weak. It is useful for theoretical considerations and for those practical cases where the knowledge about the distribution is very limited.

(c) Miscellaneous

Binomial coefficient. The expression

$$\binom{r}{n} = \frac{n!}{k!(n-k)!} = \binom{k}{n-k} \qquad 0 \le k \le n$$

is called Binomial coefficient. This number is identical to the number of combinations C_k^n without repetition.

Gamma function. Euler's Gamma function $\Gamma(x)$ is related to the factorial by $\Gamma(x) = (x-1)!$ for integer, positive x. The function is defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\,\mathrm{d}t \qquad x > 0$$

Properties are: $\Gamma(x) = 1/x \cdot \Gamma(x+1) = (x-1)\Gamma(x-1)$, and $\Gamma(x)\Gamma(1-x) = \pi/\sin(\pi x)$. Some values are:

x	$=$	$-1/2$	$1/2$	1	$3/2$	2
$\Gamma(x)$	$=$	$-2\sqrt{\pi}$	$\sqrt{\pi}$	1	$\sqrt{\pi}/2$	1

An asymptotic formula for large arguments x is Stirling's formula

$$\Gamma(x) \approx e^{-x} x^x \sqrt{2\pi/x}\left(1 + \frac{1}{12x} + \frac{1}{288x^2} + \ldots\right)$$

which can also be used to approximate the factorial $x!$ or the logarithm of the factorial (for integer x):

$$x! \approx \left(\frac{x}{e}\right)^x \sqrt{2\pi/x}\left(1 + \frac{1}{12x} + \frac{1}{288x^2} + \ldots\right) \qquad \ln x! \approx \left(x + \frac{1}{2}\right)\ln n - n + \ln\sqrt{2\pi}$$

Hessian. The symmetric $n \times n$ of second partial derivatives of a scalar function $F(\boldsymbol{x})$ of an n-vector $\boldsymbol{x}$ is called Hessian matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{pmatrix} \partial^2 F/\partial x_1^2 & \partial^2 F/\partial x_1 \partial x_2 & \ldots & \partial^2 F/\partial x_1 \partial x_n \\ \partial^2 F/\partial x_2 \partial x_1 & \partial^2 F/\partial x_2^2 & \ldots & \partial^2 F/\partial x_2 \partial x_n \\ \ldots & & & \ldots \\ \partial^2 F/\partial x_n \partial x_1 & \partial^2 F/\partial x_n \partial x_2 & \ldots & \partial^2 F/\partial x_n^2 \end{pmatrix}$$

Integrals. Some useful integrals occuring frequently in statistical calculations are given below.

$$\int_{-\infty}^{+\infty} e^{-ax^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}} \qquad \int_0^\infty x\,e^{-ax^2}\,\mathrm{d}x = \frac{1}{2a}$$

$$\int_{-\infty}^{+\infty} x^2\,e^{-ax^2}\,\mathrm{d}x = \frac{1}{2a}\sqrt{\frac{\pi}{a}} \qquad \int_0^\infty x^3\,e^{-ax^2}\,\mathrm{d}x = \frac{1}{2a^2}$$

$$\int_{-\infty}^{+\infty} x^{2n}\,e^{-ax^2}\,\mathrm{d}x = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{(2a)^n}\sqrt{\frac{\pi}{a}} \qquad \int_0^\infty x^{2n+1}\,e^{-ax^2}\,\mathrm{d}x = \frac{n!}{2a^{n+1}}$$

Jacobi matrix. Given m functions $f_1(x_1, \ldots x_n), \ldots f_m(x_1, \ldots x_n)$ of n variables $x_1, \ldots x_n$, the Jacobi matrix is the matrix of first derivatives of the functions f_i woth respect to the variables x_j:

$$\boldsymbol{J} = \frac{\partial\,(f_1, \ldots f_m)}{\partial\,(x_1, \ldots x_n)} = \begin{pmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \ldots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \ldots & \partial f_2/\partial x_n \\ \ldots & & & \ldots \\ \partial f_m/\partial x_1 & \partial f_m/\partial x_2 & \ldots & \partial f_m/x_n \end{pmatrix}$$

or short $\partial \boldsymbol{f}/\partial \boldsymbol{x}$.

The chain rule is valid: if $\boldsymbol{x} = \boldsymbol{x}(u_1, \ldots u_p)$, then the derivative $\partial \boldsymbol{f}/\partial \boldsymbol{u}$ is given by the matrix product $(\partial \boldsymbol{f}/\partial \boldsymbol{x})\,(\partial \boldsymbol{x}/\partial \boldsymbol{u})$, i.e.

$$\frac{\partial f_i}{\partial u_k} = \sum_{j=1}^{n} \frac{\partial u_i}{\partial x_j} \frac{\partial x_j}{\partial u_k} \ .$$

For a coordinate transformation $\boldsymbol{u} \to \boldsymbol{u} = \boldsymbol{u}(x_1, \ldots x_n)$ with $p = n$ the product of $\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x}$ and $\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x}$ is the unit matrix, i.e. $\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x} = (\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x})^{-1}$.

Jacobian. The Jacobian or Jacobi determinant is the factor in the transformation of a function of n variables.

$$p\,(\boldsymbol{x})\ \mathrm{d}x_1\mathrm{d}x_2 \ldots \mathrm{d}x_n = p\,(\boldsymbol{x}\,(\boldsymbol{u}))\,|\det(\mathrm{d}\boldsymbol{x}/\mathrm{d}\boldsymbol{u})\,|\,\mathrm{d}u_1\mathrm{d}u_2 \ldots \mathrm{d}u_p$$

where $(\mathrm{d}\boldsymbol{x}/\mathrm{d}\boldsymbol{u}) = (\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x})^{-1}$ is the → Jacobi matrix. The factor

$$|\det(\mathrm{d}\boldsymbol{x}/\mathrm{d}\boldsymbol{u})\,| = |\det(\mathrm{d}\boldsymbol{u}/\mathrm{d}\boldsymbol{x})\,|^{-1}$$

is the absolut value of the Jacobi determinant.

(→ Transformation of variables).

Minimization. The determination of the minimum of a function $F(\boldsymbol{a}$ is called minimization. Minimization or maximization problems arise in optimization problems (e.g. maximum likelihood method), where an objective function $F(\boldsymbol{a})$ depending on a parameter vector $\boldsymbol{a}$ is defined; the minimum or the maximum corresponds to the optimal solution of the optimization problem.

A local minimum of a function $F(\boldsymbol{a})$, depending on n variables $a_1, \ldots, a_i, \ldots, a_n$, can be determined by solving simultaneously the n equations

$$\frac{\partial F}{\partial a_1} = 0 \quad \ldots \quad \frac{\partial F}{\partial a_i} = 0 \quad \ldots \quad \frac{\partial F}{\partial a_n} = 0$$

which correspond to a vanishing gradient $\boldsymbol{\nabla} F$ and in general, define a minimum, a maximum or a saddle point. Methods for the determination are search methods or iterative methods; none of the many methods is universically applicable. There are simple methods like the simplex method of Nelder and Mead which do not use derivatives and which are useful especially in regions far from the minimum. The iterative Newton-Raphson method requires derivatives and is based on the Taylor expansion of the obejctive function at some value $\boldsymbol{a}_t$:

$$F(\boldsymbol{a}) = F(\boldsymbol{a}_t) + (\boldsymbol{a} - \boldsymbol{a}_t)^T \boldsymbol{g} + \frac{1}{2} (\boldsymbol{a} - \boldsymbol{a}_t)^T \boldsymbol{H} (\boldsymbol{a} - \boldsymbol{a}_t) + \ldots$$

where $\boldsymbol{g}$ is the gradient with elements $g_j = \partial F/\partial a_j$ and $\boldsymbol{H}$ is the Hessian with elements $H_{jk} = \partial^2 F/\partial a_j \partial a_k$ at the value $\boldsymbol{x} = \boldsymbol{x}_t$. Ignoring the higher-order terms a step $\boldsymbol{\Delta a}$ can be calculated from the condition $\boldsymbol{\nabla} F = 0$;

$$\boldsymbol{\Delta a} = \boldsymbol{a} - \boldsymbol{a}_t = -\boldsymbol{H}^{-1} \boldsymbol{g} \ .$$

which in general gives an improvement towards the solution. In a robust method a so-called line search is made for the minimum of the function $f(\alpha) = F(\boldsymbol{a}_t + \alpha \cdot \boldsymbol{\Delta a})$. The convergence towards the solution is usually very fast near the minimum; monitoring of the convergence can be done with the quantity

$$d = -\frac{1}{2} \boldsymbol{g}^T \boldsymbol{\Delta a} = \frac{1}{2} \boldsymbol{g}^T \boldsymbol{H} \boldsymbol{g}$$

which corresponds to the expected function decrease. The method requires the computation of the derivatives, which can be done analytically or by numerical methods.

A special case of minimization is the maximum likelihood method, where the function $F(\boldsymbol{a}) = -\ln \mathcal{L}(\boldsymbol{a})$ is approximately quadratic and the inverse Hessian $\boldsymbol{H}$ can be considered as a good approximation of the covariance matrix of parameters.

Standard minimization problems are impractical in case of very large number of parameters, or if the number of possible discrete solutions is given by permutations. Different methods are available in those cases like *simulated annealing* and *genetic algorithms*.

Monte Carlo method. A statistical process can be simulated by mathematical methods, where all random variables of the statistical process are artificially generated. Standard methods are used to generated random variables u from a uniform distribution $U(0,1)$ ($\rightarrow$ random number generator), and these random variables are transformed to the random variables x of the process. This method is called Monte Carlo method.

The Monte Carlo method allows to simulate experiments and to study the statistical properties of the relevant variables, e.g. the distribution of an $\rightarrow$ estimator.

Optimization. Optimization according to a certain criterium is usually expressed in terms of an objective function, which depends on parameters. The optimality condition corresponds to the minimum or maximum of the objective function.

Permutation. The number of permutations on n distinguishable objects is equal to $P_n = n!$ (with $0! = 1! = 1$). For example for the three objects a, b and c there are $3! = 6$ permutations abc, bca, cab, cba, bac and acb.

The number of combinations of n distinguishable objects to r objects is equal to the $\rightarrow$ Binomial coefficient $C_n^r = \binom{r}{n}$.

Schwarz inequality.

$$\int f^2 \, \mathrm{d}x \int g^2 \, \mathrm{d}x \geq \left(\int fg \, \mathrm{d}x \right)^2$$

and for vectors

$$|\boldsymbol{x}| \; |\boldsymbol{y}| \geq |\boldsymbol{x} \cdot \boldsymbol{y}|$$

Index

Bayes theorem, **23**
Bayesian probability, **23**
Bayesian statistics, **23**
Beta distribution, **23**
Bias, **23**
Binomial coefficient, **37**
Binomial distribution, **24**
Binomial theorem, **35**
Breit Wigner distribution, **24**

Cauchy distribution, **24**
central limit theorem, **35**
characteristic function, **35**
χ^2 distribution, **24**
classical statistics, **25**
conditional distribution, **25**
conditional probability, **25**
confidence interval, **25**
confidence level, **25**
confidence limits, **25**
continuous variable
    random, **32**
correlation coefficient, **26**
    global, **30**
covariance, **26**
Cramer-Rao inequality, **35**

discrete variable
    random, **32**
distribution
    multivariate Gaussian, **29**
    Beta, **23**
    Binomial, **24**
    bivariate Gaussian, **28**
    Breit Wigner, **24**
    Cauchy, **24**
    χ^2, **24**
    conditional, **25**
    exponential, **27**
    F, **27**
    gamma, **27**
    Gaussian, **27**
    Lorentz, **24**
    marginal, **30**
    multinomial, **31**
    normal, **27**
    point Binomial, **31**
    Poisson, **31**
    Student's, **33**
    uniform, **34**

error propagation, **26**, **33**
estimator, **26**
expectation value, **26**
exponential distribution, **27**

F distribution, **27**
factorial, **37**
function
    characteristic, **35**
    gamma, **37**

gamma distribution, **27**
gamma function, **37**
Gauß-Markov theorem, **36**
Gaussian distribution, **27**
    multivariate, **29**
    bivariate, **28**
global correlation coefficient, **30**
goodness-of-fit test, **30**

Hessian, **37**

independence, **30**
inequality
    Cramer-Rao, **35**
    information, **35**
    Schwarz, **39**
    Tschebyschew, **36**
information, **30**
Integrals, **38**

Jacobi matrix, **38**
Jacobian, **38**
Jacobian peak, **30**

law of large numbers, **36**
Limit
    lower, **25**
    upper, **25**
Limits, **25**
Lorentz distribution, **24**

marginal distribution, **30**
mean value, **26**
median, **30**
minimization, **38**
mode, **31**
moment, **31**
Monte Carlo method, **39**
multinomial distribution, **31**

normal distribution, **27**

optimization, **39**

permutation, **39**
point Binomial distribution, **31**
Poisson distribution, **31**
probability
    Bayesian, **23**
    conditional, **25**
    subjective, **23**

quantile, **32**

random variable, **32**

sample, **32**
 mean, **32**
 variance, **32**
Schwarz inequality, **39**
standard deviation, **33**
statistics
 Bayesian, **23**
Student's distribution, **33**
Subjective probability, **23**

theorem
 Bayes, **23**
 Binomial, **35**
 central limit, **35**
 Gauß-Markow, **36**
transformation of random variable, **33**
Tschebyschew inequality, **36**

uniform distribution, **34**

variable
 random, **32**
 transformation, **33**
variance, **34**