Fortgeschrittenen-Praktikum

Statistical Methods
in Experimental Physics

Physikalisches Institut

Ruprecht-Karls-Universität
Heidelberg

Markus Köhli, Klaus Reygers

Version 1.03
22 May 2023

Contents

1 Introduction

1.1 Preface

Every measured value is accompanied by an uncertainty' - an iron rule that should have been internalized after the first semester at the latest. This premise seems intuitive and well-founded.

In practice, however, this principle often first falls victim to the most diverse justifications, which are not infrequently based on incomplete knowledge of the experiment, but are usually justified by a qualitative success.

Hubble[1] based the thesis about the expansion of the universe on the observation of surrounding galaxies in the near range, which was too small in astronomical scales, respectively superimposed by other relative motions. The assertion turned out afterwards to be correct in his favor. Wrong, on the other hand, were countless discoveries such as that of polywater[2], of the polymer structure of water. It would be wrong to attribute historical negative examples to a rather qualitative way of working, which would seem unthinkable today. The statistically correct and comprehensible way of working is the foundation of the building of science today as it was then. But since the departure from this basic idea appears again and again in various form, a discreet reference to its importance should be made now and then [3] - quite in the spirit of the experimenter.

[1] Hubble, E.P.: *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*, (1929) Proc. Natl. Acad. Sci. USA 15, p. 168–173

[2] Fedyakin, N.N.: *Change in the Structure of Water during Condensation in Capillaries*, (1962) Colloid Zhournal 24, p. 497

[3] Vaux, D.L.: *Research methods: Know when your numbers are significant*, (2012) Nature 492, pp. 180–181

1.2 Acknowledgements

This script is based on the lecture '*Statistische Methoden im Fortgeschrittenen-Praktikum*', by Volker Büscher given at Albert-Ludwigs-Universität Freiburg.

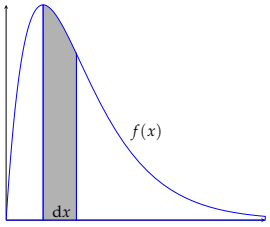
1.3 *Literature*

The following books are recommended for further reading:

COWAN, G.: *Statistical Data Analysis*, Oxford Science Publications
BRANDT, S.: *Datenanalyse*, Spektrum Akademischer Verlag
BARLOW, R.J.: *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley-VCH

2 Basics

2.1 Probability density



The probability P to find a value in the interval $[x, x + \mathrm{d}x]$ is given by the **probability density function** (*pdf*) $f(x)$:

$$P(x' \in [x, x + \mathrm{d}x]) = f(x) \cdot \mathrm{d}x.$$

The *pdf* by definition is normalized to unity

$$\int_{\Omega} f(x)\,\mathrm{d}x = 1$$

on the sample space Ω.

The probability for x' to be below $b \in \Omega[-\infty, \infty]$ is

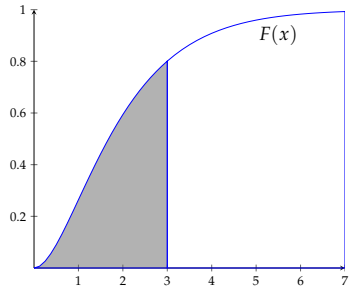
$$P(x' \leq b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x = F(b),$$

and correspondingly, the probability to find x' in the interval $[a, b]$ is:

$$P(a \leq x' \leq b) = \int_{a}^{b} f(x)\,\mathrm{d}x = F(b) - F(a).$$

With the primitive function constructed in this way, we obtain the **cumulative distribution** $F(x)$ of the probability density function $f(x)$:

$$F(x) := \int_{-\infty}^{x} f(x')\,\mathrm{d}x'.$$



$\lim_{x\to\infty} F(x) = 1$

$\lim_{x\to-\infty} F(x) = 0$

2.2 Expectation values

The expectation value $E(x)$ of a random variable x following a probability density function $f(x)$ is defined as:

$$E(x) := \mu = \int_{-\infty}^{\infty} x \cdot f(x)\,\mathrm{d}x.$$

The expectation value $E(h(x))$ of an arbitrary function $h(x)$ is:

$$E(h(x)) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,\mathrm{d}x.$$

2.3 Central Moments and Variance

Expectation values of the function

$$h_l(x) = (x - c)^l$$

are called **l-th moments** of the variable x at point c.

The **special cases** of the l-th moments α_l around the expectation value μ are given by:

$$\alpha_0 = \int_{-\infty}^{\infty} (x-\mu)^0 f(x)\,\mathrm{d}x = 1, \qquad \text{normalization}$$

$$\alpha_1 = \int_{-\infty}^{\infty} (x-\mu)^1 f(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x - \mu = 0, \qquad \text{(expectation value)}$$

$$\alpha_2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x.$$

variance
Contains information about the scatter of a random variable x around its mean value μ

The variance is defined as

$$\alpha_2 = E\left((x-\mu)^2\right) = \mathrm{Var}(x) = \sigma^2(x).$$

The square root of the variance $\sqrt{\mathrm{Var}(x)}$ is called the **standard deviation** $\sigma(x)$. Mean and standard deviation are generally the most important quantities for the statistical description of a series of measurements. The uncertainty in an experiment, the measurement error, is usually identified with the standard deviation.

The variance can be expressed in terms of expected values as follows:

$$\sigma^2(x) = E\left((x-\mu)^2\right) = E(x^2) - (E(x))^2\,.$$

Higher Moments

The moments α_l for $l > 2$ are called higher moments. Higher powers put more and more weight on the tails of the distribution.

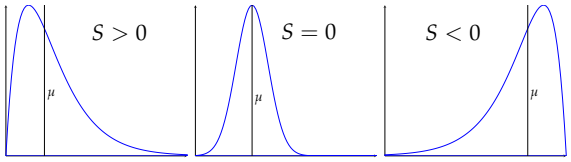
$$\alpha_3 = E\left((x-\mu)^3\right).$$

skewness
a measure of the asymmetry of the probability distribution of x

The quantity

$$S := \alpha_3/\sigma^3$$

is called **skewness**. It is the first non-zero odd central moment. It weights values to the right and left of the expected value with different signs. If S is exactly zero, the distribution is symmetric. For $S < 0$ a distribution is called left-skewed, which means it decreases more slowly to the left than to the right. For $S > 0$ it is called right-skewed.



$$\alpha_4 = E\left((x-\mu)^4\right)$$

kurtosis
A measure of the "tailedness" of the probability distribution due to the 4-th power in the exponent
von gr. $\kappa\upsilon\rho\tau\omega\sigma\iota\varsigma$ = curved, arching

The **kurtosis** K is defined as

$$K := \alpha_4/\sigma^4 - 3.$$

The fourth central moment of a Gaussian distribution is exactly 3. By subtracting 3, the kurtosis is normalized to the extent to which a distribution is narrower ($K < 0$), that is, more centered around the mean, or wider ($K > 0$) than a Gaussian distribution.

2.4 *Functions of two Random Variables*

For a two-dimensional probability density function $f(x,y)$ of the random variables x, y we can write

$$P(x' \in [x, x+\mathrm{d}x], y' \in [y, y+\mathrm{d}y]) = f(x,y) \cdot \mathrm{d}x\,\mathrm{d}y$$

with the normalization

$$\int\int_{\Omega} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1.$$

The one-dimensional projections are called **marginal distributions**

$$f_y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x,$$
$$f_x(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y,$$

with the expectations values μ_x and μ_y.

The expectation value of a two-dimensional function $h(x,y)$ is defined analogously to the one-dimensional case:

$$E(h(x,y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

The variances with respect to one variable are given by:

$$\sigma^2(x) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

$$\sigma^2(y) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y-\mu_y)^2 f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

If one considers both variables at the same time, one speaks of the **covariance**:

$$\mathrm{cov}(x,y) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)\,(y-\mu_y)\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y. \qquad \textbf{covariance}$$

Covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values (that is, the variables tend to show similar behavior), the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, the covariance is negative. The strength of the correaltion is quantified with the dimensionless **correlation coefficient** ρ which is normalized to the relative widths $\sigma_{x,y}$: **correlation coefficient**

$$\rho(x,y) := \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)}.$$

ρ takes values from -1 (negative correlation) to 1 (positive correlation). In these cases, a variation in x leads to an equally large variation in y, and the opposite for negative ρ.

The variables x and y are said to be **independent** if the joint distribution can be written as

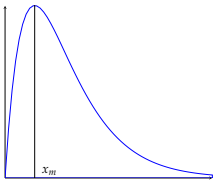
independent variables

$$f(x,y) = f_x(x) \cdot f_y(y).$$

The correlation coefficient for independent variables is $\rho = 0$, i.e. independent variables are uncorrelated. However, the converse does not hold, i.e., a correlation coefficient $\rho = 0$ does not mean that x and y are independent.

Independent variables are uncorrelated. The converse does not hold.

2.5 *Further parameters: Median und Quantiles*

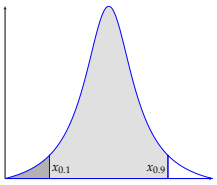


- The **mode** x_m is the value at which the distribution takes its maximum, i.e.,

$$x_m = \arg\max_{x \in \Omega} f(x).$$

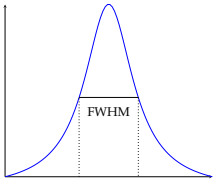
- The value $x_{0.5}$ at which the cumulative distribution function takes the value $1/2$ is called the **median**:

$$F(x_{0.5}) = \int_{-\infty}^{x_{0.5}} f(x)\,\mathrm{d}x = 0.5.$$



- More generally, the **quantile** x_q is the value at which the cumulative distribution function takes the value $q \leq 1$:

$$F(x_q) = \int_{-\infty}^{x_q} f(x)\,\mathrm{d}x = q.$$



- **The Full Width Half Maximum** (FWHM) specifies the width of the distribution at half the height of the maximum. In this way, the tails of the distribution are ignored.
  For a Gaussian distribution one has:

$$\mathrm{FWHM} = 2.35\,\sigma.$$

2.6 *Description of Discrete Data*

The data taken from an experiment (sample) constitute a data set $x_1, ..., x_N$. The underlying probability density $f(x)$ is not always known.

Distribution and parameters of the distribution must be determined from the measured data.

	Underlying distr.	data set
probability density	$f(x), F(x)$	$h(x)$ (frequency)
expectation value	$E(x) = \mu$	mean $\bar{x}$
variance	$\sigma^2(x) = \mathrm{Var}(x)$	variance $\sigma^2(x_1, ..., x_N)$

Sample Mean:

$$\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i$$

The sample mean is an unbiased estimator for the true mean μ:

$$\lim_{N \to \infty} = \mu$$

Variance:

$$\mathrm{Var}(x_1, ..., x_N) = \sigma_N^2(x_i, \mu) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

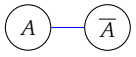
In general, the expected value μ is not known a priori, so μ must be estimated using the mean $\bar{x}$. However, the quantity $\sigma_N^2(x_i) = 1/N \sum_{i=1}^{N} (x_i - \bar{x})^2$ does not provide an unbiased estimator of the variance. It can be shown that an unbiased estimator is given by:

$$\sigma_{N-1}^2(x_i) := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

Covariance:

$$\mathrm{cov}(x, y) := \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

3 Probability distributions



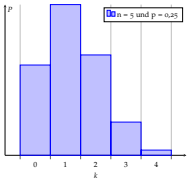
3.1 Binomial distribution

Consider an experiment with two possible outcomes A and $\overline{A}$:

$$P(A) = p,$$
$$P(\overline{A}) = 1 - p = q.$$

The probability that in n experiments the outcome A occurs k times is given by:

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



Binomial distribution

The binomial distribution has the following properties:

$$E(k) = n \cdot p,$$
$$\sigma^2(k) = n \cdot p(1-p),$$
$$\sigma(k) = \sqrt{n \cdot p(1-p)}.$$

Variance und standard deviation of the binomial distribution

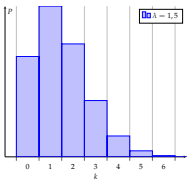
3.2 Poisson distribution

In the limit of an infinite number n of experiments, a vanishing probability p, and a finite product $n \cdot p = \lambda$, the binomial distribution approaches the **Poisson distribution**[1]:

$$\lim_{n\to\infty} P(k, p, n) = P(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

[1] after Siméon Denis Poisson, France, mathematician and physicist

A classical application example for the Poisson distribution is the number of radioactive decays in a given time period. For the Poisson distribution $P(k, \lambda)$ holds: $P(k, \lambda)$ heißt Wahrscheinlichkeitsdichte der Poisson-Verteilung. Es gilt:



$$\sum_{k=0}^{\infty} P(k, \lambda) = 1,$$

Normierung

$$E(k) = \lambda,$$

Erwartungswert

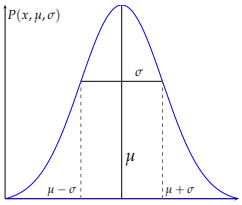
$$\sigma^2(k) = \lambda.$$

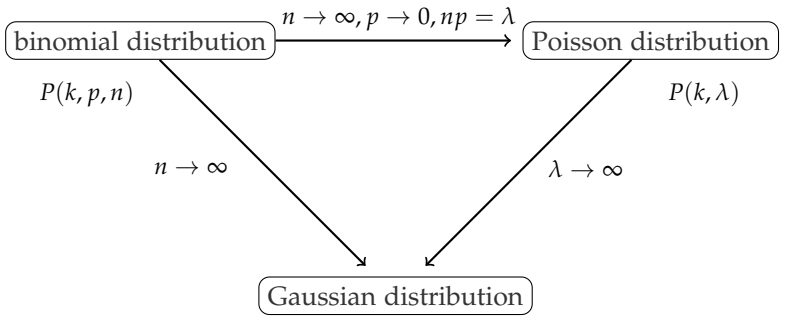
Varianz

3.3 *Gaussian distribution*

The **Gauss distribution**[2] or also **normal distribution** is of central importance in physics. Deviations of measured values from the mean value can be described by a Gaussian distribution in good approximations. Therefore, the error analysis as well as the error propagation is based to a large extent on the Gaussian distribution. It arises from the bimomial distribution for a large number if samples n and from the Poisson distribution for large expectation values λ.

[2] after Johann Carl Friedrich Gauss, Holy Roman Empire of German Nations, mathematician, astronomer, geodesist and physicist





The Gaussian distribution is an example of continous probabilty distribution. It is a symmetric distribution characterized by its mean μ and standard deviation σ:

$$P(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$
$$E(x) = \mu,$$
$$\mathrm{Var}(x) = \sigma^2.$$

Gaussian distribution

For a normally distributed measured value x, the probability to find x within $\pm n\sigma$ around the true value μ is given by:

$[\mu-\sigma,\mu+\sigma]$	$[\mu-2\sigma,\mu+2\sigma]$	$[\mu-3\sigma,\mu+3\sigma]$
68,3 %	95,4 %	99,7 %

3.3.1 *Central limit theorem*

If x_i are independently distributed random variables with mean μ and variance σ^2, then in the limit $n\to\infty$ the sum

$$X := \frac{1}{n}\sum_{i=1}^{n} x_i$$

is normally distributed with mean μ and variance σ^2/n.

3.3.2 *Two-dimensional Gaussian distribution*

For **independent** random variables x, y following Gaussian distributions for which (without loss of generality) $\mu_x = \mu_y = 0$, the probability density is given by

$$P(x,y) = P(x)P(y)$$

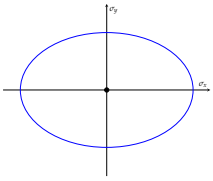
and one can thus write:

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)}.$$

Contours of constant probability density are ellipses. The ellipse enclosed by $[-\sigma_x, +\sigma_x]$ and $[-\sigma_y, +\sigma_y]$ is given by

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = 1.$$

Ellipsengleichung



In vector notation the equation for the ellipse reads:

$$(x,y)\begin{pmatrix} \frac{1}{\sigma_x^2} & 0 \\ 0 & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1,$$

that is

$$\vec{x}^{\mathsf{T}}\ \mathbf{B}\ \vec{x} = 1.$$

The matrix $\mathbf{B}$ is the inverse of the covariance matrix $\mathbf{C}$[3].

The probability density for a two-dimensional Gaussian can then be written as[4]:

$$P(x,y) = \frac{1}{2\sqrt{\det \mathbf{C}}} e^{-\frac{1}{2}\vec{x}^{\mathsf{T}}\ \mathbf{C}^{-1}\ \vec{x}}.$$

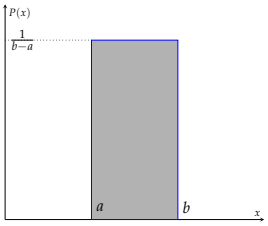
[3] for which we assume here that x and y are uncorrelated

[4] this also holds for non-vanishing covariances, see chapter 4

3.4 *Uniform distribution*

A large number of measurements can be described by the **uniform distribution**. Thus, it represents the simplest case of a detector which has a homogeneous response probability on the sample space in the interval $[a, b]$:

$$P(x) = \begin{cases} c = \frac{1}{b-a} & a \le x \le b \\ 0 & \text{sonst} \end{cases}$$



The uniform distribution has the following probabilities:

$$E(x) = \frac{1}{2}(a+b),$$

expectation value

$$\sigma^2(x) = \frac{(b-a)^2}{12}.$$

variance

The standard deviation of a uniform distribution is $\sigma = \frac{b-a}{\sqrt{12}}$.

3.5 Breit-Wigner distribution

By a **Breit[5]-Wigner[6] distribution** or also **Lorentz curve[7]** resonances can be described. This is particularly relevant when the natural linewidth can be resolved, as in the case of spectral lines or energy spectra of short-lived particles. The special case for an unshifted curve $a = 0$ with a half-width of $\Gamma/2 = 1$ is also called Cauchy distribution[8].

$$P(x, a, \Gamma) = \frac{1}{2\pi} \frac{\Gamma}{(x-a)^2 + \left(\frac{\Gamma}{2}\right)^2}.$$

[5] after Gregory Breit, Ukraine, physicist

[6] after Eugene Wigner, Kingdom of Hungary, physicist

[7] Hendrik Antoon Lorentz, Netherlands, mathematician and physicist

[8] after Augustin-Louis Cauchy, France, mathematician

The Breit-Wigner distribution distribution has the following properties:

$$E(x) = a,$$

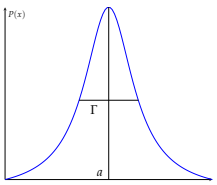
expectation value

$$\sigma^2(x) = \infty = \int_{-\infty}^{\infty} x^2 P(x)\,\mathrm{d}x = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x^2}{1+x^2}\,\mathrm{d}x.$$

variance

The variance and all higher moments diverge and are thus undefined because the function does not decay fast enough. Therefore, for the Breit-Wigner distribution, the width is given by the Full-Width-Half-Maximum (FWHM):

$$\mathrm{FWHM} := |x_2 - x_1| = \Gamma \qquad f(x_1) = f(x_2) = \frac{1}{2} f(x)$$



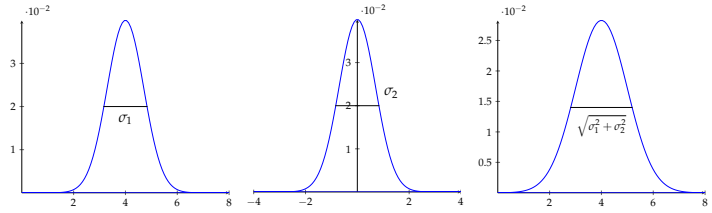
3.6 *Convolution of probability distributions*

A **convolution**[9] describes the effect of the resolving power of an apparatus on an observable. If the probability density of the observables is $f(x)$ and the one of the measurement uncertainty $g(y)$ and if the measured value $z = x + y$, then the convolution integral describes the probability density $h(z)$

$$h(z) = \int_{-\infty}^{\infty} f(t)g(z-t)\,\mathrm{d}t = \int_{-\infty}^{\infty} f(z-t)g(t)\,\mathrm{d}t.$$

An important example is the convolution of two Gaussian distributions $N(x;\mu_1,\sigma_1)$ and $N(y;\mu_2,\sigma_2)$. This results in a Gaussian distribution $N(z;\mu,\sigma)$ with

$$\mu = \mu_1 + \mu_2, \quad \sigma = \sqrt{\sigma_1^2 + \sigma_2^2}.$$



The covolution of a exponential distribution with a Gaussian results in an exponential distribution with a modified scale parameter.

[9] from lat. *convolvere* = to roll up

4 Treatment of measurement uncertainties

Preliminary remark: Measured values as interpreted in a statistical sense are often assumed to follow a Gaussian distribution around the true value. The measurement uncertainty σ is therefore identified with the standard deviation σ of a Gaussian distribution.

4.1 Uncertainty of the mean

For N repeated measurements $x_1, ..., x_N$ with uncertainties $\sigma_i = \sigma$ the **arithmetic mean** $\overline{x}$ is defined as

arithmetic mean

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

The uncertainty of the arithmetic mean is given by

$$\sigma(\overline{x}) = \frac{\sigma}{\sqrt{N}}.$$

For measured values with different uncertainties σ_i one used the **weighted mean**:

$$\overline{x}_G = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}.$$

The uncertainty of the weighted mean is

$$\sigma^2(\overline{x}_G) = \frac{1}{\sum \frac{1}{\sigma_i^2}}.$$

The weighted mean is often used to combine different independent measurements.

4.2 Error propagation

Suppose the observables x and y are independent and normally distributed around the true values μ_x and μ_y with standard deviations σ_x and σ_y, respectively. We are interested in the variation of a dependent quantity $z = f(x,y)$. With **Gaussian error propagation** one estimates the variance σ_z^2 by a first order Taylor expansion[1] :

$$f(x,y) = f(\mu_x,\mu_y) + \left.\frac{\delta f}{\delta x}\right|_{\mu_x,\mu_y}(x-\mu_x) + \left.\frac{\delta f}{\delta y}\right|_{\mu_x,\mu_y}(y-\mu_y) + \dots$$

[1] Under the assumption that
- higher order terms can be neglected
- derivatives at μ_x, μ_y can be approximated by the derivatives at x, y

In this approximation the expectation value $E[z] = E[f(x,y)]$ can be written as

$$E[f(x,y)] \approx f(\mu_x,\mu_y)$$

The variance can be written as

$$\begin{aligned}\mathrm{Var}[f(x,y)] &= E[(f(x,y) - f(\mu_x,\mu_y))^2] \\ &= E\left[\left(\left.\frac{\delta f}{\delta x}\right|_{\mu_x,\mu_y}(x-\mu_x) + \left.\frac{\delta f}{\delta y}\right|_{\mu_x,\mu_y}(y-\mu_y)\right)^2\right] \\ &= \left(\frac{\delta f}{\delta x}\right)^2 E[(x-\mu_x)^2] + \left(\frac{\delta f}{\delta y}\right)^2 E[(y-\mu_y)^2] + 2\frac{\delta f}{\delta x}\frac{\delta f}{\delta y}E[(x-\mu_x)(y-\mu_y)].\end{aligned}$$

The covariance $\mathrm{cov}(x,y) \equiv E[(x-\mu_x)(y-\mu_y)]$ vanishes for independent x and y and we obtain

$$\mathrm{Var}[f(x,y)] = \left(\frac{\delta f}{\delta x}\right)^2 \sigma_x^2 + \left(\frac{\delta f}{\delta y}\right)^2 \sigma_y^2.$$

The uncertainty of $z = f(x,y)$ can hence be written as

$$\sigma_z = \sqrt{\left(\left.\frac{\delta f}{\delta x}\right|_x\right)^2 \sigma_x^2 + \left(\left.\frac{\delta f}{\delta y}\right|_y\right)^2 \sigma_y^2}.$$

4.3 Covariance und correlation

If x and y are **not independent** of each other, their correlation ρ must be considered in the error propagation. It indicates how the change of one parameter relates to the other and can also be calculated via the multivariate variance. For the **covariance** we have

$$\mathrm{cov}(x,y) = \rho\,\sigma_x\sigma_y,$$

covariance

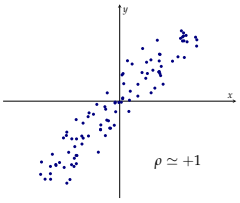
so that the variance of f can be calculated as

$$\sigma^2(f(x,y)) = \left(\frac{\delta f}{\delta x}\right)^2 \sigma_x^2 + \left(\frac{\delta f}{\delta y}\right)^2 \sigma_y^2 + 2\left(\frac{\delta f}{\delta x}\right)\left(\frac{\delta f}{\delta y}\right)\underbrace{\mathrm{cov}(x,y)}_{\rho\,\sigma_x\sigma_y}.$$

The covariance results either

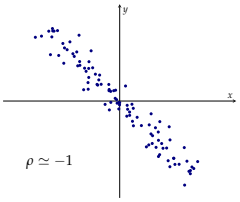
- from the experimental setup. An arrangement of two detectors, for example, where an event is detected either in one detector or necessarily in the other, implies a correlation of $\rho = -1$ on the count rates.

- from a known functional relationship. This is particularly relevant when multiple fit parameters are used from a curve fit, as these are rarely uncorrelated.
- from the scatter of the data. The correlation is then calculated as described in chapter 2.6.



Examples for correlated data:
The correlation coefficiant can take values

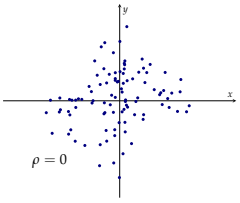
$$-1 \leq \rho \leq +1.$$



In the case of $\rho > 0$ one speaks of positive correlation, in the case of $\rho < 0$ of negative correlation. If $\rho = 0$ the variables are uncorrelated.

4.4 *Covariance matrix*

For the case where a function f depends on the variables $x_1, ..., x_n$, the bivariate dependence of the parameters is expressed by the covariance matrix C. This contains on its diagonal the variances of the measured variables x_i. The off-diagonal components are the covariances $\mathrm{cov}(x_i, x_j)$:



$$\mathbf{C} = \begin{pmatrix} \sigma_{x_1}^2 & & \mathrm{cov}(x_i, x_j) \\ & \ddots & \\ \mathrm{cov}(x_j, x_i) & & \sigma_{x_n}^2 \end{pmatrix} = \begin{pmatrix} \sigma_{x_1}^2 & & \rho\,\sigma_{x_i}\sigma_{x_j} \\ & \ddots & \\ \rho\,\sigma_{x_j}\sigma_{x_i} & & \sigma_{x_n}^2 \end{pmatrix}.$$

In linear approximation the uncertainty of f is then given by

$$\sigma^2(f(x,y)) = \sum_{i,j=1}^{n} \left(\frac{\delta f}{\delta x_i} \frac{\delta f}{\delta x_j} \mathbf{C}_{ij} \right)$$

oder in Vektorform:

$$\sigma^2(f(x,y)) = \nabla f^{\mathsf{T}} \cdot \mathbf{C} \cdot \nabla f.$$

For the general case of a set of m functions $\vec{y} = (f_1, ..., f_m)$ depending on $x_1, ..., x_n$ one has a (co-)variance for each pair f_k, f_l and the uncertainty $\sigma^2(f(x,y))$ is generalized to the error matrix $\mathbf{E}_{kl}$:

$$\mathbf{E}_{kl} = \sum_{i,j=1}^{n} \left(\frac{\delta f_k}{\delta x_i} \frac{\delta f_l}{\delta x_j} \underbrace{\rho_{ij}\,\sigma_{x_i}\sigma_{x_j}}_{\mathbf{C}_{ij}} \right).$$

With the transformation matrix $\mathbf{G}$

$$\mathbf{G}_{ki} := \frac{\delta f_k}{\delta x_i}$$

one obtains in vector form:

$$\mathbf{E} = \mathbf{G} \cdot \mathbf{C} \cdot \mathbf{G}^{\mathsf{T}}.$$

$\dim \mathbf{C} = (n, n)$
$\dim \mathbf{G} = (m, n)$
$\dim \mathbf{E} = (m, m)$

4.5 Covariance matrix and systematic uncertainties

If two measured quantities x_1, x_2 have statistical uncertainties $\sigma_{x_1}, \sigma_{x_2}$ and a common systematic uncertainty s

$$x_1 \pm \sigma_{x_1} \pm s,$$
$$x_2 \pm \sigma_{x_2} \pm s,$$

the two qantities are correlated

$$\mathrm{cov}(x_1, x_2) = E(x_1, x_2) - E(x_1)E(x_2) = s^2.$$

The corresponding covariance matrix is

$$\mathbf{C} = \begin{pmatrix} \sigma_{x_1}^2 + s^2 & s^2 \\ s^2 & \sigma_{x_2}^2 + s^2 \end{pmatrix}.$$

4.6 Estimators

The task in the evaluation of an experiment is to compare a general and simple model, which represents some kind of physical law or regularity, with the measured data. The distribution function describing the data is generally not known - but its form is, which is defined by a set of parameters $\vec{\lambda} = (\lambda_1, ..., \lambda_n)$. A simple example is radioactive decay - the complete data set depends only on the variable lifetime. From the sample one has to determine an estimator $\hat{\vec{\lambda}}$ of $\vec{\lambda}$ and its variance $\sigma(\hat{\vec{\lambda}})$. The requirements for a good estimator $S(x_1, ..., x_n)$ are that it has the following properties:

$N(t) = N_0 \cdot \mathrm{e}^{-t/\tau}$

- **unbiased**: The expectation value for the estimator of a parameter should be equal to the parameter (regardless of the number of measured values n):

$$E\left[S_\lambda(x_1, ..., x_n)\right] = \lambda.$$

- **consistent:** The estimator should converges to the true value in the limit of an infinite number of measurements:

$$\lim_{n\to\infty} \hat{\lambda} = \lambda.$$

- **efficient:** The standard deviation should be as small as possible:

$$E\left[(S-\lambda)^2\right] = \sigma^2(S) < \sigma^2(S_i),$$

where S_i is any estimator of λ.

Estimator for the mean of a Gaussian distribution

For a Gaussian distribution, the mean $\overline{x} = 1/n \sum x_i$ is an unbiased estimator for the true mean. It can be shown that its variance σ^2 is also minimal and thus the mean meets the above criteria.

Estimator for the variance of a Gaussian distribution

The estimator for the variance

$$\sigma^2 = \frac{1}{N-1}\sum_i^N (x_i - \overline{x})^2$$

would not be an unbiased estimate with the factor $1/N$. Illustratively, one degree of freedom of the data set has already been used for the determination of the mean $\overline{x}$, which is no longer available for the determination of further parameters.

4.7 χ^2 distribution

The χ^2 distribution can be defined as follows: If the random variables $x_1, ..., x_n$ follow a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, then the sum z of the n squares

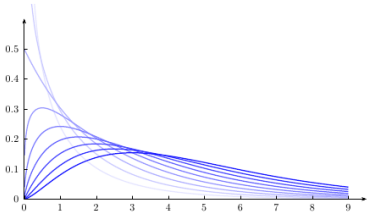
$$z = \chi^2 = x_1^2 + ... + x_n^2$$

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

$$f(z;n) = \frac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \qquad (z \geq 0)$$

where $\Gamma(x)$ is the Gamma function. Mean and variance are

$$E(z) = n, \quad \mathrm{Var}(z) = 2n.$$



The χ^2 distribution plays an important role as a measure of the quality of the description of measured data by a model. If one has a data set of n normally distributed measurements y_i at the locations x_i with standard deviations σ_i and a model prediction $f(x;\vec{\lambda})$ with predefined parameters, then for repeated measurements of the data set the quantity

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i;\vec{\lambda}))^2}{\sigma_i^2}$$

follows a χ^2 distribution with n degrees of freedom if the model correctly describes the expected values $\langle y_i \rangle$. If the m parameters have been determined by a fit to the data, the χ^2 defined in the previous equation follows an χ^2 distribution with $n - m$ degrees of freedom.

number of degrees of freedom n_F when fitting a model to data: number of measured values n minus the number of parameters m: $n_F = n - m$

Application: Hypothesis testing

Assuming that the model is correct and purely statistical errors are present, the χ^2 value of the fit should follow a χ^2 distribution in repeated measurements of the data set. The probability P of obtaining a worse (i.e., larger) value than the observed χ^2 for a given number of n degrees of freedom is called the p-value. It results from the cumulative distribution function:

$$p\text{-value} = P(\chi^2_{n_F}) = 1 - F(\chi^2_{n_F}).$$

Thus, the p value is the probability of obtaining a χ^2 value larger than the observed χ^2 value, assuming that the model used is correct. If the p-value is small, the hypothesis that the model used describes the data can be rejected. An arbitrary but commonly used criterion is $p\text{value} < 0.05$. Each degree of freedom is expected increase the χ^2 by 1:

$$\frac{\chi^2}{n} \approx 1.$$

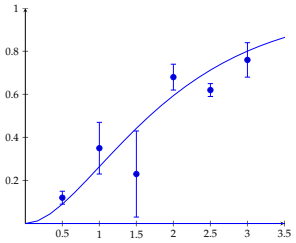
The value χ^2/doF[2], also called χ^2_{red}, reduced Chi^2, should thus be close to 1 and is often used to estimate the quality of a model. If the value is significantly different from 1, this can be due to several reasons:

[2] *dof = degrees of freedom*

- the used model is wrong or insufficient,
- the Gaussian statistic for the distribution of uncertainties is an incorrect assumption,
- there are unconsidered systematic uncertainties,
- the assumed standard deviation σ is too large (χ^2 becomes too small) or too small (χ^2 becomes too large).

4.8 *The Method of least squares*

The least squares method is intended to provide an unbiased and consistent estimator. For n data points $x_i, y_i \pm \sigma_i$, the function $f(x;\vec{\lambda})$ is determined by **minimizing** χ^2 . This provides the best estimate for the parameters $\vec{\lambda}$:



$$\arg\min_{\vec{\lambda}}(\chi^2) = \arg\min_{\vec{\lambda}}\left(\sum_{i=1}^{n}\frac{(y_i - f(x_i;\vec{\lambda}))^2}{\sigma_i^2}\right).$$

For m parameters $\lambda_1, ..., \lambda_m$ you get m equations:

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}\lambda_i} = 0.$$

Example: **Straight line fit**

$$f(x;\vec{\lambda}) = f(x;a_1,a_0) = a_1 x + a_0$$
$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - a_1 x_i - a_0)^2}{\sigma_i^2}.$$

Under the assumption that $\sigma_i \equiv \sigma$ for $i = 1...n$ one obtains

$$\frac{\mathrm{d}\chi^2}{\mathrm{d}a_0} = \frac{1}{\sigma^2}(-2)\sum(y_i - a_1 x_i - a_0) = 0,$$
$$\frac{\mathrm{d}\chi^2}{\mathrm{d}a_1} = \frac{x_i}{\sigma^2}(-2)\sum(y_i - a_1 x_i - a_0) = 0,$$

as estimator

$$\hat{a}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{cov}(x,y)}{\sigma^2(x)}$$
$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$$

with uncertainties

$$\sigma^2(\hat{a}_1) = \sum\left(\frac{\delta\hat{a}_1}{\delta x_i}\right)^2\sigma_i^2 \longrightarrow \sigma^2(\hat{a}_1) = \frac{\sigma^2}{n\left(\overline{x^2} - \bar{x}^2\right)},$$
$$\sigma^2(\hat{a}_0) = \frac{\sigma^2\overline{x^2}}{n\left(\overline{x^2} - \bar{x}^2\right)},$$
$$\mathrm{cov}(\hat{a}_1,\hat{a}_0) = -\frac{\sigma^2\bar{x}}{n\left(\overline{x^2} - \bar{x}^2\right)}.$$

Important: The covariance of $\hat{a}_1$ und $\hat{a}_0$ depends on the expectation value $\bar{x}$. For a straight line given by

$$f(x;a_1,a_0) = a_1\,(x - \bar{x}) + a_0$$

one obtains the uncorrelated solutions

$$\hat{a}_1 = \frac{\overline{xy}}{\overline{x^2}}, \qquad \sigma^2(\hat{a}_1) = \frac{\sigma^2}{n\overline{x^2}},$$
$$\hat{a}_0 = \bar{y}, \qquad \sigma^2(\hat{a}_0) = \frac{\sigma^2}{n}.$$

Generalization:
arbitrary functions und covariance matrices

Consider values y_i measured at loactions x_i

$$(x_i, y_i) \qquad \text{with } i = 1...n,$$

which are not independent, i.e., the covariance matrix $\mathbf{C}$ has non-vanishing off-diagonal components $\mathrm{cov}(y_i, y_j)$. These measured values y_i shall be described by a function $f(x;\vec{\lambda})$ which takes the values $\mu_i = f(x_i;\vec{\lambda})$ at the positions x_i. The χ^2 function then reads:

$$\chi^2 = (\vec{y} - \vec{\mu})^\mathsf{T}\, \mathbf{C}^{-1}\, (\vec{y} - \vec{\mu})\,.$$

where $\vec{\mu} = (\mu_1, ..., \mu_n)$. The minimum χ^2 is determined by $\frac{\mathrm{d}\chi^2}{\mathrm{d}\lambda_i} = 0$, i.e., the gradient of χ^2 with respect to the parameters vanishes.

In general, numerical minimization algorithms are needed to find the minimum of the χ^2 distribution. However, for functions which are linear in the parameters, a closed-form solution exists. For such a function we can write

$$f(x, \vec{\lambda}) = \sum_{k=1}^{m} a_k\,(x)\,\lambda_k,$$

and we obtain

$$\vec{\mu} = \mathbf{A}\,\vec{\lambda} \qquad \text{with } A_{ij} = a_j(x_i),$$

so that the χ^2 function reads

$$\chi^2 = (\vec{y} - \mathbf{A}\vec{\lambda})^\mathsf{T}\mathbf{C}^{-1}(\vec{y} - \mathbf{A}\,\vec{\lambda}).$$

The minimum χ^2 solution then is:

$$\vec{\hat{\lambda}} = \underbrace{(\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}}_{\mathbf{G}}\,\vec{y}$$

The covariance matrix $\mathbf{C}_\lambda$ of the parameters results from the transformation of the covariance matrix of the measured values:

$$\mathbf{C}_\lambda = \mathbf{G}\,\mathbf{C}\,\mathbf{G}^\mathsf{T}.$$

One obtains:

$$\mathbf{C}_\lambda = (\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A})^{-1}.$$

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\vec{\mu} = \begin{pmatrix} f(x_1;\vec{\lambda}) \\ \vdots \\ f(x_n;\vec{\lambda}) \end{pmatrix}$$

covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \begin{pmatrix} \sigma_{y_1}^2 & & \mathrm{cov}(y_i, y_j) \\ & \ddots & \\ \mathrm{cov}(y_j, y_i) & & \sigma_{y_n}^2 \end{pmatrix}$$

parameter
$\vec{\lambda} = (\lambda_1, ..., \lambda_m)$

example of a function which is linear in the parameters
$f(x, \vec{\lambda}) = \lambda_0 + \lambda_1 x + \lambda_2 x^2$
but not $f(x, \lambda) = \mathrm{e}^{-\lambda t}$

$$\nabla_\lambda \chi^2 = -2\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\left(\vec{y} - \mathbf{A}\,\vec{\lambda}\right) = 0$$
$$\mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\mathbf{A}\,\vec{\lambda} = \mathbf{A}^\mathsf{T}\mathbf{C}^{-1}\vec{y}$$

4.9 *Confidence intervals*

The result of a measurement is

$$\hat{\lambda} \pm \sigma\left(\hat{\lambda}\right)$$

What does this mean?
What is the correct interpretation of this result?
Where is the true value?

The frequentist approach:
This approach is based on the frequentist definition of probability. The sample space Ω is defined in advance. The probability of an event E is defined as the limit of the relative frequency n_E/n_G, where n_E is the observed number of event E and n_G is the total number of observations:

$$P(E) = \lim_{n_G \to \infty} \frac{n_E}{n_G}.$$

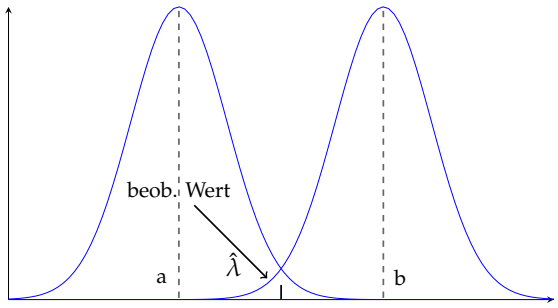
In the frequentist approach, $\hat{\lambda} \pm \sigma\left(\hat{\lambda}\right)$ is a statement about the constructed interval, which is to be understood in such a way that when the measurement is repeatedly performed, on average a proportion P of the constructed intervals contains the true value. In other words, the statement $\lambda \in [\hat{\lambda} - \sigma, \hat{\lambda} + \sigma]$ is true with a probability of 68%. The concept of a probability for the value of a parameter does not exist in the frequentist approach.

The Bayesian approach[3]**:**
This approach is based on the Bayesian notion of probability, according to which $P(H)$ indicates the degree of belief of a subject in a hypothesis H. In the case of a parameter λ to be determined, the degree of belief for a given value of λ is given by a probability density $P(\lambda)$. In order to interpret a result, it must always be considered under its predefined probability, the prior probability density. This is fixed by the experimenter and is thus subject to his subjective judgment. A measurement of a parameter λ together with the prior distribution leads to a new probability distribution $P'(\lambda)$. The apparent arbitrariness in the choice of the a prior distribution is often cited as a criticism of the Bayesian approach.

[3] after Thomas Bayes, England, mathmatician and Presbyterian minister

Confidence intervals:

In the frequentist approach, the probability density of the measured values with a fixed parameter is used to determine confidence intervals. In this method, the limits are determined in which the measured value lies with the given probability.



The confidence interval $[a, b]$ is defined in a way that

a is a lower limit for λ
b is an upper limit for λ

$$\int_{\hat{\lambda}_{\text{beob}}}^{\infty} g\left(\hat{\lambda}, a\right) \mathrm{d}\hat{\lambda} = \alpha \qquad \int_{-\infty}^{\hat{\lambda}_{\text{beob}}} g\left(\hat{\lambda}, b\right) \mathrm{d}\hat{\lambda} = \beta$$

hold for the given probabilites α, β. The interval $[a, b]$ contains the true value with a probability $P = 1 - \alpha - \beta$. The probability P is called confidence level.

4.10 *Maximum likelihood method*

While the χ^2-minimization method is based on the assumption of normally distributed measured values, the *maximum likelihood* method applies to general probability densities. If measured values $x_1, ..., x_n$ are distributed according to a probability density $f(x, \vec{\lambda})$, the likelihood function L is the product of the probability densities for each x_i:

$$L(x, \vec{\lambda}) = \prod_{i=1}^{n} f(x_i, \vec{\lambda}).$$

likelihood function

For the likelihood function, one considers the measured values as fixed and the parameters as variables. The maximum likelihood principle now states that the best estimate $\vec{\hat{\lambda}}_{\text{ML}}$ for the parameters $\vec{\lambda}$ is the one that maximizes the likelihood function:

$$\vec{\hat{\lambda}}_{\text{ML}} = \arg\max_{\vec{\lambda}} L(x, \vec{\lambda})$$

Numerically, it is often advantageous to use the logarithm of the likelihood function, the **log-likelihood function**:

$$\mathcal{L}(x, \vec{\lambda}) = \ln L(x, \vec{\lambda}) = \sum_{i=1}^{n} \ln f(x_i, \vec{\lambda}).$$

log-likelihood function

The best estimator $\vec{\hat{\lambda}}$ is given by the solutions to the equations

$$\frac{\delta}{\delta\lambda_j} L(x_i, \lambda_j) = 0, \quad j = 1, ..., m$$

where m is the number of parameters.

Example:

Radioactive decay is described by $f(t, \tau) = 1/\tau \cdot \mathrm{e}^{-t/\tau}$ where the only parameter is the mean lifetime τ. The log-likelihood function then reads:

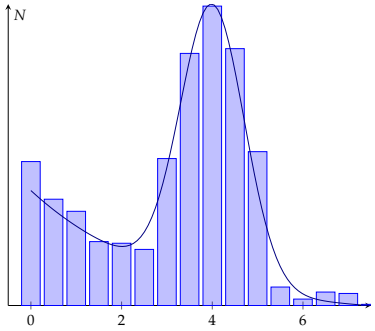
$$\mathcal{L}(t_1, ..., t_n, \tau) = \sum_{i=1}^{n} \left(-\ln \tau - \frac{t_i}{\tau} \right).$$

Maximizing the log-likelihood function gives the best estimator $\hat{\tau}$ of the lifetime:

$$\frac{\delta \mathcal{L}}{\delta \tau} = \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \Rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i = \bar{t}.$$

The maximum likelihood estimator for the lifetime in the radioactive decay is just the arithmetic mean of the measured times.

Maximum likelihood for histogramms



For small sample sizes, such as those found in histograms with **few entries per bin**, one cannot assume a normal distribution for the number of entries per bin. Either a binomial distribution or a Poisson distribution must be assumed, see chapter 3. Thus, the χ^2 method cannot be used and the maximum likelihood method is then the method of choice.

The probability for an event to appear in bin i is

$$p_i(\vec{\lambda}) = \int_{x_i-\Delta x_i/2}^{x_i+\Delta x_i/2} f(x;\vec{\lambda})\ \mathrm{d}x \approx f(x_i)\Delta x_i$$

It is usually a good choice to assume Poisson-Statistics. The total number of events then fluctuates according to the Poisson distribution $P(N_{\mathrm{tot}};\bar{N}_{\mathrm{tot}})$ around the mean $\bar{N}_{\mathrm{tot}}$ which can be either a model prediction or a free parameter. The expectation value for the number of entries in bin i predicted by the model is $\mu_i(\vec{\lambda}) = \bar{N}_{\mathrm{tot}} p_i$. The likelihood function can then be written as

$$L(N_1,...,N_n;\vec{\lambda}) \equiv L(\vec{\lambda}) = \prod_{i=1}^{n} P(N_i;\mu_i(\vec{\lambda})) = \prod_{i=1}^{n} \frac{\mu_i(\vec{\lambda})^{N_i}}{N_i!} e^{-\mu_i(\vec{\lambda})}$$

where N_i is the number of observed events in bin i. If we now go to the log-likelihood function and drop terms that do not depend on the parameters, we obtain the function

$$\tilde{\mathcal{L}}(\vec{\lambda}) = \sum_{i=1}^{n} N_i \ln \mu_i(\vec{\lambda}) - \mu_i(\vec{\lambda}) = -\bar{N}_{\mathrm{tot}} + \sum_{i=1}^{n} N_i \ln \mu_i.$$

The parameters which maximize this function are the maximum-likelihood estimators.

Variance of the maximum-likelihood estimator

To determine the variance of the maximum likelihood estimator $\hat{\vec{\lambda}}$, one can expand the log-likelihood function around its maximum:

$$\mathcal{L}\left(x_1,...,x_n,\vec{\lambda}\right) \approx \mathcal{L}\left(x_1,...,x_n,\hat{\vec{\lambda}}\right) + \underbrace{0}_{\text{1. Ord.}} + \frac{1}{2}\sum_{j,k}\left(\lambda_j-\hat{\lambda}_j\right)\left(\lambda_k-\hat{\lambda}_k\right)\underbrace{\frac{\delta^2\mathcal{L}}{\delta\lambda_j\delta\lambda_k}\Big|_{\lambda=\hat{\lambda}}}_{-B_{jk}} + ...$$

If a maximum likelihood estimator exists, $L\left(x_1,...,x_n,\vec{\lambda}\right)$ approaches a normal distribution in the limit of large number n of measurements. Correspondingly, then $\mathcal{L}\left(x_1,...,x_n,\vec{\lambda}\right)$ is a quadratic function in $\vec{\lambda}$. This is called *asymptotic normality*. Thus, around the maximum of the log-likelihood function we get

$$\mathcal{L}\left(x_1,...,x_n,\vec{\lambda}\right) \approx \mathcal{L}_{\max} - \frac{1}{2}\left(\vec{\lambda}-\hat{\vec{\lambda}}\right)^{\mathsf{T}}\mathbf{B}\left(\vec{\lambda}-\hat{\vec{\lambda}}\right)$$

and therefore for the likelihood function:

$$L\left(x_1,...,x_n,\vec{\lambda}\right) \approx L_{\max}\, \mathrm{e}^{-\frac{1}{2}\left(\vec{\lambda}-\hat{\vec{\lambda}}\right)^{\mathsf{T}}\mathbf{C}^{-1}\left(\vec{\lambda}-\hat{\vec{\lambda}}\right)}.$$

Here, too, the covariance matrix $\mathbf{C}$ can be identified with the inverse of the matrix $\mathbf{B}$ as in chapter 3.3. For uncorrelated parameters the standard deviation results from the diagonal elements

$$\sigma^2_{\lambda_j} = C_{jj} = B^{-1}_{jj} = \left(-\frac{\delta^2\mathcal{L}}{\delta\lambda_j\delta\lambda_j}\Big|_{\lambda=\hat{\lambda}}\right)^{-1}.$$

For the special case of only one parameter this reduces to

$$\sigma^2_{\lambda} = -\frac{1}{\frac{\delta^2\mathcal{L}}{\delta^2\lambda}\Big|_{\lambda=\hat{\lambda}}}.$$

The uncertainty of λ can alternatively be estimated using the values $\hat{\lambda}-\sigma^-_\lambda$ and $\hat{\lambda}+\sigma^+_\lambda$, for which the log-likelihood function takes the value $\ln L_{\max} - \frac{1}{2}$:

$$\ln L(\hat{\lambda}\pm\sigma_\lambda) = \ln L_{\max} - \frac{1}{2}$$

This usually yields asymmetric uncertainties. However, in the limit of a large data sample $\sigma_\lambda = \sigma^-_\lambda = \sigma^+_\lambda$ holds.

4.11 *Bayesian parameter estimation*

In Bayesian parameter estimation, all the information about the parameters $\vec{\lambda} = (\lambda_1,...,\lambda_n)$ to be estimated is contained in the posterior probability function $P(\vec{\lambda}|x)$. Using Bayes' theorem, this can be written as

$$P(\vec{\lambda}|x) = \frac{L(x|\vec{\lambda})\pi(\vec{\lambda})}{\int L(x|\vec{\lambda})\pi(\vec{\lambda})\,d\vec{\lambda}}$$

where as before the likelihood function is given by $L(x|\vec{\lambda}) = \prod_{i=1}^{n} f(x_i, \vec{\lambda})$. The prior probability distribution $\pi(\vec{\lambda})$ describes the knowledge about the parameters to be estimated before the measurement. The formula for $P(\vec{\lambda}|x)$ then specifies how this knowledge is updated based on the measured values x_i. The denominator $\int L(x|\vec{\lambda})\pi(\vec{\lambda})\, d\vec{\lambda}$ does not depend on $\vec{\lambda}$ and thus can be considered a normalization factor that can be determined by normalizing the posterior probability distribution to 1. If a constant prior distribution is chosen, the parameter vector $\vec{\lambda}$ for which the posterior distribution has a maximum corresponds to the maximum likelihood estimator. A constant likelihood distribution cannot be normalized and is thus a so-called improper prior distribution.

As an example, consider two measurements of a parameter λ. Given λ, let the probability density for measuring a certain value x_1 be given by a normal distribution $N(x_1; \lambda, \sigma_1)$ with mean λ and standard deviation σ_1. If we now consider this distribution in the Bayesian approach as a function of λ at fixed x_1, we get as prior distribution

$$\pi(\lambda) = N(\lambda; x_1, \sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(\lambda - x_1)^2}{2\sigma_1^2}}.$$

Taking into account the second measurement x_2 of λ (with standard deviation σ_2) we get

$$\begin{aligned} P(\lambda|x_2) \propto L(\lambda|x_2)\pi(\lambda) &= N(\lambda; x_2, \sigma_2)\pi(\lambda) \\ &= N(\lambda; x_2, \sigma_2) N(\lambda; x_1, \sigma_1). \end{aligned}$$

The product of two normal distributions is proportional to a normal distribution and we get as posterior distribution

$$P(\lambda|x_2) = N(\lambda; \mu, \sigma)$$

with

$$\mu = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \left(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2} \right) \quad \text{und} \quad \sigma = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}.$$

Thus, in this case, the Bayesian approach yields the formula for the weighted mean from section 4.1.

5 Practical tasks

The practical tasks are provided in the form of jupyter notebooks. These notebooks are to be completed with text and code. The notebooks with complete solution will then has to be sent to the tutor. The programming language to be used is python with appropriate packages (numpy, matplotlib, ...).

The practical tasks to be worked on are selected by the tutor.. Here are some possible tasks

Error propagation

1. In this task, Sympy is to be used to determine analytical expressions for the uncertainty of a dependent quantity $y = f(x_1, ..., x_n)$, where the measured values x_i may be correlated (`S01_error_prop_01.ipynb`).

Method of least squares

1. Linear least-squares fit
   If the fit function is linear in the fit parameters, the fit can be determined analytically. This is to be done in this task (`S01_least_squares_01.ipynb`)

2. Simultaneous χ^2 fit
   Here a model is simultaneously fit to several different data sets (`S01_least_squares_02.ipynb`).

Maximum likelihood method

1. Mean lifetime in an exponential decay
   Here the maximum likelihood methods is studied in more detail using the exponential decay as an example (`S01_ml_01.ipynb`).

2. Unbinned maximum likelihood fit
   A simple maximum likelihood fit with only one parameter (`S01_ml_02.ipynb`).

Bayesian parameter estimation

1. Determination of the posterior distribution for the parameter p of a binomial distribution (`S01_bayes_01.ipynb`).

Basic knowledge of Python is required for this experiment. The AP Python Introduction Course provides a good introduction. For

numerical minimization in χ^2 and maximum likelihood fits, the iminuit package is useful. The jupyter notebooks can be processed on your own computer or on the KIP jupyter server.