Advanced Lab's course
Introduction to Statistical Analysis
and Error Analysis

1.1 Aim of the lab

This lab allows students to meet several goals:

1. Introduction to error analysis and statistical analysis of the data
2. Introduction to basics of analysis and simulations with **Matlab** (or any other software used to do data analysis)
3. Performance of basic measurements of data with low and large statistics

1.2 Introduction

Data analysis and statistical methods play an essential role in science as it is a tool which gives possibility to treat uncertainties to data and draw conclusions. For experimental physics, it is also a tool to design and plan an experiment. Statistical methods of data analysis allow to identify independent and depended variables in the system, decide which model (theory) should be chosen for data description and etc. This laboratory manual is aimed to give students a brief introduction to data analysis techniques, however it is highly recommended to read further literature [1, 2, 3].

1.3 Elements of probability theory: distribution, moments of distribution, types of distribution

During the experiment one deals with stochastic processes, during which each next trial is independent on previous result. Random processes are described by the *probability density* function $P(x)$ (term *distribution* function is also used). Depending on the nature of variable "x", distribution $P(x)$ can be continuous or discrete. If variable "x" is discrete, then the probability to obtain "x_i" equals to $P(x_i)$. If variable "x" is continuous, then the probability to obtain "x" in the region from x to $x + dx$ equals to $P(x)dx$.
Usually, distribution probability is normalised to 1. In case of continuous distribution:

$$\int_{-\infty}^{+\infty} P(x)dx = 1 \tag{1.1}$$

In case of discrete distribution:

$$\sum_i P(x_i) = 1 \tag{1.2}$$

The probability $p(x)$ to find x between certain values in case of continuous distribution is defined in the following way:

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} P(x)dx \tag{1.3}$$

In case of discrete distribution, integration should be replaced with summation:

$$p(x_1 \leq x \leq x_2) = \sum_{x=x_1}^{x_2} P(x) \tag{1.4}$$

1.3.1 Moments of distribution

Theoretical mean value μ of variable x is defined as the first moment about zero of the distribution $P(x)$. Mathematically it can be expressed in the following way:

$$\mu = \int xP(x)dx \tag{1.5}$$

The variance of the distribution σ^2 is defined as a second moment about the mean value:

$$\sigma^2 = \int (x-\mu)^2 P(x)dx \tag{1.6}$$

It has a meaning of average squared deviation of variable "x" from its mean value μ. Square root of variance, σ, is called standard deviation, and used to describe the width of the distribution. Value of σ allows to judge about the magnitude of fluctuation of random variable "x" around its mean value σ. Also one can introduce higher moments of distribution, but they are rarely used in experimental physics. In case of discrete distributions, integration in equations (1.5, 1.6) should be changed to summation.

$$\mu = \sum_i x_i P(x_i) \tag{1.7}$$

$$\sigma^2 = \sum_i (x_i-\mu)^2 P(x_i) \tag{1.8}$$

Students are welcome to study literature for more information about it [1, 2, 3].
The correlation (dependency) between parameters x and y is described by covariance:

$$cov(x,y) = \int (x-\mu_x)(y-\mu_y)P(x,y)dxdy \tag{1.9}$$

where $P(x,y)$ is a 2-dimensional distribution of variables "x" and "y". It is also convenient to introduce correlation coefficient ρ ($\rho \in [-1;1]$):

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{1.10}$$

The value of ρ gives information about the dependency between parameters. If $|\rho|= 1$, then it is a case of perfect linear correlation. If $\rho = 0$, variables are independent.

1.3.2 Types of distributions: Binomial, Poisson, Gaussian

The Binomial distribution describes processes in which outcome of a trial is dichotomous (*yes* or *no*, *head* or *tail* and etc.). The probability of r positive (or negative) outcomes in N trials is calculated in the following way:

$$P(r) = \frac{N!}{r!\,(N-r)!} p^r (1-p)^{N-r} \tag{1.11}$$

where p is the probability of success (or failure) of one single trial. Examples of binomial distribution can be found in Fig. 1.1. It is easy to see that binomial distribution is discrete. Theoretical mean value μ equals to:

$$\mu = \sum_r rP(r) = Np \tag{1.12}$$

The variance of the distribution equals to:

$$\sigma^2 = \sum_r (r-\mu)^2 P(r) = Np(1-p) \tag{1.13}$$

Binomial distribution can be approximated with Poisson or Gaussian distributions, depending on the parameters.

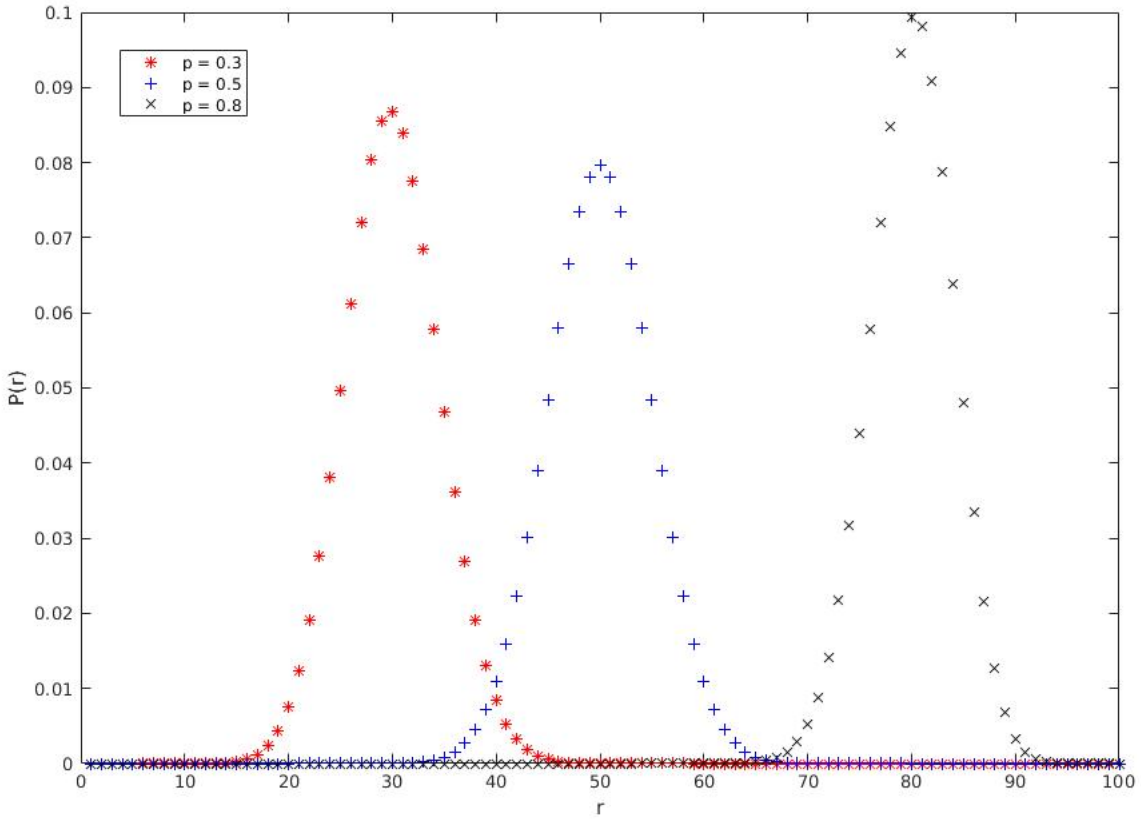

Figure 1.1: Examples of binomial distribution with different parameter "p". Details can be found in the legend.

The Poisson distribution rises as a limit case of Binomial distribution, when the following conditions are satisfied:

1. Events are independent within any time interval.
2. Probability of one event $p \to 0$, while the amount of trials $N \to \infty$.
3. Mean value $\mu = Np$ stays finite.

Probability to observe r events under the conditions listed above is calculated in the following way:

$$P(r) = \frac{\mu^r e^{-\mu}}{r!} \tag{1.14}$$

Poisson distribution has many examples in every day life, for example, the radioactive decay of ^{238}U isotope. This nucleus is an alpha-decayer with half-life $T_{1/2} \approx 4.5 \times 10^9$ years. From laws of nuclear physics, it is known that nuclei act independently from each other during radioactive decay. The probability of one uranium nucleus to decay in one second is $\lambda = \ln 2/T_{1/2} \approx 5 \times 10^{-18}$ s^{-1}. As one can see, probability of one single event is extremely small. But, suppose that we have 1 g of ^{238}U, which contains $N \approx 2.5 \times 10^{21}$ isotopes, the total number of decay events in one second can be large and can be approximately described by a binomial distribution with $p = 5 \times 10^{-18}$ s^{-1} and $N = 2.5 \times 10^{21}$. By equation 1.12, the mean value of the total number of decay events in one second is $\mu = Np \approx 12.5 \times 10^3$ decays/s $= 12.5$ decays/ms. If we want a distribution function of the total number of decay events in one millisecond, the Poisson distribution perfectly describes this situation.

Poisson distribution is discrete and has only one parameter μ, which stands for the mean value. It can be shown, that variance of the distributions equals to:

$$\sigma^2 = \mu \tag{1.15}$$

On Fig. 1.2 one can find several examples of Poisson distribution with different values of μ. As it is seen, with small values of μ, Poisson distribution is asymmetric. But with relatively large values of μ distributions becomes symmetric and can be approximated by the Gaussian distribution. During such approximation, the discrete nature of Poisson distribution can be neglected.

The Gaussian or normal distribution is one of the most important distributions in statistics and for data analysis in particular. As it was mentioned above, binomial and Poisson distributions can be approximated by Gaussian distribution. Also, it is known from central limit theorem (CLT), that if a set of independent random variables (even with distributions different from normal) are added, their sum will tend to normal distribution. On practice, it means that many processes can be approximated with Gaussian distribution. For example, instrumental errors are generally distributed normally. The Gaussian distribution is a symmetric, continuous distribution, with density given by:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big) \tag{1.16}$$

It is dependent of two parameters: mean value μ and standard deviation σ. Example of Gaussian distribution can be found in Fig. 1.3. Parameter σ characterises the width of Gaussian distribution. It is also common to use another parameter to describe the width, the full width of half maximum (FWHM). It can be shown, that:

$$\text{FWHM} = 2\sqrt{2\ln 2}\sigma \approx 2.35\sigma \tag{1.17}$$

Another important feature of normal distribution is the area under it, depending on the intervals.

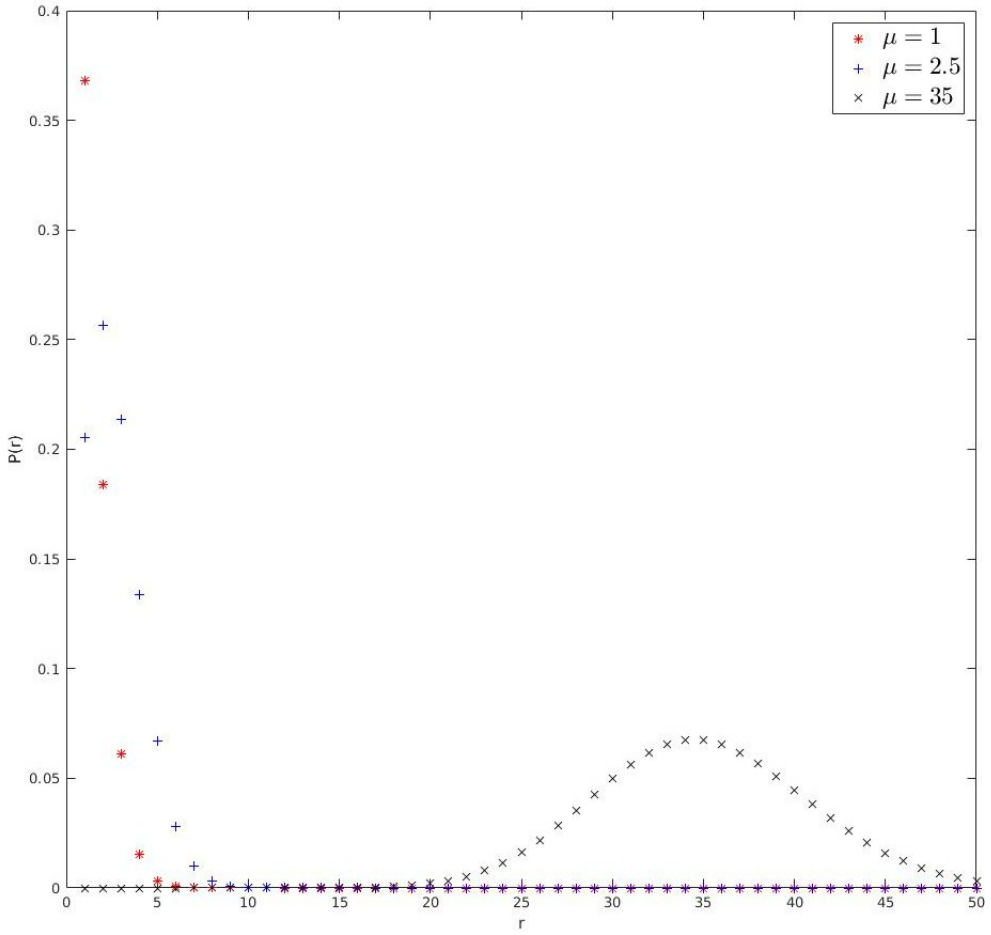

Figure 1.2: Examples of Poisson distribution with different parameter μ. Details can be found in the legend.

Example can be found in Fig. 1.4. The area in the interval $[\mu - \sigma; \mu + \sigma]$ equals to $\approx 68\%$. For the interval within $\pm 2\sigma$ and $\pm 3\sigma$, the areas are 95.5% and 99.7% respectively. This result is very important for data representation and means that if one will present experimental results within only 1σ, then there is approximately a probability of 1/3 that the true value will be outside the region. If results are presented with 2σ interval, this probability is less than 5%.

1.4 Statistical methods in experimental physics

In general case, he estimation of the population mean of a sample $x \in \{x_1, x_2, x_3, ..., x_n\}$ is calculated in the following way:

$$\bar{x} = \frac{1}{n} \sum_i x_i \tag{1.18}$$

The estimation of the population standard deviation s is given by:

$$s = \sqrt{\frac{1}{(n-1)} \sum_i (x_i - \bar{x})^2} \tag{1.19}$$

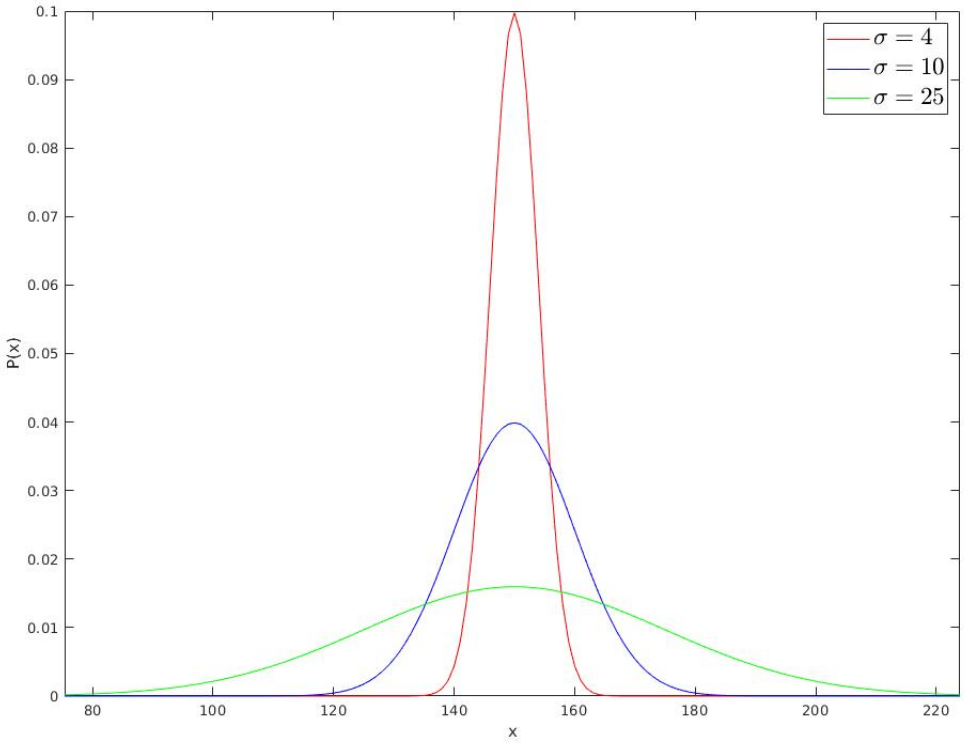

Figure 1.3: Examples of Gaussian distribution with different parameters σ. Details can be found in the legend.

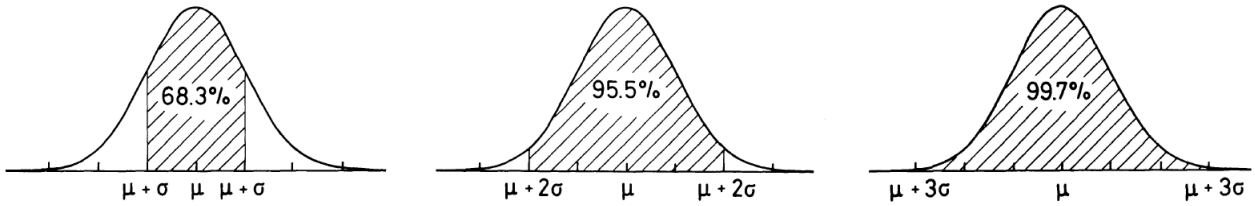

Figure 1.4: Area under Gaussian distribution, contained between different intervals. Figure is taken from [2].

The estimated standard deviation of $\bar{x}$ is given by:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{1.20}$$

The estimated mean $\bar{x}$ can be described by the normal distribution with the mean $\bar{x}$ and standard deviation $s/\sqrt{n}$, if the sample size is large, typically $n > 30$. Standard deviation $\sigma_{\bar{x}}$ has a meaning of a statistical error and gives a probability of 66% that the population mean μ will be found in the region $\bar{x} - \sigma_{\bar{x}} \leq x \leq \bar{x} + \sigma_{\bar{x}}$. In physics, confidence interval is usually within 2σ with the confidence level $\alpha = 95\%$

$$\bar{x} - 2\sigma_{\bar{x}} \leq x \leq \bar{x} + 2\sigma_{\bar{x}} \tag{1.21}$$

If the sample size is small, $n < 30$, we need to use the student's t coefficient $t_{\alpha,n}$. Its value depends on the sample size n and the confidence level α, and their values can be found from the

table 1.1 with the confidence level $\alpha = 95\%$. The confidence interval with confidence level $\alpha = 95\%$ is given by

$$\bar{x} - t_{0.95,n}\sigma_{\bar{x}} \leq x \leq \bar{x} + t_{0.95,n}\sigma_{\bar{x}} \tag{1.22}$$

When $n > 30$, $t_{0.95,n}$ is reduced to 2 and the confidence interval is reduced to 1.21.

Table 1.1: Values of student's t coefficient for confidence level $\alpha = 95\%$.

n	2	3	4	5	6	7-8	9-10	30-∞
$t_{0.95,n}$	12.7	4.3	3.2	2.8	2.6	2.4	2.3	2

During the experiment, every quantity contains two types of error: statistical and systematical. Final result for value x is written in the following way:

$$x = \bar{x} \pm \sqrt{\sigma_{stat}^2 + \sigma_{syst}^2} \tag{1.23}$$

1.4.1 Propagation of Errors

This section covers a method to estimate the total error σ_f of a function $f = f(x, y, z)$, when individual error of variables x, y and z are known, namely σ_x, σ_y and σ_z. It is considered that variables x, y and z are independent. It can be shown that σ_f can be expressed in the following way:

$$\sigma_f = \sqrt{\left(\frac{\partial f}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial f}{\partial y}\sigma_y\right)^2 + \left(\frac{\partial f}{\partial z}\sigma_z\right)^2} \tag{1.24}$$

In more general case, if variables x, y and z are dependent, the covariances between variables must be taken into account. Additional information on this case can be found in [1, 2, 3].

For example, if $f(x, y, z) = x + y + z$, their partial derivatives are 1. The total error is

$$\sigma_f = \sqrt{(\sigma_x)^2 + (\sigma_y)^2 + (\sigma_z)^2} \tag{1.25}$$

This means that if f is the sum of x, y and z, the resulting error σ_f is the square root of the sum of squares of individual errors.

Consider the case that $f(x, y, z) = xyz$. The total error is

$$\sigma_f = \sqrt{(yz\sigma_x)^2 + (xz\sigma_y)^2 + (yz\sigma_z)^2} \tag{1.26}$$

$$\frac{\sigma_f}{xyz} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + \left(\frac{\sigma_z}{z}\right)^2} \tag{1.27}$$

$$\frac{\sigma_f}{f} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + \left(\frac{\sigma_z}{z}\right)^2} \tag{1.28}$$

This means that if f is the product of x, y and z, the resulting relative error σ_f/f is the square root of the sum of squares of individual relative errors.

1.5 Fitting methods

A fitting method is a process of constructing a curve, or mathematical function, that has the best fit to the experimental data.

1.5.1 χ^2 method

The χ^2 method is useful for testing how good the fit is. χ^2 analysis [2, 4, 5] is based on minimizing the difference between experimental data and the fitting function. Suppose that we have a series of experimental data (x_i, y_i) with the error of y_i denoted by σ_i, and we use a fitting function $y = f(x; a_1, a_2, ..., a_m)$ to fit the experimental data y_i by varing different values of the m parameters a_j. The deviation bewteen the experimental data and the fitting curve can be measured by the following quantity χ^2:

$$\chi^2 = \sum_{i=1}^{n} \Big(\frac{y_i - f(x_i; a_1, a_2, ..., a_m)}{\sigma_i} \Big)^2 \tag{1.29}$$

where summation is taken over all experimental points. By finding the minimum value of χ^2 by trying all possible values of the m parameters a_j, the best values of the parameters can be found and hence the best-fit curve can be found. Every best-fit values of the parameters need to satisfy the following minimization condition:

$$\frac{\partial \chi^2}{\partial a_j} = 0 \qquad \text{for } 1 \leq j \leq m \tag{1.30}$$

1.5.2 Linear fit

We can use the linear fit as an example for using the χ^2 method. Suppose the fitting function is $y = f(x; a, b) = ax + b$, with the parameters a and b. By the equation 1.29, the χ^2 is

$$\chi^2 = \sum_{i=1}^{n} \Big(\frac{y_i - ax_i - b}{\sigma_i} \Big)^2 \tag{1.31}$$

To find the best-fit values of a and b, by the equation 1.30, we need to solve the following system of equations:

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{n} \frac{(y_i - ax_i - b)x_i}{\sigma_i^2} = 0 \tag{1.32}$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{n} \frac{(y_i - ax_i - b)}{\sigma_i^2} = 0 \tag{1.33}$$

To simplify the notation, we can define the following sums

$$S_0 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \qquad S_x = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \qquad S_y = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \qquad S_{yy} = \sum_{i=1}^{n} \frac{y_i^2}{\sigma_i^2} \qquad S_{xy} = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}$$

Equations 1.32 and 1.33 can be simplified to

$$-2(S_{xy} - aS_{xx} - bS_x) = 0 \tag{1.34}$$

$$-2(S_y - aS_x - bS_0) = 0 \tag{1.35}$$

By solving the liner equations, the best-fit values of a and b are

$$a = \frac{S_0 S_{xy} - S_x S_y}{S_0 S_{xx} - S_x^2} \tag{1.36}$$

$$b = \frac{S_y S_{xx} - S_x S_{xy}}{S_0 S_{xx} - S_x^2} \tag{1.37}$$

The errors of a and b are given by the further reading [2].

$$\sigma_a^2 = \frac{S_0}{S_0 S_{xx} - S_x^2} \tag{1.38}$$

$$\sigma_b^2 = \frac{S_{xx}}{S_0 S_{xx} - S_x^2} \tag{1.39}$$

To measure the quality of the fit, we can divide the χ^2 by the degree of freedom $n-2$.

$$\frac{\chi^2}{n-2} \tag{1.40}$$

The fit is good if the expression 1.40 is close to 1.

1.5.3 Analysis of data in case of large statistics

The technique, discussed at the beginning of section 1.4, can always be used . But in the case of large statistics of experimental data, we can analyse the data by fitting a histogram by a normal distribution, because the histogram is reliable only if the number of experimental data is large. We can do many sets of experiments. And, for each experiments, we collect a fixed number of the experimental data and calculate its means. According to the central limit theorem, the mean of a experiment can be approximately described by a normal distribution.

In figure 1.5, it shows the examples of the histograms of means of a radius for a particle r and square displacement of a Brownian particle $\langle x^2 \rangle$. The histogram allows us to measure the mean value $\bar{x}$ and its standard deviation $\sigma_{\bar{x}}$, described in equation 1.18 and 1.20, by fitting the histogram by a normal distribution.

Final result with confidence interval of confidence level $\alpha = 66\%$ is

$$x = \bar{x} \pm \sqrt{(\sigma_{stat})^2 + (\sigma_{syst})^2} \tag{1.41}$$

Final result with confidence interval of $\alpha = 95\%$ is

$$x = \bar{x} \pm \sqrt{(2\sigma_{stat})^2 + (\sigma_{syst})^2} \tag{1.42}$$

1.6 Experimental Setup and Procedure

Experimental equipment consists from following parts:

1. Standard micrometer
2. Thin aluminium plate

Below a student can find a list of "must-do" tasks, but students are highly motivated to expand to data analysis further:

1. Measure the thickness of a thin aluminium plate, using standard micrometer. Obtain data of large (200 points) and low (10 points) statistics. Estimate mean value (1.18) and standard deviation (1.20) for each case. Fit each data set with a distribution function, extract mean value and standard deviation from the fitting. Calculate χ^2 (1.29) for each case and define "goodness" of the fit. Present your results with 95% confidence level, taking into account statistical and systematical errors.

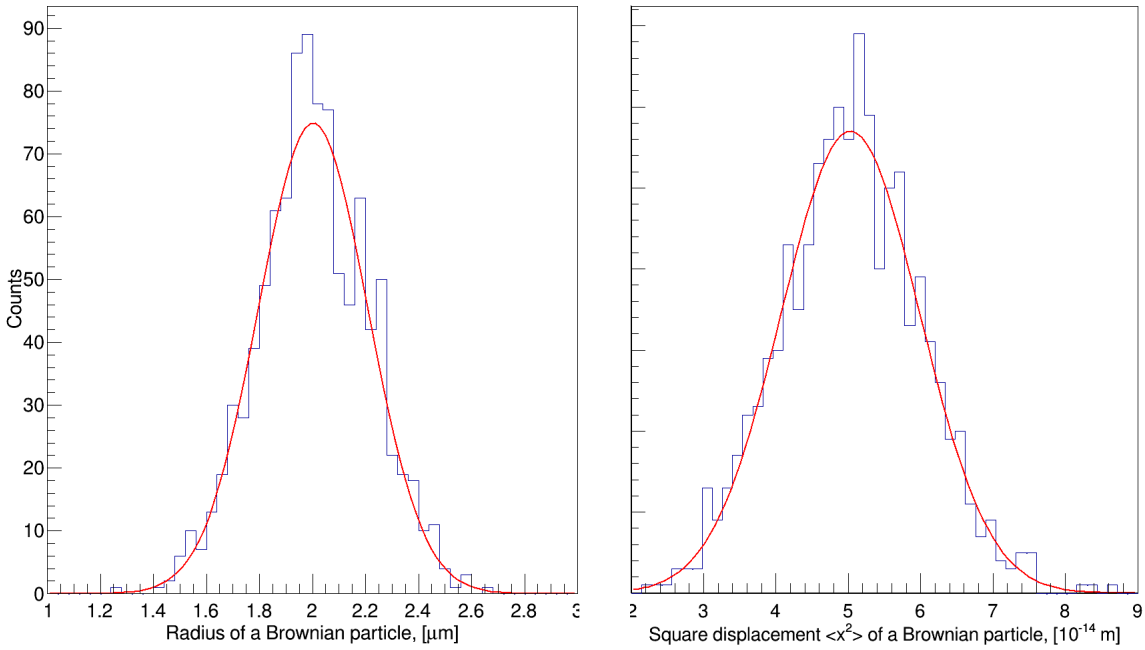

Figure 1.5: Example of the distribution of a radius for a particle and square displacement of a Brownian particle. Data is generated using Gaussian distribution in both cases.

2. Simulate Landau distribution sitting on a background curve (exponential decay). Add statistical fluctuations and error bars (take as $\sqrt{N}$) to data points. Fit the data, estimate "goodness" of the fit, extract parameters of the distribution and compare them with the simulated values. Present your results with 95% confidence level, taking into account statistical and systematical errors.

3. Simulate two Gaussian distributions party merging with each other, and sitting on the background curve (exponential decay). Simulate several cases with Gaussians of the same and different statistics, different widths of Gaussian distributions. Fit the data, estimate "goodness" of the fit, extract parameters of the distribution and compare them with the simulated values for each case. Present your results with 95% confidence level, taking into account statistical and systematical errors.

1.7 Discussion

To successfully pass the lab, a student must complete the following tasks:

1. What is the distribution function of the thickness of aluminium plate? How many parameters does it have?

2. In what cases does the χ^2 analysis method can be used? Derive χ^2 analysis method from maximum likelihood method.

3. Other questions

Bibliography

[1] James F. (2008). Statistical Methods in Experimental Physics.

[2] Leo W. R. (1994). Techniques for Nuclear and Particle Physics Experiments. p. 81

[3] Taylor J. R. (1997). An Introduction to Error Analysis.

[4] Steve Baker and Robert D. Cousins (1984). Clarification of the use of CHI-square and likelihood functions in fits to histograms. Nuclear Instruments and Methods in Physics Research, Volume 221, Issue 2, Pages 437 - 442

[5] R. Andrae, T. Schulze-Hartung, P. Melchior (2010). Dos and don'ts of reduced chi-squared. arXiv:1012.3754v1.