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 data analysis)
3. Performance of basic measurements of data with low are large statistics

1.2 Introduction

Data analysis and statistical methods play 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, than the probability to obtain "x_i" equals to $P(x_i)$. If variable "x" is continuous, than 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. 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:

$$cov(x,y) = \int (x-\mu_x)(y-\mu_y)P(x,y)dxdy \tag{1.7}$$

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

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{1.8}$$

The value of ρ gives information about the dependency between parameters. If $|\rho|= 1$, than 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:

$$P(r) = \frac{N!}{r!\,(N-r)!} p^r (1-p)^{N-r} \tag{1.9}$$

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:

$$\mu = \sum_r rP(r) = Np \tag{1.10}$$

The variance of the distribution equals to:

$$\sigma^2 = \sum_r (r-\mu)^2 P(r) = Np(1-p) \tag{1.11}$$

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

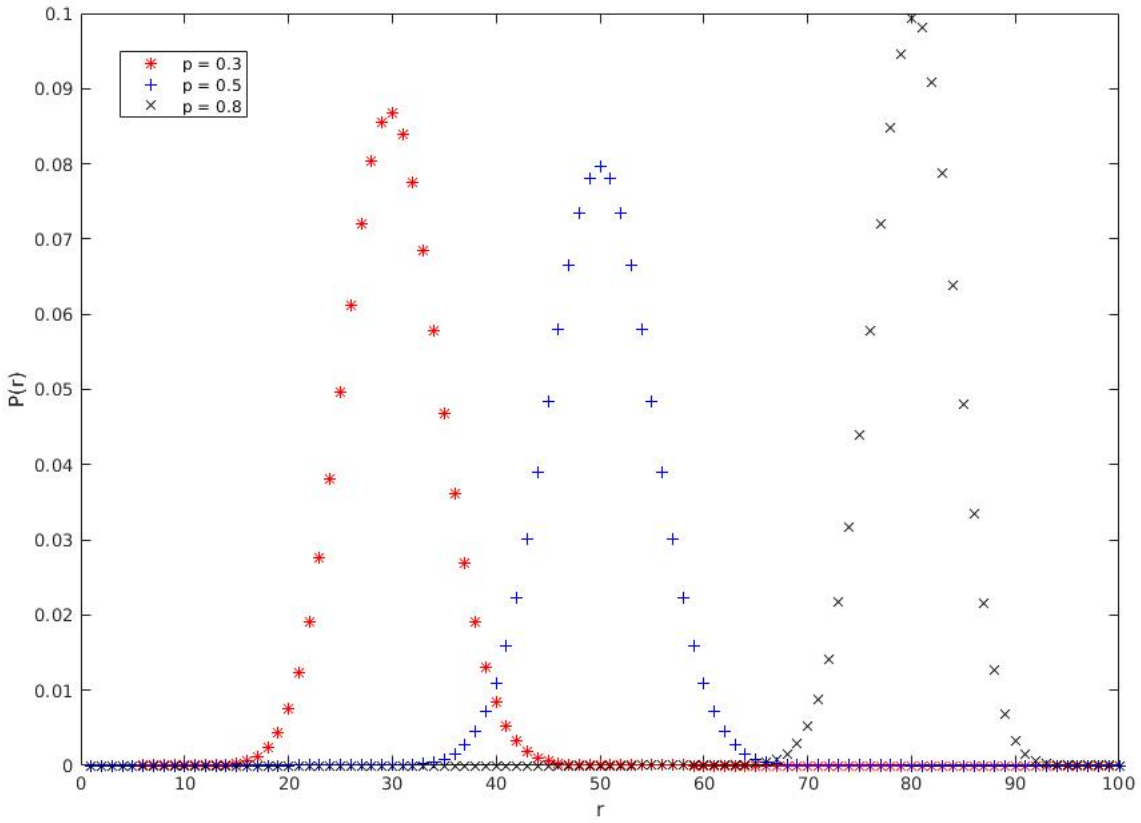

Figure 1.1: Examples of binomial distribution with different parameter "p". Details can be found in the legend.

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:

$$P(r) = \frac{\mu^r e^{-\mu}}{r!} \tag{1.12}$$

Poisson distribution has many examples in every day life. For example, radioactive decay of ^{238}U isotope. This nuclei is an alpha-decayer with $T_{1/2} \approx 4.5 \times 10^9$ years. From laws of nuclear physics, it is knows, that nuclei act independently from each other during radioactive decay. The probability of one uranium nucleus to decay is $\lambda = \ln 2/T_{1/2} \approx 5 \times 10^{-18}$ s^{-1}. As one can see, probability of a single event is very small. But, suppose, that we have 1 g of ^{238}U, which contains $N \approx 2.5 \times 10^{21}$ isotopes. The mean value remains constant in time and equals to $\mu = Np \approx 12.5 \times 10^3$ decays/s.

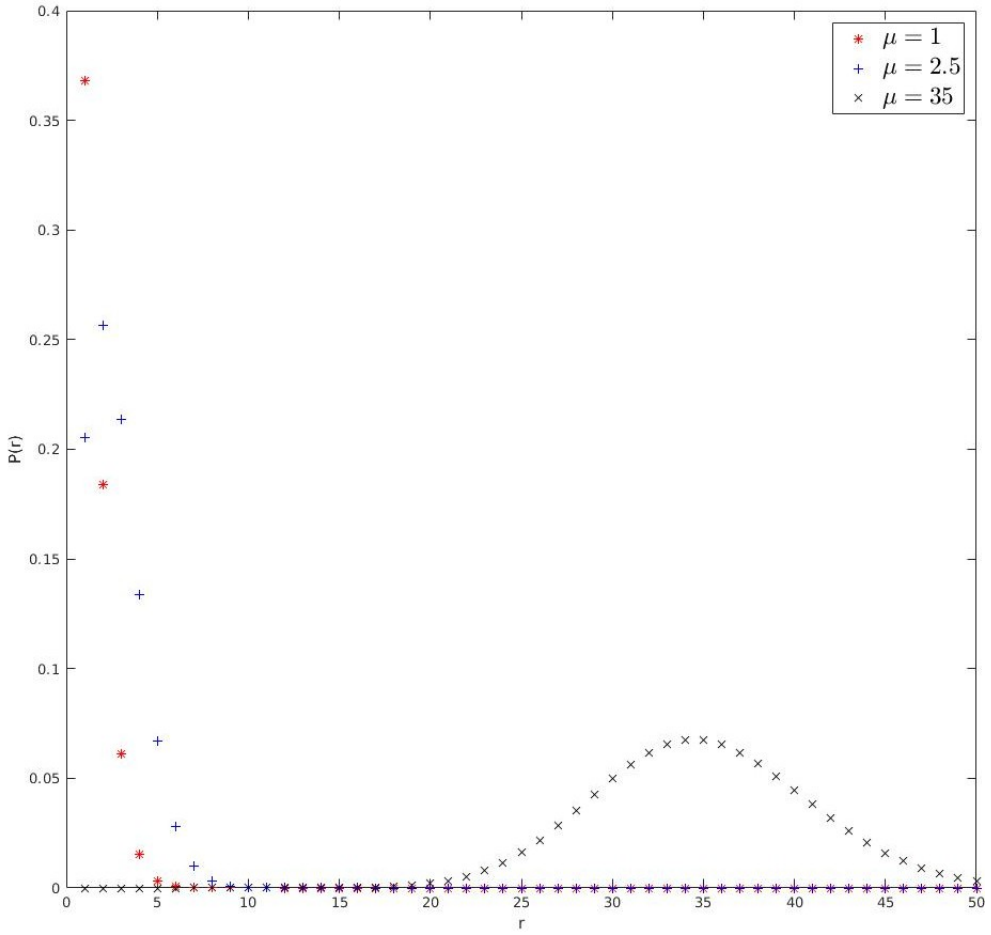

Figure 1.2: Examples of Poisson distribution with different parameter μ. Details can be found in the legend.

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:

$$\sigma^2 = \mu \tag{1.13}$$

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 should 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 one. 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:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big) \tag{1.14}$$

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.

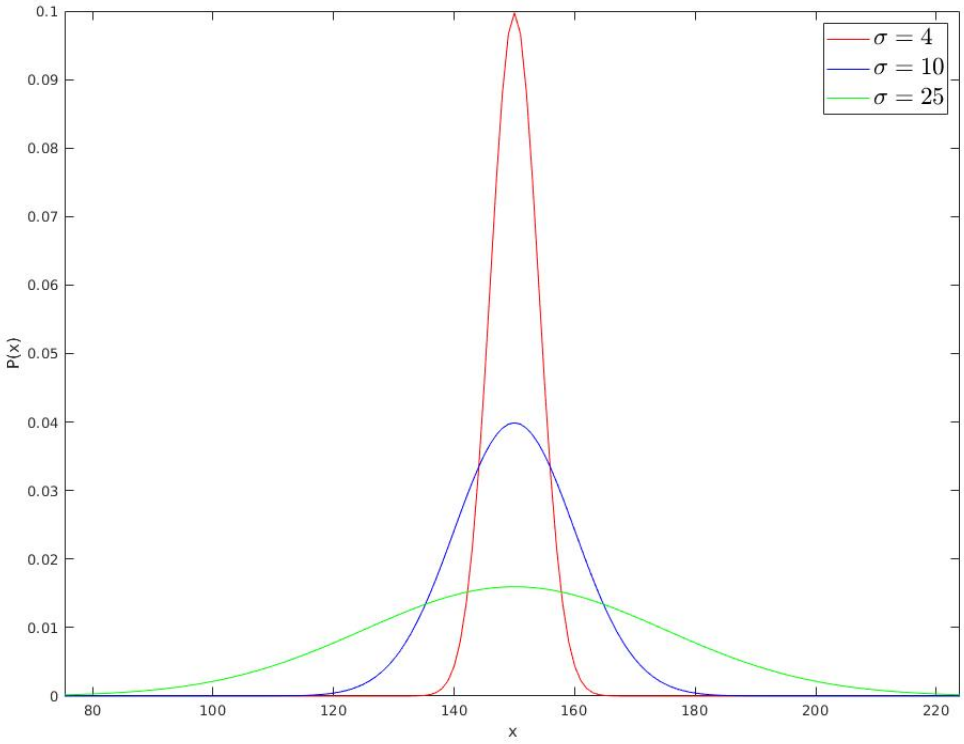

Figure 1.3: Examples of Gaussian distribution with different parameters σ. Details can be found in the legend.

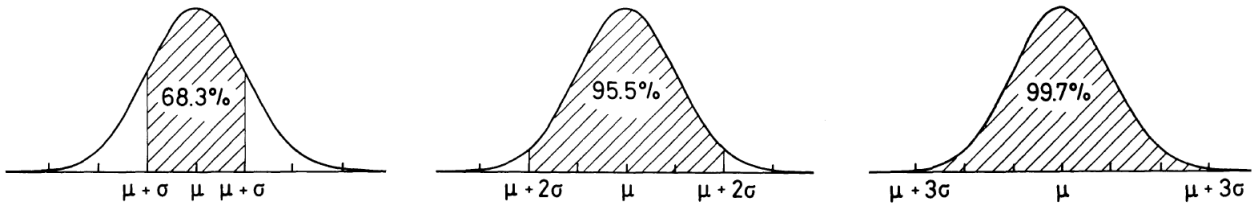

Figure 1.4: Area under Gaussian distribution, contained between different intervals. Figure is taken from [2].

It is also common to use another parameter to describe the width - full width of half maximum (FWHM). It can be shown, that:

$$\text{FWHM} = 2\sqrt{2\ln 2}\sigma \approx 2.35\sigma \tag{1.15}$$

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

1.4 Statistical methods in experimental physics

In general case, average value of a sample $x \in [x_1, x_2, x_3, ..., x_n]$ is calculated in the following way:

$$\bar{x} = \frac{1}{n} \sum x_i \tag{1.16}$$

Standard deviation $\sigma_{\bar{x}}$ of the average value $\bar{x}$ is defined as following:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n(n-1)} \sum (\bar{x} - x_i)^2} \tag{1.17}$$

Standard deviation $\sigma_{\bar{x}}$ has a meaning of a statistical error and gives probability of 66%, that the measured value x will be found in the region $\bar{x} - \sigma_{\bar{x}} \leq x \leq \bar{x} + \sigma_{\bar{x}}$. For confidence interval of α % standard deviation $\hat{\sigma}$ is written in the following way:

$$\hat{\sigma} = t_{\alpha,n} \times \sigma_{\bar{x}} \tag{1.18}$$

where $t_{\alpha,n}$ stands for student's coefficient. In physics and industry, confidence interval is usually taken as $\alpha = 95\%$. It's value depend on the number of experimental points and can be found from the table 1.1.

Table 1.1: Values of student's t coefficient for confidence interval α=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:

$$x = \hat{x} \pm \sqrt{\hat{\sigma}_{stat}^2 + \sigma_{syst}^2} \tag{1.19}$$

1.4.1 Propagation of Errors

This section covers a method to estimate an σ_{tot} of a function $f = f(x, y, z)$, when standard deviations σ_x, σ_y and σ_z are known. Here it is considered that variables x, y, z are independent. It can be shown, that σ_{tot} is expressed in the following way:

$$\sigma_{tot} = \sqrt{\left(\frac{\partial f}{\partial x}\sigma_x\right)^2 + \left(\frac{\partial f}{\partial y}\sigma_y\right)^2 + \left(\frac{\partial f}{\partial z}\sigma_z\right)^2} \tag{1.20}$$

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

1.5 Fitting methods

1.5.1 Likelihood function method

1.5.2 χ^2 method

χ^2 analysis [2, 4, 5] is based on minimising the difference between experimental data and the fitting function:

$$\chi^2 = \sum_{j=1}^{n} \left(\frac{y_j - F(E_j, \vec{\theta})}{\sigma_j} \right)^2 \tag{1.21}$$

where summation is taken over all experimental points. In formula 1.21 y_j stands for experimental data points, σ_j - standard deviation (error) of experimental data, $F(x_j, \vec{\theta})$ - fitting function with a set of parameters $\vec{\theta}$. Parameters θ are found from the minimisation condition:

$$\frac{\partial \chi^2}{\partial \theta_j} = 0 \tag{1.22}$$

1.5.3 Analysis of data in case of large statics

The technique, discussed at the beginning of section 1.4, can be used always. But the case of large statistics of experimental data gives access to analyse the data via fitting. On Fig. 1.5 one can see the examples of the distribution of a radius for a particle and square displacement of a Brownian particle. Large statistics of data allows to build the distribution of a value x. Mean value $\bar{x}$ and standard deviation $\sigma_{\bar{x}}$ of mean value $\bar{x}$ are estimated from the fit of the distribution. According to Central limit theorem a set of independent random variables tends toward a normal (Gauss) distribution. On the Fig. 1.5 one can find distribution of the value of radius r of Brownian particle and square displacement $\langle x^2 \rangle$. Final result with confidence interval of $\alpha = 66\%$ is written in the

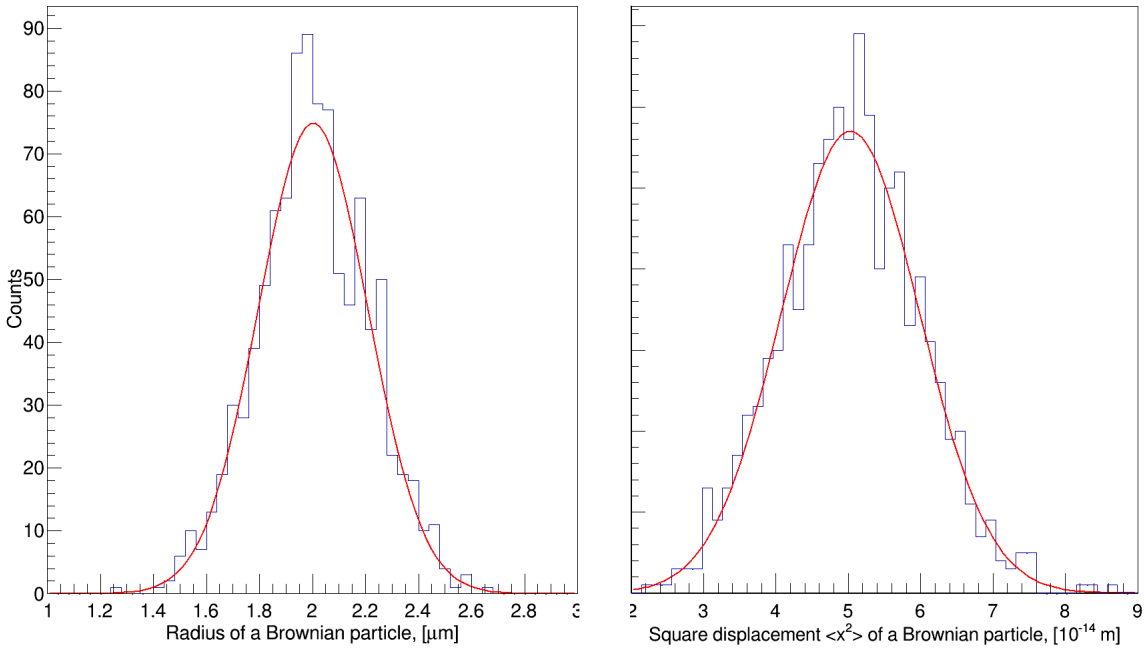

Figure 1.5: Example of the distribution of a radius for a particle and square displacement of a Brownian particle. Data is generated using Gauss distribution in both cases.

following way:

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

Final result with confidence interval of $\alpha = 95\%$ is written in the following way:

$$x = \bar{x} \pm \sqrt{2\sigma_{stat} + \sigma_{syst}} \tag{1.24}$$

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 (1.16) and standard deviation (1.17) for each case. Fit each data set with a distribution function, extract mean value and standard deviation from the fitting. Calculate χ^2 (1.21) for each case and define "goodness" of the fit. Present your results with 95% confidence level, taking into account statistical and systematical errors.
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 donts of reduced chi-squared. arXiv:1012.3754v1.