SHORT REVIEW OF DATA ANALYSIS AND STATISTICS

LEAST SQUARE METHOD

The least-square method is used to fit measurements with linear function of regression. It is also possible to fit other types of non-linear regression functions by linearizing them.
The least-square method can be used to fit linear function:

$$y(x) = a + bx \quad (1)$$

If possible, every set of measurements in which one variable depends on the other one, should be analysed by the least-square method in order to obtain the dependency.

Procedure of determination of the regression function and coefficients by the least-square method:

1. Obtain a series of measurement of two dependant quantities x_i and y_i in order to determine pairs (x_i, y_i). The number of measurements and obtained pairs should be as high and diverse as possible in order to have a sufficiently large number of fitting points. This will provide lower uncertainty and better fitting. There should be around 10 points optimally.

2. Choose the function to fit the measurements. The choice should be made according to the theoretical expectations and our hypothesis, and according to the theoretical background. If we would like to obtain general fitting and determine the dependency, it is advisable to fit the general function:

   $$y(x) = ax^b \quad (2)$$

   where *x* and *y* are measured quantities, *b* is power coefficient that determines the dependency, *a* is coefficient. If we obtain *b* = 1, the dependency is linear, for *b* = 2 dependency is quadratic (square law), etc. This procedure can be used when we do not expect y-segment.

3. Chosen function should be adapted for use with the least-square method by linearization. Non-linear functions of power type can be linearized by the use of logarithm. Exponential function such as (2) is transformed into linear function if logarithm is applied:

   $$\log y = b \log x + \log a \quad (3)$$

   We can choose logarithms *log x* and *log y* of measured quantities *x* and *y* as new variables:

   $$y' = \log y$$

   $$x' = \log x$$

   Relation (3) with these new variables *x'* and *y'* becomes linear function and the least-square method can be applied:

   $$y' = a' + b' x'$$

   We can determine new coefficients *a'* and *b'* by the least-square fitting. They are related to the original *a* and *b* coefficients from (2) as:

$$a' = \log a$$

$$b' = b$$

The above procedure shows linearization of non-linear function, adapting it for the least-square fitting.

More complex power function

$$y = ae^{bx}$$

can be also linearized by taking its logarithm:

$$\ln y = \ln a + bx$$

By the use of the least-square method, coefficients *ln a* and *b* can be determined.

4. Find new values of variables if the function is non-linear (e.g. *log x* and *log y* in (2)).

5. Determine regression coefficients and their uncertainties by the least-square fitting. If linearized non-linear function is fitted, determine the true values of coefficients you are looking for (e.g. coefficient *a* from the coefficient *log a* determined by the least-square fitting in (2)). Use the same procedure to determine uncertainty of the true value of coefficients you are looking for from the uncertainty of the coefficient determined by the least-square fitting (in equation (2), coefficient *a* is $a = 10^{a'}$, *a'* is determined by the least-square fitting, so the standard deviation of *a* is $\sigma_a = \frac{\partial a}{\partial a'}\sigma_{a'} = 10^{a'}\sigma_{a'}\ln 10$

6. Results of fitting should be presented as $a = (a \pm \sigma_a)$, $b = (b \pm \sigma_b)$, where the values in parenthesis are numerical.

7. Show linear regression together with the measurements on the same diagram. If the regression is non-linear, show linearized function of regression obtained by the least-square fitting together with adjusted measurements (in (2), show on diagram measurements *log x* and *log y* as well as fitted regression function with coefficients *a'* and *b'*). If possible, show on a separate diagram true values of measurements together with non-linear regression function (in (2), show on diagram values *x* and *y* and regression function (2) of power-type with the determined values of coefficients *a* and *b*).

Determination of coefficients in linear regression (relation 1):

$$a = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

where the summation is performed over every measured pair (x_i, y_i).

Uncertainty in coefficients *a* and *b* (standard deviations):

$$\sigma_a = \sqrt{\frac{\sigma_y^2\left(\sum x_i^2\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2}}$$

$$\sigma_b = \sqrt{\frac{n\sigma_y^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}}$$

where standard deviation σ_y can be calculated according to:

$$\sigma_y = \sqrt{\frac{\sum\left(y_i - a - bx_i\right)^2}{n-2}}$$

DIRECT MEASUREMENTS

Measurements in which the value *x* is directly obtained in *n* repeated measurements are called **direct measurements**. In that way, we obtain *n* different measurements of values x_i. **Mean value** of measurements *x* is then determined as:

$$\bar{x} = \frac{\sum x_i}{n}$$

Standard deviation (measurements uncertainty) of the mean is:

$$\sigma_x = \sqrt{\frac{\sum\left(\bar{x} - x_i\right)^2}{n(n-1)}}$$

Result of the direct measurements should be presented as:

$$x = \left(\bar{x} \pm \sigma_x\right)$$

and it can be interpreted as the following: the real value of the measured physical quantity can be found in the interval $\left[\bar{x} - 3\sigma_x, \bar{x} + 3\sigma_x\right]$ called **confidence interval** around the mean value with a probability of 99%.

Relative uncertainty is determined as:

$$R_{\sigma_x} = \frac{\sigma_x}{\bar{x}} \cdot 100\%$$

Maximal error Δx is the maximal possible value of the error we can make in a single measurement. If only random errors are present, maximal error is considered to be:

$$\Delta x = 3\sigma_x$$

If, beside the random error, the systematic error is also present in the measurements, and this systematic error is estimated to be *s*, than maximal error of direct measurements is:

$$\Delta x = 3\sigma_x + s$$

and in that case the total measurement uncertainty of *x* corresponds to maximal error Δx.

The result should be presented as:

$$x = \left(\bar{x} \pm \Delta x \right)$$

If the standard deviation of the mean cannot be determined from the measurements themselves, the uncertainty can be **estimated**. If **estimated uncertainty** is Δx, then the result of a series of measurements can be presented as

$$x = \left(\bar{x} \pm \Delta x \right)$$

If we cannot obtain multiple measurements and only one measurement x_1 with estimated uncertainty of Δx_1 is found, then the result should be present as:

$$x = \left(x_1 \pm \Delta x_1 \right)$$

Relative uncertainty with the estimated error:

$$R_{\Delta x} = \frac{\Delta x}{\bar{x}} \cdot 100\%$$

Relative uncertainty if only one measurement with estimated error is obtained:

$$R_{\Delta x_1} = \frac{\Delta x_1}{x_1} \cdot 100\%$$

INDIRECT MEASUREMENTS

We would like to determine the value F of some physical quantity which is a function of directly measured quantities $x_1, x_2, x_3, \ldots, x_k$, described by F:

$$F = F\left(x_1, x_2, \ldots, x_k \right)$$

Indirectly determined quantity F can be obtained from the value of function F by using the mean values of directly measured quantities $x_1, x_2, \ldots, x_k$:

$$\overline{F} = F\left(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_k} \right)$$

Standard deviations (uncertainty, mean error) of indirectly determined quantity $\overline{F}$ is:

$$\sigma_F = \sqrt{\sum_k \left(\frac{\partial F}{\partial x_k} \sigma_k \right)^2}$$

while the relative uncertainty of indirectly determined quantity F is:

$$R_{\sigma_F} = \frac{\sigma_F}{\overline{F}} \cdot 100\%$$

The result of indirectly determined quantity should be presented as:

$$F = \left(\overline{F} \pm \sigma_F \right)$$

MEASUREMENTS IN MULTIPLE SERIES

The physical quantities can be determined from multiple series by multiple measurements in each of the series. If this is the case, multiple series of independent measurements are obtained for which mean values and uncertainties can be determined:

$$x_1 = \left(\overline{x_1} \pm \sigma_{x_1}\right), \quad x_2 = \left(\overline{x_2} \pm \sigma_{x_2}\right), \quad ....., \quad x_i = \left(\overline{x_i} \pm \sigma_{x_i}\right)$$

Different series can have different uncertainties.

1. Consistent measurements

Measurements are **consistent** if for all the values of x_i the differences $\left|\overline{x} - \overline{x_i}\right|$ are less or of the order of uncertainty σ_{x_i}, where σ_{x_i} can be any of the values $\sigma_{x_1}, \sigma_{x_2}, ...$, and $\overline{x}$ is the mean value of the mean values of measurement series $\overline{x_1}, \overline{x_2}, ...$

If the difference between the mean value of all the series and mean value of each series is of the order of uncertainty for that series, the **general mean value** and **uncertainty** of all the series are:

General mean: $\overline{x} = \dfrac{1}{\sigma_{x_1}^{-2} + \sigma_{x_2}^{-2} + ... + \sigma_{x_i}^{-2}} \left(\dfrac{\overline{x_1}}{\sigma_{x_1}^2} + \dfrac{\overline{x_2}}{\sigma_{x_2}^2} + ... + \dfrac{\overline{x_i}}{\sigma_{x_i}^2} \right)$

General uncertainty: $\sigma_x = \dfrac{1}{\sigma_{x_1}^{-2} + \sigma_{x_2}^{-2} + ... + \sigma_{x_i}^{-2}}$

If all the means of individual series are similar and e.g. $\sigma_{x_1} \approx \sigma_{x_2} \approx \sigma_{x_4} \approx ... \approx \sigma_{x_i} >> \sigma_{x_3}$, then:

$$x_3 = \left(\overline{x_3} \pm \sigma_{x_3}\right)$$

2. Non-consistent measurements

Measurements are **non-consistent** if, for all the values of x_i, the differences $\left|\overline{x} - \overline{x_i}\right|$ are much larger than any of the uncertainties σ_{x_i} of individual series, where $x_1 = \left(\overline{x_1} \pm \sigma_{x_1}\right)$, $x_2 = \left(\overline{x_2} \pm \sigma_{x_2}\right)$, ..., $x_i = \left(\overline{x_i} \pm \sigma_{x_i}\right)$.

In this case, disregard all the uncertainties of each individual series σ_{x_i}. General mean is determined as a mean of individual series with disregarded uncertainties. All means of individual series should be considered as individual measurements, and general uncertainty determined according to this.