Error calculation for beginners

an example-oriented introduction for students of the TUHH

1. Measurements and inaccuracy

Many physical quantities such as length, temperature, mass etc. can be determined by measurement. Repeated readings of such measurements tend to spread around the true value; which is called distribution. Imagine you measure a mass precisely ten times in succession and you get ten different results - although you are sure the mass never changed. The reason is that each measuring process is inherently fraught with inaccuracy, which is usually referred to as an *error*. In order to properly interpret your precise measurement, it is necessary to specify error for example for a value of 78.34655 kg.
In scientific experiments, the specification of any inaccuracy is particularly important in order to set boundaries on the estimated values in order to determine the probability of finding the true value within this inaccuracy interval, which is "narrow". Without such a boundary, a reading of the measured value such as 78.34655 kg is only informative and meaningless. So what do we do? We need to...

- ... design the experiment in such a way that the error is as small as possible. There are not a lot within the lab that can be manipulated, because the most experiments and their procedures are already fairly concrete.
- ...determine the uncertainty of the results.
- ... find out how large the probability of finding the result within the error interval is by repetition of the experiment.

The typical procedure, like in our scale example, is to take repeated measurements in order to have a series of measurement values. Several measurement series may be combined to make a larger, single series but in doing so we must ensure that all measurements were conducted under the same conditions. If any of the test conditions were altered, such as the use of different measurement tools, the distribution of the measured values are also expected to change. Series of measurements will be considered to have been evenly conducted until the end of Chapter 5.

How do we specify errors? An error is the inaccuracy of a physical quantity which more often denoted as Δx. This, however, is not to be confused with a change in x. A particular type of error is referred to as standard deviation σ_x, which we will discuss later. The subscript indicates the physical quantity being referred to.

Error values can be specified by

- absolute (Δx),
- relative ($\Delta x/x$)
- percentage ($\Delta x/x \cdot 100\%$)

absolute: $m=(78{,}35 \pm 0{,}61)$ kg
relative: $m=78{,}35 \text{ kg} \pm 7{,}8 \cdot 10^{-3}$
percentage: $m=78{,}35$ kg $\pm 0{,}78$ %
These are simple examples in order to help make the notation clear.

In the example box on the right, it can be seen that not all decimal places of our previous weight example have been taken into account. If the error is Δm=0.61 kg, it makes no sense to record more decimal places for m as they are smaller than the inaccuracy. So it is important to note that going to more decimal places does not make the value more accurate! This holds also for the small scales as well as particle-detectors, which are as big as a house block.

Error calculation provides the tool to determine the error magnitude of the measurements of a physical quantity. Although a calculator may spit out many decimals after calculation, how many should we take into account? How accurate should an error be? As a general rule of thumb, in our example, it should be taken to two significant digits (significant decimal numbers) and rounded off at the end.
Here is a summary of all that has been already mentioned in this example:

Value read from the measurement tool (scale):	m=78,34655 kg
Calculated error:	Δm=0,612549 kg
Measured value with (absolute) error:	m=(78,35±0,61) kg

Here, Δm was not calculated but simply considered as a proper value for illustration purposes.

2. An error is an error, isn't it?

Simple answer: Nope! Because the inaccuracies that occur during measurement can have different fundamental causes. Any error of a measured value or a calculated magnitude belongs to one of these two categories:

2.1 Statistical Error

- These kinds of error affect the result of the measurement and are unpredictable and uncontrollable; hence, we call them random errors. What causes such errors?

- Deficiency in the human sense organs (such as the limited resolving power of the eye, when it comes to the question of whether two fine lines are superimposed or easily side by side),
- Clumsiness when measuring and reading (parallax error),
- Statistically acting external influences (i.e. vibrations).

Statistical errors can be treated mathematically using the tools of statistics. Note that these errors have both signs ("±" in the box above). The values from Repeated measurements, like weighing a mass ten times, are all distributed around a mean value. This average is not actually the value that we want. However, the larger a series of measurements is, the closer the average approaches the actual value. For researchers and trainees, this means that test series should be as large as possible. You wouldn't trust a medicine that has only been tested on five people.
It can be stressed enough that a single measurement in principle tells us nothing. Only once you have determined the accuracy of the method of measurement by many repetitions, you can understand how the result of each individual measurement fluctuates around an

approximation of the true value. This approximation, which is calculated from the individual measurements, must then be useful and suitable.

2.2 Systematic Error

Systematic errors affect all the individual measurements in the same way. They are reproducible; i.e. they occur in the same size and with the same sign by repeating the measurements under the same conditions. This is quite convenient, as the results of measurement can and must be corrected accordingly. Note that the statistical errors are not reproducible!

So what leads to a systematic error in a measurement? Issues arise most commonly through faulty measuring instruments, such as a scale not displaying zero despite having nothing on it. Other causes may include the environmental influences that are often neglected in the internship, as temperature and pressure that change the values of a series of measurements, or even weak electric and magnetic fields, from wall outlets or nearby power lines.
For the purposes of error calculation, only statistical errors are taken into account and not systematic errors.

3. Frequency Distribution of Results (Measured Values)

After sweating and cursing in getting proper measurement and calculating the error, my results is 32 ± 2 cm. So far so good, but unfortunately the reference value in the text book is 35 cm. I've messed up somewhere, this whole thing is stupid! This is the reaction of many students when they compare their results with value out of a book, so we will clarify just what '± 2 cm' actually means. Does that mean none of the values measured were greater than 34 cm and none less than 30 cm? Simple answer is no. What is going in an experiment?

A physical quantity is measured, for simplicity, is called x. It could be a temperature, time, length, or anything else. The measurement is performed exactly N times under precisely the same conditions. The result of the ith measurement is x_i. It may happen that the same reading occurs more than once. Then, the number N_i indicates how many of N measurements have the same result x_i. The next step is to divide N_i by the total number of measurements N to get the relative frequency $n(x_i)$ of the value x_i. As a formula:

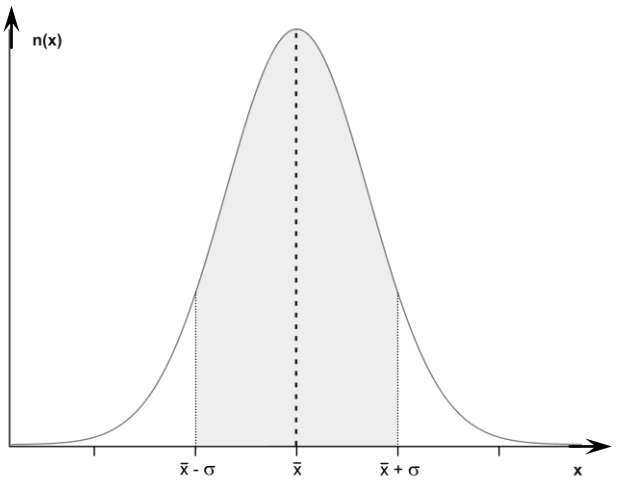
$$n(x_i) = \frac{N_i}{N} = \frac{\text{measurements with the same magnitude } x_i}{\text{Total number of measurements}} \tag{1}$$

Let us remember that the measured values are not random numbers, but are somehow distributed around a mean value. We plot the relative frequency of a measured value xi versus the possible values x. We see that the resulting curve becomes ever more prominent with a higher N value and can be described using Gaussian (or normal) approximation, so the values have Gaussian (or normal) distribution. The Gaussian curve is represented by:

$$n(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \tag{2}$$

Important: The Gaussian distribution is obtained only for N → ∞. The distribution of your measurements will no doubt seem quite different if too few measurements were taken.

The following figure shows a graphic representation of the Gaussian distribution:



Which properties does this distribution have?

- n(x) is the distribution of a *probability density* while n(x)dx determines the probability of the measured value being located in the interval [x, x + dx] .
- $\bar{x}$ is the most occurring value. For $N \to \infty$ it is equal to the true value of x.
- It can be seen in the figure that the curve has two inflection points that are located at $\bar{x} - \sigma$ and $\bar{x} + \sigma$. So, σ indicates the distance between the inflection points and $\bar{x}$.

Since n(x) is a probability density, the area under the curve is a probability. Accordingly, we have

$$\int_{-\infty}^{\infty} n(x)dx = 1, \tag{3}$$

We consider the probability of a measurement within -∞ to +∞ boundary; which of course must be 1 (normalization condition).
The two defined variables $\bar{x}$ and σ in the context of Gaussian distribution, as you may have guessed, have practical significance for the evaluation of your measurements:

- $\bar{x}$ is the mean value of the distribution. It is the most common measurement result and of particular interest, because it is a good approximation of the true value of a physical quantity out of a large series of measurements ($N \to \infty$).
- σ is the standard deviation of a distribution. It is a measure of the scattering of the individual measurement results around the mean value. A large σ value means that the Gaussian distribution is quite spread and a small value describes a narrow distribution. In the first case, the results of measurement deviate too far from x̄ whereas in the latter, results lie close in vicinity to the mean value. Thus, the standard deviation is appropriate in specifying the precision of a measurement method.

When various physical quantities are measured in an experiment, we write the standard deviation as σ_x in correlation to the value x. But what does σ_x show quantitatively? It's quite simple:
$[\bar{x} - \sigma, \bar{x} + \sigma]$ is an interval around $\bar{x}$, in which 68.3% of the all measured values lie, and describes the probability of finding a measurement within this interval is

$$\int_{\bar{x}-\sigma}^{\bar{x}+\sigma} n(x)dx = 0{,}683\,. \qquad (4)$$

Accordingly, we have:

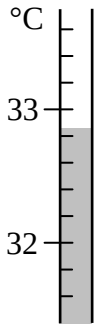
95,4% of the measurement results lie within $[\,\bar{x}-2\sigma,\,\bar{x}+2\sigma\,]$ and

99,7% of the measurement results lie within $[\,\bar{x}-3\sigma,\,\bar{x}+3\sigma\,]$.

This clearly shows that we cannot say that all measurements locate within the 3, 4 or 5-times of the standard deviation. At any large or small interval in which $\bar{x}$ lies, there is a probability greater than zero that the next measured value lies outside of this interval.
If we return to the beginning of this chapter; it was said that a value out of a textbook is never the true value, but a mean value, which was obtained just like in our experiments (unless it is theoretically derived, in which case it should be expressly denoted as a theoretical value). Incidentally, textbook references are only useable when the errors are also given. So if you get (32 ± 2) cm and the literature says 35 cm (often error values are not given in), you may argue that the values are within the two times of the standard deviation.

The next step is to clarify how to calculate mean values and standard deviations from a series of measurements. One thing should first be made clear that often causes confusion when taking your first error calculations. Error calculation is about determining inaccuracies from values that are either obtained directly by reading them off of a measurement instrument or calculated by means of a formula. The *input* we have comes from the measured values themselves, as well as known and already existing inaccuracies in the tools of measurement. The latter however, is not the goal of the error calculation. Often the accuracy of a measurement device is specified by the manufacturer, if not, their accuracy be estimated.

Here, there is the sketch of a mercury thermometer calibration:



What does the thermometer show? We read 32.8 ° C and a little bit… No matter what we read, it is definitely not precise because we cannot accurately read off the scale. In this example, it can only be said that the mercury lies either exactly on a line or somewhere in-between, in this case the latter. So we estimate the temperature to be T = 32.9 ° C with an inaccuracy in reading of ΔT = 0.1 ° C, which is the distance between a ‘line’ and the point exactly between two lines. It doesn’t get any more accurate than that. Note that the error was estimated and not calculated!

4. Mean Value and Standard Deviation

As explained above, we have already seen the main outcome of error calculation, the true value of a measured quantity x, calculated as the arithmetic average in a limiting case of an

infinitely large N number of measurements, i.e. an infinitely large series of measurements. Mathematically, we express this definition as:

$$x_{wahr} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i \tag{5}$$

where x_i is the result of the *i*th measurement of value x.
Obviously it is impractical to take an infinitely large series of measurements, but we are able to make a reasonable amount of measurements N within a practical timeframe. In statistics this is referred to as taking a sample set.

What do we want to get out of this sample set? We want...

1. ... have the best possible estimation of $\bar{x}$ for the true value,
2. ...calculate the standard deviation σ_x, which describes the distribution of the individual measurements around the average value (see previous section) and
3. ... the inaccuracy of the mean value $\sigma_{\bar{x}}$, which tells us how all possible means are distributed around the true value.

Let's start with the first point; the best estimate of the true value is obtained from the arithmetic mean value of all measurement results:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \ldots + x_N}{N} \tag{6}$$

i	*T* (°C)
1	38,6
2	38,8
3	38,9
4	38,9
5	39,1
6	38,8
7	38,7
8	38,4

On the left is a series of measurements of my body temperature *T* measured with a thermometer (I was ill and I was bored ☺).
The true value of T cannot be calculated because of the finite number of measurements N in the sample set, but as a mean value, we have:

$$\bar{T} = \frac{1}{8} \cdot (38{,}6\ °\text{C} + \ldots + 38{,}4\ °\text{C}) = 38{,}775\ °\text{C}.$$

The second point tells us of the standard deviation σ_x, which describes how the individual measurements spread around the mean value $\bar{x}$. We calculate that from the average of the square of the distance between a measurement and the mean value $\bar{x}$. The sum of these squares is divided by N-1 instead of N, because one of the individual measurement deviations from the Mean value is zero (because the values are distributed over and below the mean). The mathematical formula looks somehow cumbersome:

$$\sigma_x = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}$$

$$= \sqrt{\frac{1}{N-1} \cdot \left[\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots + \left(x_N - \bar{x}\right)^2 \right]}. \qquad (7)$$

This is the standard deviation *of an individual measurement* from its *mean-value*. It is also called the **standard** error of the sampling distribution.

As an example, let's focus on the temperature-measurement series from the box:

We had determined $\bar{T}$ = 38,775°C calculated from N = 8 measurements.

T (°C)	$\left(T - \bar{T}\right)^2$ (°C^2)
38,6	0,030625
38,8	0,000625
38,9	0,015625
38,9	0,015625
39,1	0,105625
38,8	0,000625
38,7	0,005625
38,4	0,140625

The sum of the right-hand column is 0,315 ° C^2 (do not be confused by "degrees Celsius squared"!). This value is then divided by N-1 = 7, which gives 0.045 ° C2. We then take the average of this square deviation of individual measurements from the mean $\bar{T}$ (i.e., the standard deviation of an individual measurement).

$\sigma_T = 0,21$°C (the accuracy is with two significant places)

The mean value calculated by using (6) is an estimation of the true value of the size of x. Here, we take several series of measurements with differing averages. It shows the average itself has an error. All averages that are theoretically possible are distributed around the true value x_{wahr}. So, we calculate the standard deviation of the *mean value* from the *true value*, which leads us to the third point. The true value itself remains as always hidden, but amazingly we are still able to calculate this standard deviation as follows:

$$\begin{aligned} \sigma_{\bar{x}} &= \sqrt{\frac{1}{N \cdot (N-1)} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2} \\ &= \sqrt{\frac{1}{N \cdot (N-1)} \cdot \left[\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots + \left(x_N - \bar{x}\right)^2 \right]} \qquad (8) \\ &= \frac{1}{\sqrt{N}} \cdot \sigma_x \end{aligned}$$

Pay attention to the small difference between the standard deviation of a mean-value and that of an individual measurement! In fact, the root mean square deviation of the mean-value from the true value decreases with increasing the sample size N by a factor of $\frac{1}{\sqrt{N}}$. Here, we see mathematically why a series of measurements should be as large as possible.

If we take $\sigma_T = 0.21$ °C from above and divide it by √8, we get the *standard deviation* of the *mean-value*:

$$\sigma_{\bar{T}} = 0{,}074\ °C$$

What if an average value, which is indeed an estimate of the true value, is determined, but we are interested more in mean square deviation from the true value? So, it is useful to write down the whole calculation steps while you are calculating standard deviation. (take care that measurement value and error calculation should be written with the same amount of decimal places):

$$\bar{T} = (38{,}775 \pm 0{,}074)\ °C.$$

That temperature is high for my body, so I might have to leave you now and seek medical diagnose.

In statistics, beside error calculation we can calculate another concept like the variance Var(x) of a measured variable x. This is represented by σ^2 and called the square of the standard deviation.

5. Propagation of Error

In the lab, you may often need to measure a number of different physical quantities, in order to calculate the value of interest. This is because the required quantity is not accessible by a direct measurement. Here, the measured quantities are called x_1, x_2, ..., x_m respectively. Pay attention to the difference in the indices compared to the previous chapters. Before, the index showed which measurement was being represented from a series of measurements of the same physical quantity, but here each index represents a new quantity, for example, x_1 could be a temperature, x_2 an electrical voltage, etc. As z can be calculated from the different quantities, it is written as a function of $x_1, \ldots, x_m$. In other words $z = f(x_1, \ldots, x_m)$. Now we want to know how to calculate the error of z, when we already know the errors of $x_1, \ldots, x_m$. These errors are usually the inaccuracy of the measuring devices, often stated in the manufacturer's information attached to the device.

5.1 The Error Propagation Law

In order to properly apply an error calculation, two conditions must be fulfilled:

1) The measurement results of the measured values $x_1, ..., x_m$ must be normally distributed in order to have Gaussian distribution.
2) The individual values $x_1, ..., x_m$ must be statistically independent.

The second condition means that one measured vale x_i must not have any analytical dependency on the other measured value x_j, otherwise x_i is a function of x_j or $x_i = f(x_j)$ in which case there would be too many measurable variables for the calculation. Logically, we want to apply the smallest possible amount of measurements. The following example illustrates this definition:

A student wishes to calculate the area *A* of a circular disc and so measures the radius *r*, as well as the diameter *d* just to be certain. From a series of measurements they then calculate the mean values $\bar{r}$ and $\bar{d}$. As s/he wants to evaluate both variables, the area A gets divided into two halves:

$$A = A_1 + A_2 = \frac{1}{2}\pi\bar{r}^2 + \frac{1}{2}\pi\left(\frac{\bar{d}}{2}\right)^2$$

This formula is undoubtedly correct, there is no reason the area can't be calculated this way. However, the two variables r and d are not independent, since the are related as $d = 2.r$. In this case, the second condition is not satisfied and the error propagation law is not applicable. If however, the equation $A = 2\pi r^2$ were used, both conditions are fulfilled and we work with the smallest possible number of variables.

We then assume that the mean value of each series of measurements regarding to the different variables $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m$ are known as well as standard deviations of the mean values $s_{\bar{X}_1}, \ldots, s_{\bar{X}_m}$. Then, the best estimate of the mean-value of z is

$$\bar{z} = f\left(\bar{X}_1, \ldots, \bar{X}_m\right), \tag{9}$$

The optimum estimation for the error of $\bar{z}$ is the standard deviation $\sigma_{\bar{z}}$. It is calculated according to the error propagation law, which plays a very important role in the practical internship.

$$\sigma_{\bar{z}} = \sqrt{\sum_{j=1}^{m}\left(\frac{\partial \bar{f}}{\partial x_j}\right)^2 \cdot \sigma_{\bar{x}_j}^2}$$

$$= \sqrt{\left(\frac{\partial \bar{f}}{\partial x_1}\right)^2 \cdot \sigma_{\bar{x}_1}^2 + \ldots + \left(\frac{\partial \bar{f}}{\partial x_m}\right)^2 \cdot \sigma_{\bar{x}_m}^2}\,. \tag{10}$$

Here, $\partial/\partial x_j$ is the partial derivative with respect to x_j (formed as a normal derivative), and the horizontal bar variables indicates that mean values of them must be considered in the bracket. You learn how to use this formula in the following example.

We calculate the density ρ of a cube, here. The side length *a* and mass *m* are measured. The formula for density is:

$$\rho = \frac{m}{a^3}.$$

In order to calculate the average density, the average values m and a are used:

$$\bar{\rho} = \frac{\bar{m}}{\bar{a}^3}$$

The errors of the averages $\sigma_{\bar{m}}$ and $\sigma_{\bar{a}}$ are calculated in accordance with Equ. (8). Let's assume that they are already known. The error of the density is calculated according to the error propagation law, Equ. (10):

$$\begin{aligned}\sigma_\rho &= \sqrt{\left(\frac{\partial\rho}{\partial m}\right)^2 \cdot \sigma_m^2 + \left(\frac{\partial\rho}{\partial a}\right)^2 \cdot \sigma_a^2} \\ &= \sqrt{\left(\frac{1}{a^3}\right)^2 \cdot \sigma_m^2 + \left(\frac{-3m}{a^4}\right)^2 \cdot \sigma_a^2} \\ &= \sqrt{\frac{1}{a^6} \cdot \sigma_m^2 + \frac{9m^2}{a^8} \cdot \sigma_a^2}\end{aligned}$$

To get the error of *mean-density* we simply substitute the mean values of the variables as well as the *standard deviations* of each *mean-value*:

$$\sigma_{\bar{\rho}} = \sqrt{\frac{1}{\bar{a}^6} \cdot \sigma_{\bar{m}}^2 + \frac{9\bar{m}^2}{\bar{a}^8} \cdot \sigma_{\bar{a}}^2}\ .$$

5.2 Maximum Error

It can sometimes happen that one of the conditions (or both) of the error propagation law is not fulfilled. If this is the case, Equ. (10) cannot be used to specify the error of the physical quantity z. This does not mean that an error cannot be determined, but that the interpretation of the maximum error Δz is instead applicable. Maximum error mathematically calculates as follows:

$$\begin{aligned}\Delta z &= \sum_{j=1}^{m}\left|\frac{\partial f}{\partial x_j}\cdot\Delta x_j\right| \\ &= \left|\frac{\partial f}{\partial x_1}\cdot\Delta x_1\right| + \left|\frac{\partial f}{\partial x_2}\cdot\Delta x_2\right| + \ldots + \left|\frac{\partial f}{\partial x_m}\cdot\Delta x_m\right|\end{aligned} \tag{11}$$

The absolute value lines are due to the partial derivative of f, which can be negative; but we need to sum positive contributions of each various quantities.
For any value x_j, whose mean-values and standard deviations are known out of a series of measurements, the error is

$$\Delta x_j = \sigma_{\bar{x}_j},$$

where the standard deviation of the mean-value is also plugged in. For all other variables, a realistic estimation of the error must be taken (an example of this can be found on page 5). As a simple example, we consider the density calculation from the previous page.

The only difference in this calculation from the last is that the errors of the mass *m* and side length *a* are not calculated from a series of measurements (standard deviation), but instead are estimated values Δm and Δa. Even if the law of error propagation were used, the standard deviation of $\bar{z}$ would not b determined, but only an estimate of the error $\Delta\bar{z}$. Here, we are interested specifically in the maximum error of the density ρ, calculated according to Equ. (11):

$$\begin{aligned}\Delta\rho &= \left|\frac{\partial\rho}{\partial m}\cdot\Delta m\right| + \left|\frac{\partial\rho}{\partial a}\cdot\Delta a\right| \\ &= \left|\frac{1}{a^3}\cdot\Delta m\right| + \left|\frac{-3m}{a^4}\cdot\Delta a\right| \\ &= \frac{1}{a^4}\cdot\left(a\cdot\Delta m + 3m\cdot\Delta a\right)\end{aligned}$$

The last step is considering the absolut values in order to omit the minus sign. Since m and a and their errors Δm and Δa are greater than zero, we should sum the errors. The maximum error of the *density mean-value* is obtained by inserting the mean values of *m* and *a*:

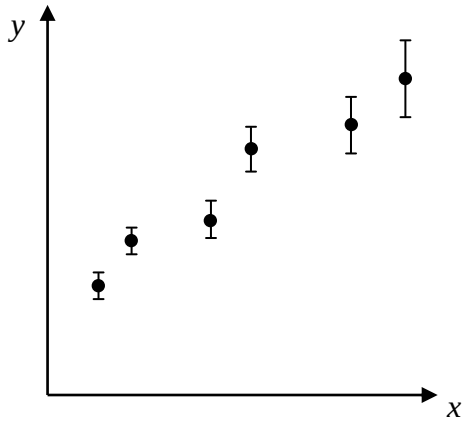
$$\Delta\bar{\rho} = \frac{1}{\bar{a}^4}\cdot\left(\bar{a}\cdot\Delta m + 3\bar{m}\cdot\Delta a\right)$$

6. Linear Regression

In many cases, the two physical quantities such as x and y are linearly dependent, and the relationship is shown with a linear equation

$$y = a \cdot x + b, \tag{12}$$

where a is the slope, b is the y-intercept, and both are constant.
Often during experiments we change the magnitude of a quantity x for N sequentially different values x_i and then we read the dependent quantity y_i from a measurement device; each y_i value has inherently a random error. An experimenter is interested in parameters a and b (usually materials or fundamental constants), and how to get those values is the discussion target of this section.
After N measurements of successive different x_i, we have a series of measurements. It is a common question for the interns to find out whether the measured values, which are in this series of measurements, show a linear relation or not. Remembering a reliable tip can save a lot of time and effort: Just plot it! Put the data points (x_i, y_i) in a graph to make it easier to recognize what kind of curve they make. This takes us to a crucial point; even if theoretically a linear relation for x and y can be derived, it still may be that the points (x_i, y_i) do not lie exactly on a line. This is of course due to the error of the y_i values. Of course all values x_i also contain errors, but they are assumed to be negligible. In principle, the error of x_i may be considered in linear regression but this increases computational complexity dramatically and we cannot calculate that anymore in this internship.



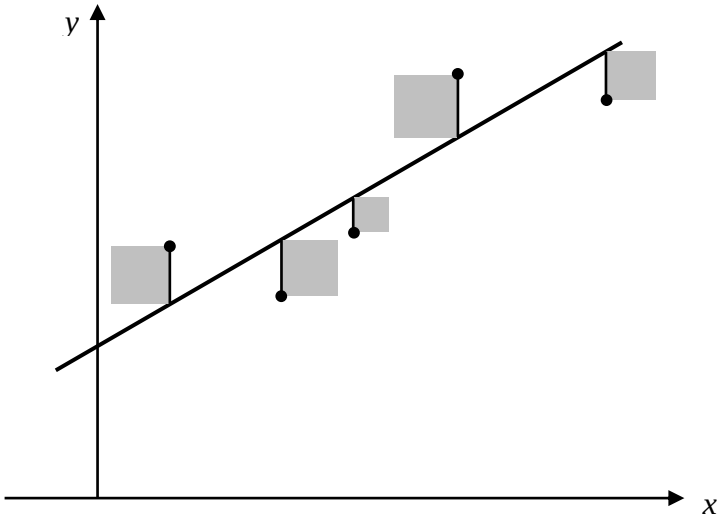
Typical position of the data points when x and y theoretically have a linear relation. Only the errors of the y-coordinates are taken into account (error bars).

The task is now to find the best fitting straight line to describe the data points, and then to get those parameters a and b.
The easiest way to approximate a best-fit line is simply to draw a straight line using a ruler, in a way that by visual inspection 'balances' the positioning of the data points. It is not so much the number of points above and below the line that is important, but more so that their distances to the line are as small as possible. This method of finding a best fit is likely familiar to many students as it is often taught in school. However, do keep in mind that the common practice of simply connecting the first and last data points to make a line is incorrect! It must be incorrect because in order to find a best-fit line, each pair of values must be taken into account (Unfortunately, it seems there are still teachers who encourage the use of a wrong method!!!).

The next step is easier, b is read off of the y-axis and the slope a is determined by means of drawing a triangle.
So far, so good. There is of course a mathematical, quantitative method to find the desired linear equation. The method is called linear regression, which is based on the fact that the sum of the squares of the distances between measured points and the straight line should be minimal (distances in the y-direction):

$$\sum_{i=1}^{N}\left[y_i-\left(a\cdot x_i+b\right)\right]^2 \overset{!}{=} \text{minimal} \tag{13}$$

Here, we see that we have an extreme-value problem. You may be wondering why we take the sum of the squares and not simply the sum of distances. The reason lies within the depths of mathematical statistics:. The Gauss-Markov law states that the best-fit model, i.e.
Linear equation is obtained when the method of the least-square is used.



Method of the least-square: The squares of the distances are represented with the gray areas. The linear regression is the straight line for which the sum of the gray areas is the smallest. The error bar of the data points has been disregarding for simplification.

After lengthy calculations which will not be explained in detail, the solution to the extreme-value problem Equ. (13) gives us the best estimations *a*' and *b*' of the parameters *a* and *b*:

$$a' = \frac{S_{xy}}{S_x^2}; \qquad b' = \bar{y} - a'\cdot\bar{x} \tag{14}$$

with the following definitions:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \; ; \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right) \; ; \qquad S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 \tag{15}$$

This puts us in a difficult situation in that we have to calculate averages for *x* and *y*, although the *x*-values vary deliberately during the series of tests, in other words the values of *x* and *y* are continuously changing.

It is always a good idea and highly recommended in the evaluation of results to write down all intermediate values (15) when conducting a linear regression - it makes keeping track of your workings and progress considerably easier.

Is that all? Not entirely! Although the estimates (14) for the parameters *a* and *b* are already sufficient to formulate the linear equation for the regression line, it will undoubtedly, as always, contain errors. If we for example - by means of the linear regression - would like to determine a physical constant, its error value will be of upmost interest.

From Equ. (11) and applying the error propagation (we again skip the details here, because you learn nothing new out of it.) we find the best estimate for the error of y.

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left[y_i - \left(a' \cdot x_i + b'\right)\right]^2} \; . \tag{16}$$

This is the standard deviation of the *individual measurement*. A quick look back to Equ. (7) in Chapter 4 highlights a slight difference in calculation, in that the standard deviation of an individual measurement divided by *N*-2. Make sure it is understood, that earlier we discussed the deviation of the individual measurements from the *mean*, while here we have the deviation of the individual measurements from the *linear regression*. In the first case we have only one parameter *x* which changes, in the second case we juggle two parameters *a* and *b*. This clarifies the slight difference. Also, note that Equ. (16) does **not** give the y-error, which is shown as error bars in the chart!
The best estimates of the error $\sigma_{a'}$ and $\sigma_{b'}$ for the estimated values a' and b' are calculated with:

$$\begin{aligned} \sigma_{a'} &= \sigma_y \cdot \sqrt{\frac{N}{N \cdot \sum x_i^2 - \left(\sum x_i\right)^2}} \\ &= \sigma_y \cdot \sqrt{\frac{1}{\sum x_i^2 - N \cdot \bar{x}^2}} \\ &= \sigma_y \cdot \sqrt{\frac{1}{\sum \left(x_i - \bar{x}\right)^2}} \end{aligned} \tag{17}$$

and for the errors of the slope and the y-intercept.

$$\begin{aligned} \sigma_{b'} &= \sigma_y \cdot \sqrt{\frac{\sum x_i^2}{N \cdot \sum x_i^2 - \left(\sum x_i\right)^2}} \\ &= \sigma_y \cdot \sqrt{\frac{1}{N} \cdot \frac{\sum x_i^2}{\sum x_i^2 - N \cdot \bar{x}^2}} \\ &= \sigma_y \cdot \sqrt{\frac{1}{N} \cdot \frac{\sum x_i^2}{\sum \left(x_i - \bar{x}\right)^2}} \end{aligned} \tag{18}$$

All sums are from i = 1 to N.

If we have a line crossing the origin with the form of y = a.x, then we apply this formula for calculating the slope

$$a = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sum_{i=1}^{n} x_i^2} \tag{19}$$

and the following one for the error of the slope:

$$\sigma_a = \sqrt{\frac{1}{n-1} \frac{\sum_{i=1}^{n} d_i^{\,2}}{\sum_{i=1}^{n} x_i^{\,2}}} \quad \text{, where} \quad d_i = y_i - a \cdot x_i \,. \tag{20}$$

There have been many large formulas introduced in this text, but do not panic, it is much easier than it looks. Since you will encounter the linear regression quite often in lab, we attached another example. Have you ever wondered how thick a piece of paper in your textbook is? Probably not (and why should you have?). This is a good example of something that cannot be particularly measured directly, but can be determined by linear regression.

----------Example----------

A book can be quantized like in quantum physics ☺: The thickness of the book is quantized, and a "book quanta" is the thickness of a sheet. the thickness is measured with a vernier caliper, not a single sheet, nor the whole book, but succesively 10, 20, 30, etc pages including the cover (the reason the cover is measured will become clear in a moment). The series of measurements are shown in the table, in which m is the number of the sheets:

Anzahl m	Dicke d (mm)
10	3,2
20	4,2
30	5,1
40	5,8
50	6,8
60	7,7
70	8,8
80	9,7
90	10,8
100	11,7

First, the intermediate results are calculated: With formula (6) we obtain the average values of m and d:

$\bar{m}$ =55 und $\bar{d}$ =7,38 mm.

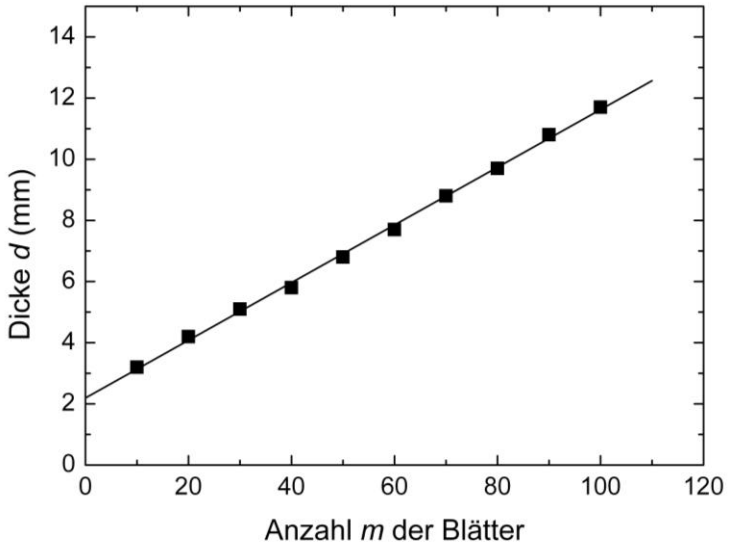
From Equ. (15) we get:

S_{md}=86,4444 mm and
S_m^2=916,6666.

The best estimates of the parameters a and b are then obtained from Equ. (14)

a'=0,0943 mm and
b'=2,1935 mm.

This is already enough to plot a graph with linear regression.



The image was created with the program ‘Origin’. Programs such as Excel or Origin can perform the entire linear regression themselves and are able to provide not only the diagram but also a', b' and their errors.

The linear equation of the plotted data is

$$d = 0{,}0943\ \text{mm} \cdot m + 2{,}1935\ \text{mm}.$$

We do not have the error information yet. For this we have a table of intermediate values:

m_i^2	$(m_i - \bar{m})^2$	$(d_i-(a'\cdot m_i+b'))^2$ (mm^2)
100	2025	0,00405
400	1225	0,01457
900	625	0,00604
1600	225	0,02732
2500	25	0,01172
3600	25	0,02289
4900	225	0,00003
6400	625	0,00139
8100	1225	0,01433
10000	2025	0,00588

Only the values in the right column have units as m is only a number.

First of all, we apply the formula (16)

$$\sigma_d = \sqrt{\frac{1}{8} \cdot \sum_{i=1}^{10} \left[d_i - \left(a' \cdot m_i + b' \right) \right]^2} = 0{,}173 \text{ mm.}$$

Using equ. (17) and (18) we get the errors

$\sigma_{a'}$=0,0019 mm und $\sigma_{b'}$=0,12 mm

In summary, we get also :

a'=(0,0943±0,0019) mm
b'=(2,19±0,12) mm

The slope a' specifies the average thickness of a sheet in the book. So we have determined the "book quanta", and also the error! The y-intercept b' is the thickness of the book back cover, i.e. the thickness of the book if it had no pages. Puting all the calculations together gives us the equation of the book:

Book thickness $D = a' \cdot m + 2b'$.

On the right side, there is a factor of 2, because we should also consider both the top and the bottom covers of the book.

The book equation in this example describes a fundamental law, and a' and b' correspond to the material parameters. The equation will always hold for a particular book, in this case it was Kohlrausch, Practical Physics, 22nd edition, BG Teubner, Stuttgart., 1968. ☺

----------End of Example ----------

So far we have discussed the general form of linear regression. There is an important special case in which we may use slightly different formulas to save some work; when the *y*-intercept is zero and the linear regression can be written as $y = a \cdot x$. But be careful! Consider thoroughly beforehand whether this really is the case! The simplified calculations may not be used if the *y*-intercept is just small but still present. It must be *exactly* zero, like the relationship between the length and time it takes for a beam of light to travel a certain distance. For a special case linear regression where the *y*-intercept is zero, the following simplified formulas may be used to calculate the best approximation *a*' of the slope and its error:
The standard deviation of the individual measurements of the linear regression is analogous to Equ. (16)

$$\sigma_y = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} \left[y_i - a' \cdot x_i \right]^2} \,. \qquad (21)$$

Here, the square is again divided by N-1, because we are dealing with only one parameter. The best estimate for the slope is:

$$a' = \frac{\sum x_i \cdot y_i}{\sum x_i^2}, \tag{22}$$

and as the best estimate of the error we have

$$\sigma_{a'} = \sigma_y \cdot \sqrt{\frac{1}{\sum x_i^2}}, \tag{23}$$

where the sums are always from i = 1 to N.

Sometimes computer programs calculate a coefficient, usually called *R*, in addition to the slope and y-intercept. It comes from Pearson's empirical correlation coefficient. It is defined as

$$R = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}} \tag{24}$$

with the definition of Equ. (15) and $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2$. R can have values between -1 and 1 and is used to check whether the measured data actually have a linear relation. For R=±1, all measuring points are exactly on a straight line. The sign indicates only whether the slope is positive or negative. When R = 0 there is no linear relation. The emphasis here is on linear, because the variables x and y can be even nonlinear but still correlates clearly visible in the diagram.

Bibliography:

Kamke,
Der Umgang mit experimentellen Daten, insbesondere Fehleranalyse, im physikalischen Anfänger-Praktikum: Eine elementare Einführung.
Verlag und Herausgeber: Wolfgang Kamke
ISBN: 978-3-00-031620-3
→ An excellent introduction, a detailed text without being a thick textbook and pretty easy to understand for beginners. Also it is suitable for future reference.

Bronstein, Semendjajew, Musiol, Mühlig,
Taschenbuch der Mathematik.
Verlag Harri Deutsch
ISBN: 3-8171-2005-2
→ The mother of all Math formulas! What is not there is either unsolvable or trivial.

Ralf Dinter, Institut für Angewandte Physik, Universität Hamburg, Februar 2011
Translated by
Samaneh Javanbakht, I. Institut für Theoretische Physik, Universität Hamburg, March 2015