6 Graphical handling of data with errors

Often you perform a series of experiments in which you vary an independent variable, such as temperature. What you are really interested in is the relation between the measured values and the independent variables, but the trouble is that your experimental values contain statistical deviations. You may already have a theory about the form of this relation and use the experiment to derive the still unknown parameters. It can also happen that the experiment is used to validate the theory or to decide on a modification. In this chapter a global view is taken and functional relations are qualitatively evaluated using simple graphical presentations of the experimental data. The trick of transforming functional relations to a linear form allows quick graphical interpretations. Even the inaccuracies of the parameters can be graphically estimated. If you want accurate results, then skip to the next chapter.

6.1 Introduction

In the previous chapter you have learned how to handle a series of equivalent measurements that should have produced equal results if there had been no random deviations in the measured data. Very commonly, however, a quantity y_i is measured as a function $f(x_i)$ of an independent variable x_i such as time, temperature, distance, concentration or bin number. The measured quantity may also be a function of several such variables. Usually the independent variables – which are under the control of the experimenter – are known with high accuracy and the dependent variables – the measured values – are subject to random errors. In that case

$$y_i = f(x_i) + \varepsilon_i, \tag{6.1}$$

where x_i is the independent variable (or the set of independent variables) and ε_i is a random sample from a probability distribution.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

Generally, you already have a theory about the function f, although that theory may contain unknown parameters $\theta_k (k = 1, \ldots, m)$:

$$y = f(x, \theta_1, \ldots, \theta_m). \tag{6.2}$$

An example is the linear relation

$$y = ax + b, \tag{6.3}$$

but the relation can be more complex like

$$y = c \exp(-kx). \tag{6.4}$$

It is often possible to *linearize* the relation by a simple transformation. For the latter case:

$$\ln y = \ln c - kx \tag{6.5}$$

yields a linear relation between $\ln y$ and x. It is usually recommended to make such a linearization, as a simple graphic plot will show a straight line, permitting a quick judgment of the suitability of your presumed functional relation. In Section 6.2 a few examples will be worked out.

Let us return to the linear relation $y = ax + b$. Suppose you measured n data points (x_i, y_i), $i = 1, \ldots, n$, and expect the measured values y_i to satisfy *as closely as possible* the relation

$$y_i \approx f(x_i), \tag{6.6}$$

where $f(x) = ax + b$ is the expected relation. Your task is to determine the parameters a and b such that the measured values y_i deviate as little as possible from the function values. But what does that mean? The deviations ε_i of the measured values with respect to the function:

$$\varepsilon_i = y_i - f(x_i) \tag{6.7}$$

should be the sole consequence of random errors and we expect in general that the deviations ε_i are random samples from a probability distribution with zero mean. In practice this distribution is often normal. The correct method for this kind of parameter estimation is the *least-squares fit*, which is treated in Chapter 7. A computer program is needed to perform a least-squares fit.

It is not always necessary to perform a precise least-squares fit. It is always meaningful to plot the data in such a way that you expect a linear relation. A straight line can be adequately judged by visual inspection. A straight line drawn "by eye" to fit the points often gives sufficiently accurate results and even the inaccuracies in the parameters a and b can be estimated by

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

varying the line within the cloud of measured data points. There is nothing wrong with making a quick sketch on old-fashioned graph paper! Computer programs are useful when there are many data points, when different points have different weights or when high accuracy is required, but they are never a substitute for bad measurements and almost never give you more insight into the functional relations. Be careful with computer programs that are not well-documented or do something you don't quite understand!

This chapter is devoted to simple graphical processing of experimental data with a simple discussion of the inaccuracies in the results. Always ask yourself if such a simple graphic analysis can be useful for your problem: often you get a better insight into the relation between model and data. After having done a simple analysis, a more accurate and elaborate computer analysis can (and should) be made.

6.2 Linearization of functions

In this section a few examples of the linearization of functions are given.

(i) $y = ae^{-kx}$: $\ln y = \ln a - kx$ (examples: concentration as function of time for a first-order reaction, number of counts per minute for a radioactive decay process). Plot $\ln y$ on a linear scale versus x, or plot y on a logarithmic scale versus x. If you do this by hand, use semi-log paper (one coordinate linear, the other logarithmic with e.g. two decades). Or use a simple Python plot. Figure 2.7 on page 16 is an example. The slope ($-k$ in this example) is read from the graph by selecting a segment (take a large segment for better accuracy) and read the coordinates of the end points (x_1, y_1) and (x_2, y_2); the slope equals $\ln(y_2/y_1)/(x_2 - x_1)$. If you take a full decade for the end points (e.g. passing through $y = 1$ and $y = 10$), then the slope is simply $\ln 10/(x_2 - x_1)$.

(ii) $y = a + be^{-kx}$: $\ln(y - a) = \ln b - kx$. First estimate a from the values of y for large x and then plot $y - a$ versus x on a logarithmic scale. If the plot doesn't yield a linear relation, adjust a somewhat (within reasonable bounds).

(iii) $y = a_1 e^{-k_1 x} + a_2 e^{-k_2 x}$. This is difficult to handle graphically, unless k_1 and k_2 are very different. A computer program also has difficulties with this kind of analysis! First estimate the "slow" component (with smallest k), subtract that component from y and plot the difference on a logarithmic scale. Figure 6.1 gives the result for the data given in Table 6.1; the standard error in each y is ± 1 unit.

The column z in Table 6.1 gives the differences between y and the values given by the line in the left panel of Fig. 6.1. This line has been drawn "by eye" and goes through the points (0, 25) and (100, 2.5), yielding $k_2 = [\ln(25/2.5)]/100 = 0.023$. Hence the equation for this line is

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

Table 6.1 *Measured values y that result from a sum of two exponentials. The column z results from subtraction of the "slowest" exponential. The standard uncertainty in y equals one unit.*

x	y	z	x	y	z
0	90.2	65.2	40	11.7	1.7
5	62.2	39.9	50	8.8	0.9
10	42.7	22.9	60	6.9	0.6
15	30.1	12.4	70	4.6	−0.4
20	23.6	7.8	80	5.0	1.1
25	17.9	3.8	90	2.9	−0.3
30	14.0	1.5			

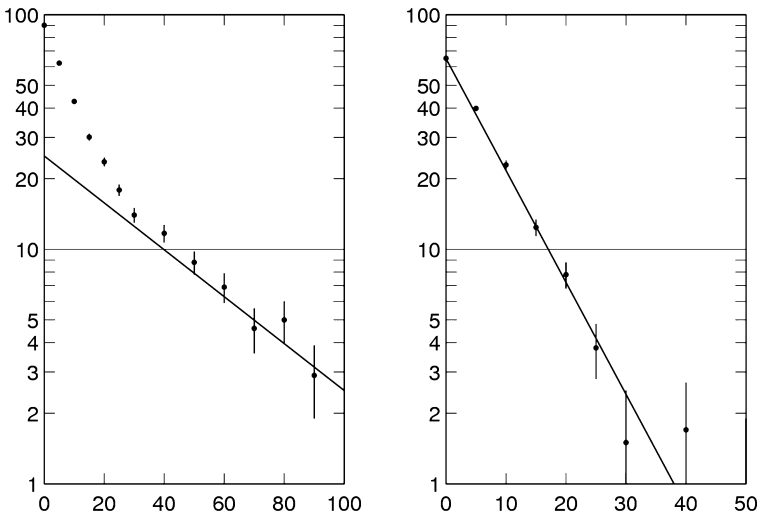

Figure 6.1 Graphical analysis of data which represent the sum of two exponentially decaying quantities. In the left panel the data points y have been plotted on a logarithmic scale versus the independent variable x and the "slowest" component is approximated by a straight line. In the right panel the differences z between the data y and the "slow" component are plotted. Note the different scales for x.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

$25\exp(-0.023x)$. In the right panel of this figure z has been plotted: the points approximately follow a linear relation. The drawn line goes through the points (0,65) and (38, 1), yielding $k_1 = (\ln 65)/38 = 0.11$. Therefore, the function that approximates the behavior of all data points is given by

$$f(x) = 65\, e^{-0.11\,x} + 25\, e^{-0.023\,x}. \tag{6.8}$$

This simple graphical approach does not provide a solid basis to make a reliable guess of the uncertainties in the parameters of this equation. But it provides an excellent basis for the *initial guess* of the parameters in a *nonlinear least squares fit*. The latter is the subject of Chapter 7. Such a fit must be carried out by computer; a suitable program not only provides the best fit, but also gives an estimate of the inaccuracies and correlations of the parameters.

(iv) $y = (x - a)^p$ (example: the isothermal compressibility χ of a fluid in the neighborhood of the critical temperature behaves as a function of temperature according to $\chi = C(T - T_c)^{-\gamma}$, where γ is the *critical exponent*). Plot $\log y$ versus $\log(x - a)$ (or y versus $(x - a)$ on a double-logarithmic scale); if a is not known beforehand, then vary a somewhat until the relationship becomes a straight line. The slope of the line yields p.

(v) $y = ax/(b + x)$ (examples: adsorbed quantity n_{ads} of a solute versus concentration c in solution or versus pressure p in the gas phase in the case of Langmuir-type adsorption: $n_{ads} = n_{max}c/(K + c)$; reaction rate v as function of substrate concentration $[S]$ in the case of Michaelis–Menten kinetics[1] $v = v_{max}[S]/(K_m + [S])$). By taking the reciprocal of both sides, this equation becomes a linear relation between $1/y$ and $1/x$:

$$\frac{1}{y} = \frac{1}{a} + \frac{b}{a}\frac{1}{x}. \tag{6.9}$$

In enzyme kinetics a graph of $1/v$ versus $1/[S]$ is called a *Lineweaver–Burk plot*.[2] There are two other ways to produce a linear relation: plot x/y versus x (the *Hanes method*):

$$\frac{x}{y} = \frac{b}{a} + \frac{x}{a}, \tag{6.10}$$

or plot y/x versus y (the *Eadie–Hofstee method*):

$$\frac{y}{x} = \frac{a}{b} - \frac{y}{b} \tag{6.11}$$

[1] This will be familiar to you if you are a biochemist, but sound as abacadabra if you are a physicist or mechanical engineer. You may consult any textbook on biochemistry for details. Or think of an application in your own field that leads to this kind of equation.

[2] See e.g. Price and Dwek (1979).

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

Table 6.2 *Conversion rate v of urea by the enzyme urease as function of the urea concentration [S]. The reciprocal values are given to produce a Lineweaver–Burk plot. The standard uncertainty in* $1/v$ *equals* σ_v/v^2.

[S] mM	1/[S] mM^{-1}	v mmol min^{-1}	σ_v mg^{-1}	$1/v$ $mmol^{-1}$ min	$\sigma_{1/v}$ mg
30	0.03333	3.09	0.2	0.3236	0.0209
60	0.01667	5.52	0.2	0.1812	0.0066
100	0.01000	7.59	0.2	0.1318	0.0035
150	0.00667	8.72	0.2	0.1147	0.0026
250	0.00400	10.69	0.2	0.09355	0.0018
400	0.00250	12.34	0.2	0.08104	0.0013

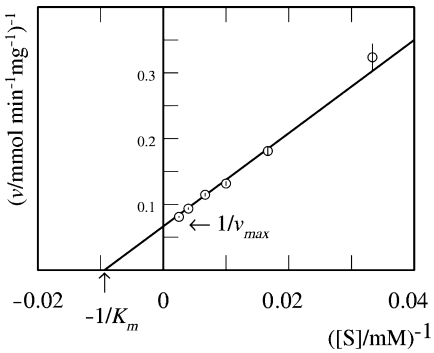

Figure 6.2 Lineweaver–Burk plot of the tabulated data.

Which method to choose depends on the inaccuracies of the data points: whenever a reciprocal of x or y is used, small values get relatively more importance in the plot.

Example: urease kinetics

With the experimental values for the rate of conversion $v = y$ of urea by the enzyme urease[3] as a function of the urea concentration [S]$= x$ as given in Table 6.2, the plots of Fig. 6.2 and Fig. 6.3 are obtained. In a Lineweaver–Burk plot the value of $K_m = b$ can be obtained from the intersection with

[3] Example taken from Price and Dwek (1979), with additional noise.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

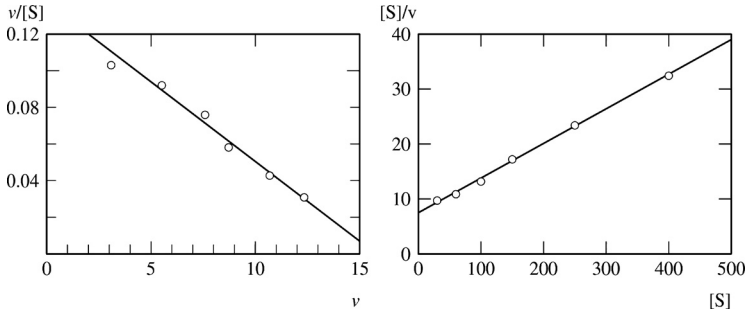

Figure 6.3 Eadie–Hofstee (left) and Hanes (right) plot of the tabulated data.

the horizontal (x) axis and the value of $v_{max} = a$ can be obtained from the intersection with the vertical (y) axis. The estimation of inaccuracies of the parameters from these graphs is not reliable; also in this case it is better to perform a nonlinear least-squares analysis using the graphical estimates for the initial guess of the parameters.

6.3 Graphical estimates of the accuracy of parameters

In the previous section you have seen how you can plot your data in such a way that a linear relationship is obtained and how you can estimate the two parameters of a linear function by drawing the "best" line through the data points. In this section you will see how you can make a simple estimate of the uncertainties in those parameters. Sometimes such estimates are sufficient. If they are not, a more accurate *least-squares fit* is required.

In order to be able to estimate the uncertainties, you need to include error bars in the graphs. When the uncertainty in the independent variable x, plotted on the horizontal scale, is negligible, it suffices to use vertical error bars from $y - \sigma_y$ to $y + \sigma_y$. When there is a sizeable uncertainty in x, a horizontal error bar from $x - \sigma_x$ to $x + \sigma_x$ must be included as well. A clear presentation is an *ellipse* with principal axes of length $2\sigma_x$ and $2\sigma_y$.

The best straight line through the data points fits as closely as possible to all (x_i, y_i). The first requirement is that the line be drawn such that the sum of the deviations (sign included) is (close to) zero. But that does not determine the line! Any line through the "center of mass"[4] of the points $(\langle x\rangle, \langle y\rangle)$ fulfills this criterion. We need this criterium to be fulfilled not only globally, but also locally. A good guess is the line constructed through *two* centers of mass, each of a group of data points, see Fig. 6.4.

[4] "Mass" is to be interpreted as "statistical weight."

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

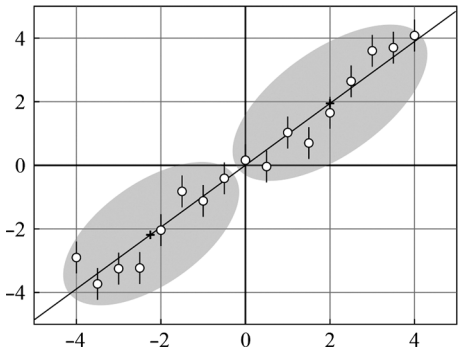

Figure 6.4 A line drawn through two "centers of mass" of two clouds of points approximates a linear fit to all points.

After drawing a straight line $f(x) = ax + b$, the parameters a and b can be determined from the slope and the value at which the line intersects the y-axis. The latter may be difficult to determine when the value $x = 0$ is outside the range of x-values of the data points. A much better method is to determine the "center of mass" $(\langle x\rangle, \langle y\rangle)$ of the points. The best fit should go through this point, as you shall see in Chapter 7. Only the slope a needs to be estimated:

$$f(x) = a(x - \langle x\rangle) + b, \tag{6.12}$$

$$b = \langle y\rangle. \tag{6.13}$$

The use of this relation has the advantage that uncertainties in the slope and the additive constant are uncorrelated with each other (see page 90). It is now much easier to estimate the uncertainties in a and b.

In order to estimate the uncertainties in the parameters, the line can be varied in slope a (Fig. 6.5) or in additive constant b (Fig. 6.6). As we know from the properties of a normal distribution, about 2/3 of the points should remain within the lines if a parameter is varied by $\pm\sigma$. So, as a rule of thumb, vary the parameters (one at a time) symmetrically such that about 15 percent of the points fall outside the lines on each side. Be aware of possible outliers that deviate conspicuously from the line. How to handle outliers has been treated in Section 5.7 on page 63: either eliminate or measure again!

6.4 Using calibration

Suppose you work with an instrument or method that produces a *reading* y (e.g. a digital number, a needle deflection, a meniscus height) from which

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

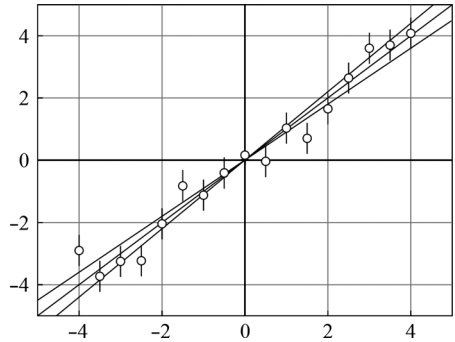

Figure 6.5 Linear fit through "center of mass" with slope varied by $\pm 10\%$ ($a = 1.0 \pm 0.1$).

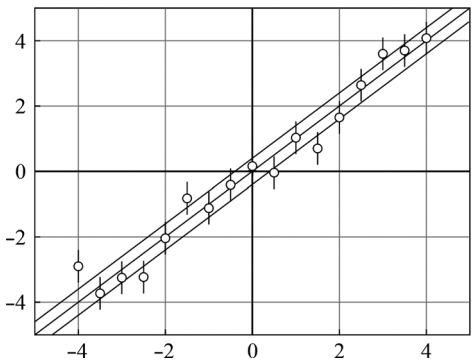

Figure 6.6 Linear fit through "center of mass" with additive constant varied by ± 0.4 ($b = 0.0 \pm 0.4$).

a quantity x (e.g. a concentration, an electrical current, a pressure) must be deduced. When the instrument is not properly calibrated, i.e., when the reading does not correspond directly and reliably to the measured quantity, you should calibrate the instrument yourself. For this purpose you produce a *calibration table*, and preferably a *calibration curve*, by measuring the reading for a number of accurately known values of x. These data you either tabulate, or plot and interpolate in a curve, or express the relation between y and x in a mathematical function. Often you will tabulate a *correction table* or plot a *correction curve* that contains the differences between the readings and the correct values. Be sure to be explicit what the difference means: usually the correction is to be added to the reading to obtain the true value. In all cases you can deduce the value of x for any measured reading by inversion

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

Table 6.3 *Compass deviation chart of the U.S.S. Cleveland (1984) listing the compass deviations (dev) from the true magnetic bearing for various ship headings (head). The deviation W (West) means negative and E (East) means positive; the deviation is to be added to the compass reading to obtain the true magnetic bearing of the ship.*

head	dev	head	dev	head	dev	head	dev
0	1.5W	90	1.0W	180	0.0	270	1.5E
15	0.5W	105	2.0W	195	0.5E	285	0.0
30	0.0	120	3.0W	210	1.5E	300	0.5W
45	0.0	135	2.5W	225	2.5E	315	2.0W
60	0.0	150	2.0W	240	2.0E	330	2.5W
75	0.5W	165	1.0W	255	2.5E	345	2.0W

of the calibration relation. How do you proceed and how do you determine the uncertainty in x?

Be explicit!

Mariners and navigators have coped with magnetic compass corrections for centuries, although modern electronic aids have diminished their problems. The compass reading (C) must be corrected first by adding the *deviation* due to the influence of magnetic and ferrous materials in the ship itself to obtain the magnetic bearing (M); then the latter must be corrected by adding the *variation* due to the position of the magnetic north pole – that does not coincide with the true geometric north pole – to obtain the true bearing (T). Traditionally deviation and variation are expressed as E (East) if positive, or W (West) if negative. Since a sign error can have catastrophical consequences, sailors of all nations have invented mnemonics to remind them of the proper sequence to add or subtract the corrections. A mannerly English mnemonic is CADET: "**C**ompass **AD**d **E**ast (to get) **T**rue (bearing)", which applies equally to deviation and variation. In the Dutch Navy Reserve (KMR) the mnemonic "**K**ies **d**e **M**eisjes **v**an **R**otterdam" ("choose the girls of Rotterdam"): **K**ompas + **d**eviatie → **M**agnetisch + **v**ariatie → **R**echtwijzend (True bearing) is more popular. But beware: American navigators reverse the correction by the mnemonics "**T**rue **V**irgins **M**ake **D**ull **C**ompany" (**T**rue + **V**ariation → **M**agnetic + **D**eviation → **C**ompass), which is wrong unless the sign of the correction is also reversed. To remember this, they also memorize "**A**dd **W**hiskey" to **A**dd **W**esterly corrections. So be careful and explicit in all cases. See Table 6.3[5] and Fig. 6.7.

[5] Data from www.tpub.com/context/administration/14220/css/14220_64.htm.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

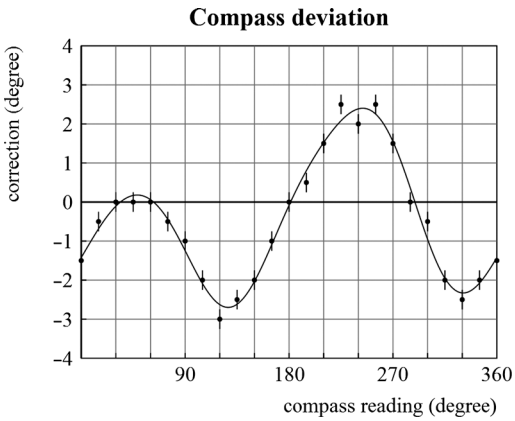

Figure 6.7 Graph of the compass deviations (Table 6.3). The error bars are ± 0.25 degree, as the corrections are given with 0.5 degree precision. The drawn line is a least squares fit to a sum of Fourier components up to and including the fourth harmonic.

Python code 6.1 on page 182 shows how the least-squares Fourier components in Fig. 6.7 are computed. For general least square fits see Section 7.3.

Make sure in the calibration procedure that you cover the whole range of values for which the method will be used. Extrapolation is generally unreliable, but there is also no need to cover values that in practice will never occur. Draw the best line through the points; if the line is not straight, hopefully you can build it up from straight segments between calibrated points. If you want to be sophisticated, compute a *cubic spline* fitting function. Now, for any new measurement of x, given by a reading y, the quantity x can simply be read back from the calibration curve.

Now consider the inaccuracy of a measurement. There are two sources of error: one is the inaccuracy Δy in the reading y; the other is the inaccuracy in the calibration curve itself, due to inaccuracies of the calibration measurements. You should also be aware of additional errors that may occur, e.g. resulting from aging of the instrument after the last calibration. Both types of error lead to an uncertainty in x and both sources add up quadratically, because they are independent of each other. The two contributions are depicted in Fig. 6.8, which shows how a concentration in solution is deduced from a measurement of the optical density in a spectrometer. The

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

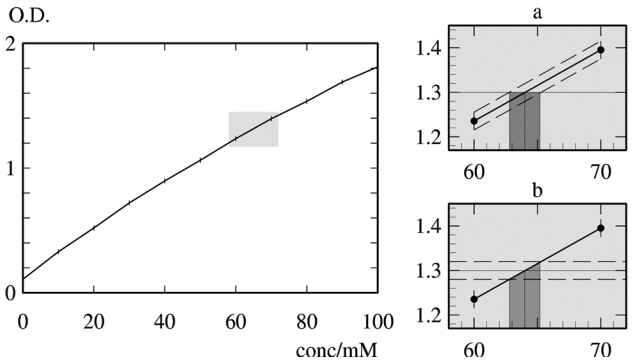

Figure 6.8 Example of a calibration line for spectrometric determination of the concentration of a chromophore in solution: optical density O.D. = log (*incident intensity/transmitted intensity*) as a function of the concentration of the solute. The gray area is magnified in the panels on the right: (a) the calibration error in the concentration, (b) the inaccuracy in the concentration resulting from the inaccuracy of the measured O.D.

calibration error is visualized by drawing two parallel sections of the calibration curve at distances representing the standard uncertainty in the calibration itself.

If the calibration has been very carefully performed, the calibration error is likely to be smaller than the direct error in the reading. In that case only the standard uncertainty σ_y of the reading counts. It leads to a standard uncertainty σ_x in the measured quantity by the relation

$$\sigma_x = \frac{\sigma_y}{\left|\left(\frac{dy}{dx}\right)_{\text{cal}}\right|}. \tag{6.14}$$

Summary *In this chapter you have learned how to plot your data in such a way that a functional relation becomes visible, preferably as a straight line. From simple plots you can roughly estimate the parameters of your function and – by varying the lines in position or slope – you can even get an idea of the inaccuracies of the parameters. You have also seen how calibrations are used to interpret instrument readings. You will not make errors in the sign when you apply calibrated corrections. The treatment in this chapter was rather sloppy, as its purpose was to provide a quick insight into your data. For more precision, proceed to the next chapter.*

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007

Exercises

6.1 Draw a straight line "through" the points of Fig. 2.7 on page 16 and determine the parameters in $c(t) = c_0\, e^{-kt}$.

6.2 From Figs. 6.2 and 6.3, determine the values of v_{max} and K_m. The straight lines drawn "by eye" go through the points $(-0.0094, 0)$ and $(0.04, 0.35)$ (Lineweaver–Burk), $(0.04, 0.35)$ and $(15, 0.007)$ (Eadie–Hofstee); $(0, 7.5)$ and $(500, 39)$ (Hanes).

6.3 Draw the best straight line through the data points of the logarithmic graph of k versus $1000/T$, made in Exercise 3.2 (page 25). Determine the constant E in the relation $k = A \exp(-E/RT)$ (which units?). Estimate the inaccuracy in E.

6.4 Using Fig. 6.8, determine the concentration (with s.d.) when the measured optical density equals 1.38 ± 0.01, assuming that the calibration error is negligible.

Downloaded from https://www.cambridge.org/core. Universitaetsbibliothek Duisburg-Essen, on 13 Jul 2018 at 17:16:17, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511921247.007