PART I

Data and error analysis

- 1 Introduction *page* 3
- 2 The presentation of physical quantities with their inaccuracies 5
- 3 Errors: classification and propagation 18
- 4 Probability distributions 27
- 5 Processing of experimental data 53
- 6 Graphical handling of data with errors 71
- 7 Fitting functions to data 84
- 8 Back to Bayes: knowledge as a probability distribution 111

1 Introduction

It is impossible to measure physical quantities without errors. In most cases errors result from deviations and inaccuracies caused by the measuring apparatus or from the inaccurate reading of the displaying device, but also with optimal instruments and digital displays there are always fluctuations in the measured data. Ultimately there is random thermal noise affecting all quantities that are determined at a finite temperature. Any experimentally determined quantity therefore has a certain inaccuracy. If the experiment were to be repeated, the result would be (slightly) different. One could say that the result of a particular experiment is no more than a *random sample* from a probability distribution. When reporting the result of an experiment, it is important to also report the extent of the uncertainty, e.g. in terms of the best estimate of some measure of the *width* of the probability distribution. When experimental data are processed and conclusions are drawn from them, knowledge of the experimental uncertainties is essential to assess the reliability of the conclusion.

Ideally, you should specify the probability distribution from which the reported experimental value is supposed to be a random sample. The problem is that you have only one experiment; even if your experiment consists of many observations of which you report the average, you have only one average to report. So you have only one sample of the reported item and you could naively conclude that you have no knowledge at all about the underlying probability distribution of that sample. Fortunately, there is the science of statistics that tells us differently. When your experiment consists of a series of repeated observations of a variable x, with outcomes $x_1, x_2, \ldots, x_n$, and you report the result of the total experiment as the average of the x_i's, statistics tells you how to *estimate* certain properties of the probability distribution of which the reported result is supposed to be a random sample. Thus you can estimate the mean of the distribution or – if you prefer – the most probable value of the distribution, which then is the result of your measurement. You can also estimate the width of the distribution, which indicates the random uncertainty in the result.

The result of an experiment is generally not equal to a directly measured quantity, but is derived from measured quantities by some functional relation.

For example, the area of a rectangle is the product of the measured length and width of two sides. Each measurement has its estimated value and random error and these errors *propagate* through the functional relation (here a product) to the final result. The contributing errors must be properly combined to one error estimate in the result.

The purpose of this book is to indicate how one can arrive at the best estimates of both the value(s) and the random error(s) in the result, based on the measurements from which the result is derived. In order to maintain its usefulness as a practical guide, the main part of this book simply states the equations and procedures, without proper derivations. Thus the practical applicant is not bothered by unnecessary detail. However, several appendices are included that provide further details and give a proper background in statistics with derivations of the equations used. For further reading many textbooks are available.[1]

Chapter 2 describes the proper presentation of results of measurements with their accuracies and with their units. Chapter 3 classifies the various types of error and describes how contributing errors will propagate and combine into a more complex result. Chapter 4 describes a number of common probability distributions from which experimental errors may be sampled. In Chapter 5 it is shown how the characteristics of a *data series* can be defined and then be used to arrive at estimates of the best value and accuracy of the result. Chapter 6 is concerned with simple graphic treatment of data, while Chapter 7 treats the more accurate *least-squares* fitting of model parameters to experimental data. Chapter 8, finally, discusses the philosophical basis of statistical methods, confronting traditional hypothesis testing with the more intuitive but powerful *Bayesian* method to determine the probability distribution of model parameters.

[1] Most textbooks aim at a wider audience and are therefore less useful for physical scientists and engineers. For the latter interest group see Bevington and Robinson (2003), Taylor (1997), Barlow (1989) and Petruccelli *et al.* (1999).

2 The presentation of physical quantities with their inaccuracies

This chapter is about the *presentation* of experimental results. When the value of a physical quantity is reported, the uncertainty in the value must be properly reported too, and it must be clear to the reader what kind of uncertainty is meant and how it has been estimated. Given the uncertainty, the value must be reported with the proper number of digits. But the quantity also has a unit that must be reported according to international standards. Thus this chapter is about reporting your results: this is the last thing you do, but we'll make it the first chapter before more serious matters require attention.

2.1 How to report a series of measurements

In most cases you derive a result on the basis of a series of (similar) measurements. In general you do not report all individual outcomes of the measurements, but you report the best estimates of the quantity you wish to "measure," based on the experimental data and on the model you use to derive the required quantity from the data. In fact, you use a *data reduction method*. In a publication you are *required* to be explicit about the method used to derive the end result from the data. However, in certain cases you may also choose to report details of the data themselves (preferably in an appendix or deposited as "additional material"); this enables the reader to check your results or apply alternative data reduction methods.

List all data, a histogram or percentiles

The fullest report of your experimental data is a list or table of all data. Almost[1] equivalent is the report of a *cumulative distribution* of the data (see Section 5.1 on page 54). Somewhat less complete is reporting a *histogram* after collecting data in a limited number of intervals, called *bins*. Much less

[1] Not quite, because one loses information on possible sequential correlation between data points.

Table 2.1 *Thirty observations, numbered in increasing order.*

1	6.61	6	7.70	11	8.35	16	8.67	21	9.17	26	9.75
2	7.19	7	7.78	12	8.49	17	9.00	22	9.38	27	10.06
3	7.22	8	7.79	13	8.61	18	9.08	23	9.64	28	10.09
4	7.29	9	8.10	14	8.62	19	9.15	24	9.70	29	11.28
5	7.55	10	8.19	15	8.65	20	9.16	25	9.72	30	11.39

complete is to report certain *percentiles* of the cumulative distribution, usually the 0, 25%, 50%, 75% and 100% values (i.e., the full range, the median and the first and third quartiles). This is done in a *box-and-whisker* display. See the example below.

List properties of the data set

The methods above are *rank-based* reports: they follow from ranking the data in a sequence. You can also report *properties* of the set of data, such as the number of observations, their average, the mean squared deviation from the average or the root of that number, the correlation between successive observations, possible outliers, etc. Note that we do not use the names *mean, variance, standard deviation*, which we reserve for properties of probability distributions, not data sets. Use of these terms may cause confusion; for example, the *best estimate* for the variance of the parent probability distribution – of which the data set is supposed to be a random sample – is not equal to the mean squared deviation from the average, but slightly larger ($n/(n-1)\times$). See Section 5.3 on page 58.

Example: 30 observations

Suppose you measure a quantity x and you have observed 30 samples with the results as given in Table 2.1. Figure 2.1 shows the cumulative distribution function of these data and Fig. 2.2 shows the same, but plotted on a "probability scale" which should produce a straight line for normal-distributed data. A histogram using six equidistant bins is shown in Fig. 2.3. It is clear that this sampling is rather unevenly distributed.

These numbers and cumulative distributions were generated with **Python code** 2.1 on page 171

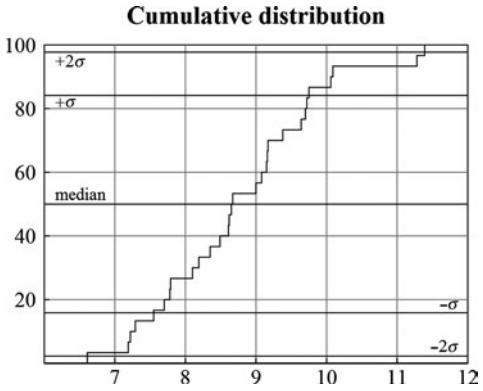

Figure 2.1 The cumulative distribution function of thirty observations. The vertical scale represents the cumulative percentage of the total.

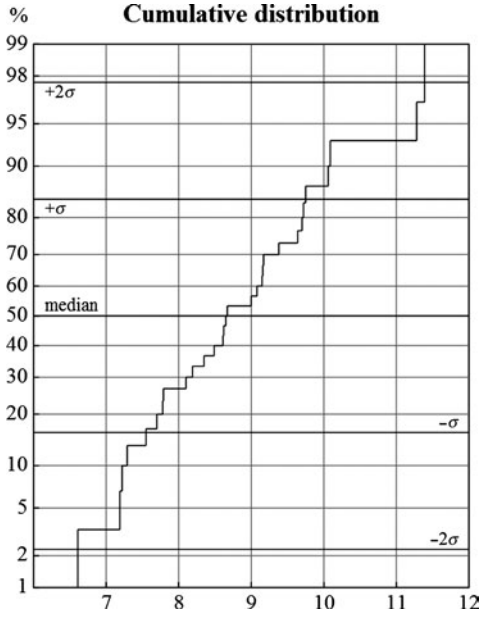

Figure 2.2 The cumulative distribution function of thirty observations. The vertical scale represents the cumulative percentage of the total on a probability scale, designed to produce straight lines for normal distributions.

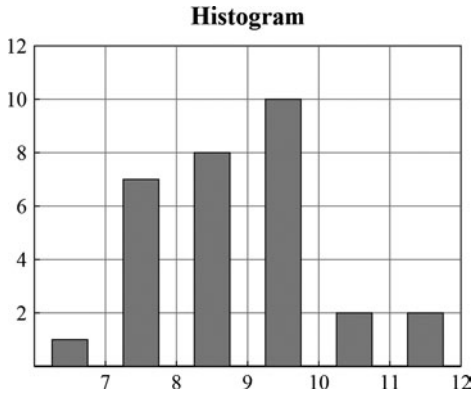

Figure 2.3 A histogram of thirty observations. The data have been gathered in six equidistant bins. The vertical scale gives the number of observations in each bin.

The histogram of Fig. 2.3 was generated with **Python code** 2.2 on page 171

The *properties* of the data set you can report are:

(i) *number of observations*: $n = 30$
(ii) *average*: $m = 8.78$
(iii) *mean squared deviation from average*: msd $= 1.28$
(iv) *root-mean-square deviation from average*: rmsd $= 1.13$

The properties are available as array methods or functions. See **Python code** 2.3 on page 171

Other rank-based properties of the data set are values that exceed a given fraction of the data, such as the *median* (at 50%), the first and third *quartile* (at 25 and 75%) or the p-th *percentile*. The latter is a value x_p such that $p\%$ of the data has a value $\leq x_p$ and $(100 - p)\%$ has a value $> x_p$.[2] The total *range* is the interval between the minimum and maximum values. Figure 2.4 shows the data as a box-and-whisker display of the total range (the whisker) and the quartiles (the box).

A simple program to determine a series of percentiles is **Python code** 2.4 on page 172

[2] There may be an ambiguity here. The p-th percentile may be exactly one of the data values, e.g. the median equals the 5th value out of a set of 9. In general, the percentile will fall in a range between two values, e.g. the median lies between the 5th and the 6th value out of a set of 10 values. In that case linear interpolation is used.

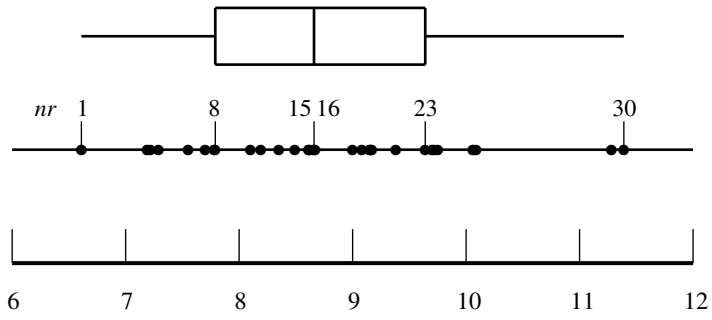

Figure 2.4 A *box-and-whisker* display of the total range, the median and the first and third quartile of thirty observations. Note that the median falls between nr 15 and nr 16 (15 observations on both sides); the average between the two values is taken.

2.2 How to represent numbers

Decimal separator: comma or period?

In the English language and in all "computer languages" (and among others also in China, Israel and Switzerland) the decimal *point* is used as separator between the integer and fractional parts of a real decimal number. In many other languages (all other European languages, Russian and related languages) the decimal *comma* is used instead. Be consistent and adhere to what your language requires! In order to avoid confusion, scientists are strongly advised *not* to use periods or commas to divide long numbers into groups of three digits, like 300,000 (English) or 300.000 (e.g. French). Instead, use a *space* (or even better, if your text editor allows it, a *thin space*) to separate groups of three digits: 300 000.[3]

Significant figures

The end result of a measurement must be presented with as many digits as are compatible with the accuracy of the result. Also when a number ends with zeros! These are the *significant figures* of the result. However, intermediate results in a calculation should be expressed with a higher precision in order to prevent accumulation of rounding errors. Always indicate the accuracy of the end result! If the accuracy is not explicitly given, it is assumed that the error in the last digit is ± 0.5.

[3] This is the IUPAC recommendation, see http://old.iupac.org/reports/provisional/guidelines.html#printing

Examples, for the English language

(i) 1.65 ± 0.05
(ii) 2.500 ± 0.003
(iii) $35\,600 \pm 200$: better as $(3.56 \pm 0.02) \times 10^4$
(iv) 5.627 ± 0.036 is allowed, but makes sense only when the inaccuracy itself is known with sufficient accuracy. If not, this value should be written as 5.63 ± 0.04.
(v) Avogadro's number is known as $(6.022\,141\,79 \pm 0.000\,000\,30) \times 10^{23}$ mol^{-1} (CODATA 2006). The notation $6.022\,141\,79(30) \times 10^{23}$ mol^{-1} is a commonly accepted abbreviation.
(vi) 2.5 means 2.50 ± 0.05
(vii) 2.50 means 2.500 ± 0.005
(viii) In older literature one sometimes finds a subscript $_5$, indicating an inaccuracy of about one quarter in the last decimal: $2.3_5 = 2.35 \pm 0.03$, but this is not recommended.

When inaccuracies must be rounded, then do this in a conservative manner: when in doubt, round up rather than down. For example, if a statistical calculation yields an inaccuracy of 0.2476, then round this to 0.3 rather than 0.2, unless the statistics of your measurement warrants the expression in two decimals (0.25). See Section 5.5 on page 60. Be aware of the fact that calculators know nothing about statistics and generally suggest a totally unrealistic precision.

2.3 How to express inaccuracies

There are many ways to express the (in)accuracy of a result. When you report an inaccuracy it must be absolutely clear which kind of inaccuracy you mean. In general, when no further indication is given, it is assumed that the quoted number represents the *standard deviation* or *root-mean-square error* of the estimated probability distribution.

Absolute and relative errors

You can indicate inaccuracies as *absolute*, with the same dimension as the reported quantity, or as a dimensionless *relative* value. Absolute inaccuracies are often given as numbers in parentheses, relating to the last decimal(s) of the quantity itself.

Examples

(i) 2.52 ± 0.02
(ii) $2.52 \pm 1\%$
(iii) 2.52(2)
(iv) $N_A = 6.022\,141\,79(30) \times 10^{23}$ mol^{-1}