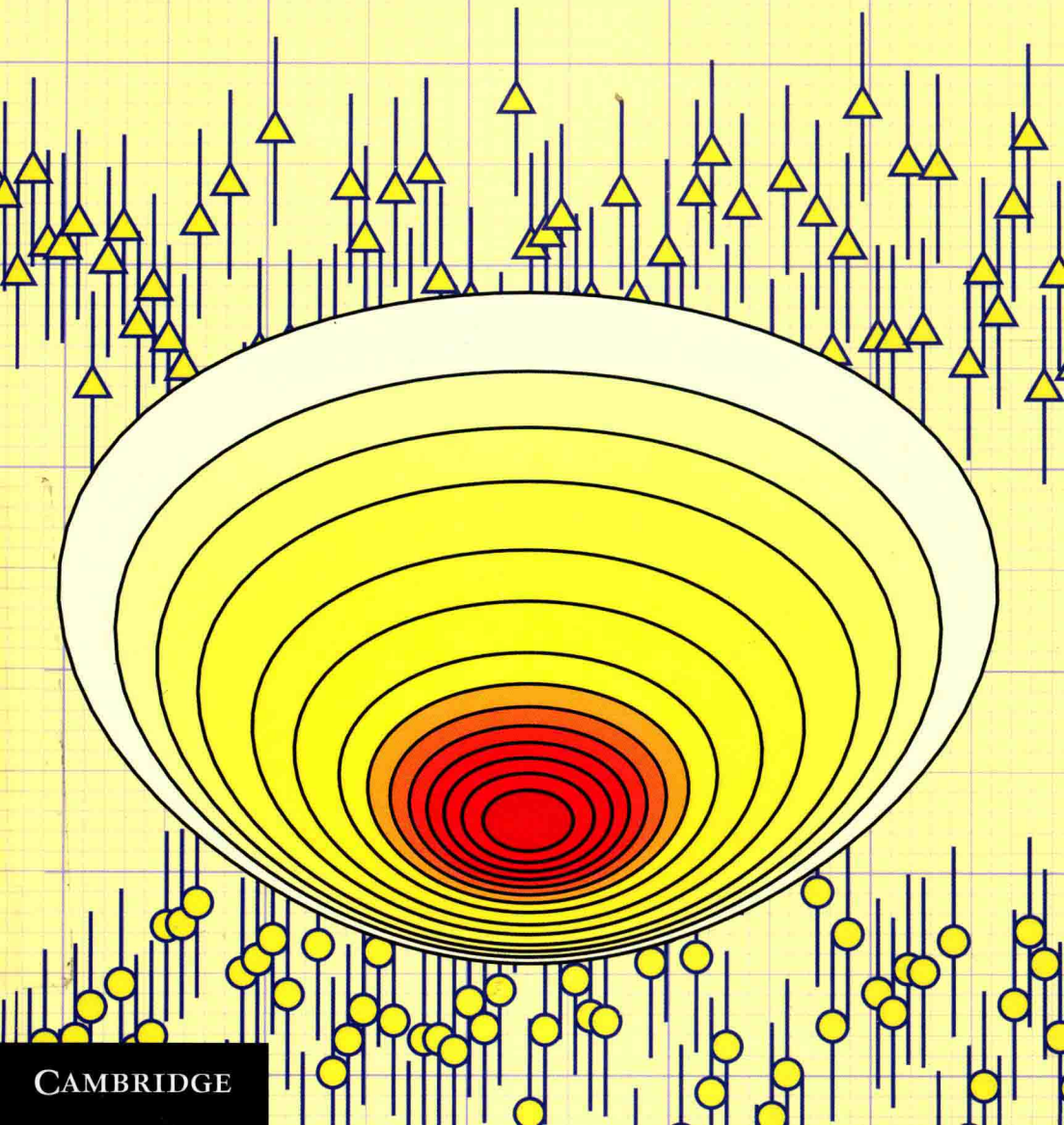
A Student's Guide to Data and Error Analysis

HERMAN J. C. BERENDSEN



CAMBRIDGE

A Student's Guide to Data and Error Analysis

HERMAN J. C. BERENDSEN

Emeritus Professor of Physical Chemistry,
University of Groningen, the Netherlands

CAMBRIDGE
UNIVERSITY

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521119405

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Berendsen, Herman J. C.
A student's guide to data and error analysis / Herman J.C. Berendsen.
p. cm.
ISBN 978-0-521-11940-5 (Hardback) – ISBN 978-0-521-13492-7 (pbk.)
1. Error analysis (Mathematics) I. Title.
QA275.B43 2011
511′.43–dc22

2010048231

ISBN 978-0-521-11940-5 Hardback
ISBN 978-0-521-13492-7 Paperback

A Student's Guide to Data and Error Analysis

All students taking laboratory courses within the physical sciences and engineering will benefit from this book, whilst researchers will find it an invaluable reference. This concise, practical guide brings the reader up to speed on the proper handling and presentation of scientific data and its inaccuracies. It covers all the vital topics with practical guidelines, computer programs (in Python), and recipes for handling experimental errors and reporting experimental data. In addition to the essentials, it also provides further background material for advanced readers who want to understand how the methods work. Plenty of examples, exercises, and solutions are provided to aid and test understanding, whilst useful data, tables, and formulas are compiled in a handy section for easy reference.

HERMAN J. C. BERENDSEN is Emeritus Professor of Physical Chemistry at the University of Groningen, the Netherlands. His research started in nuclear magnetic resonance, but focused later on molecular dynamics simulations on systems of biological interest. He is one of the pioneers in this field and, with over 37 000 citations, is one of the most quoted authors in physics and chemistry. He has taught courses in molecular modeling worldwide and authored the book *Simulating the Physical World* (Cambridge University Press, 2007).

To my wife and daughters

Preface

This book is written as a guide for the presentation of experimental data including a consistent treatment of experimental errors and inaccuracies. It is meant for experimentalists in physics, astronomy, chemistry, life sciences and engineering. However, it can be equally useful for theoreticians who produce simulation data: they are often confronted with statistical data analysis for which the same methods apply as for the analysis of experimental data. The emphasis in this book is on the determination of best estimates for the values and inaccuracies of parameters in a theory, given experimental data. This is the problem area encountered by most physical scientists and engineers. The problem area of experimental design and hypothesis testing – excellently covered by many textbooks – is only touched on but not treated in this book.

The text can be used in education on error analysis, either in conjunction with experimental classes or in separate courses on data analysis and presentation. It is written in such a way – by including examples and exercises – that most students will be able to acquire the necessary knowledge from self study as well. The book is also meant to be kept for later reference in practical applications. For this purpose a set of "data sheets" and a number of useful computer programs are included.

This book consists of parts. Part I contains the main body of the text. It treats the most common statistical distributions for experimental errors and emphasizes the error processing needed to arrive at a correct evaluation of the accuracy of a reported result. It also pays attention to the correct reporting of physical data with their units. The last chapter considers the inference of knowledge from data from a Bayesian point of view, hopefully inducing the reader to sit back and think. The material in Part I is kept practical, without much discussion of the theoretical background on which the various types of analysis are based. This will not at all satisfy the eager student who has sufficient background in mathematics and who wishes to grasp a fuller understanding of the principles involved. Part II is to satisfy the curious: it contains several Appendices that explain various issues in more detail and provide derivations of the equations quoted in Part I. The Appendices in

Part II obviously require more mathematical skills (in particular in the field of linear algebra) than Part I. Part III contains Python code examples and Part IV provides answers to exercises. Finally, Part V contains practical information in the form of a number of "data sheets" which provide reference data in a compact form.

Throughout the book computer programs are included to facilitate the computations needed for applications. There are several professional software packages available for statistical data analysis. In the context of an educational effort, I strongly advise against the use of a specialized "black-box" software package that can be easily misused to produce ill-understood results. A "black-box" computer program should never be a magic substitute for a method that is not understood by the user! If a software package is to be used, it should provide general mathematical and graphical tools, preferably in an interactive way using an interpreter rather than a compiler. The commercial packages MATHEMATICA, MATLAB and MATHCAD are suitable for this purpose. However, most readers of this book will not have access to any or all of these packages, or – if they have temporary access through their institution – may not be able to continue access at a later point in time. Therefore for this book the choice was made to use the generally available, actively developing, open-source interpretative language PYTHON. With its array-handling and scientific extensions NUMPY and SCIPY the capabilities of this language come close to those of the commercial packages. Software related to this book, including a Python module `plotsvg.py` providing easy plotting routines, can be found on www.hjcb.nl/.

This book is the successor of the Dutch textbook *Goed meten met fouten* (Berendsen, 1997) that has been used in courses at the departments of physics and chemistry of the University of Groningen since 1997. The author is indebted to Emile Apol, A. van der Pol and Ruud Scheek for corrections and suggestions. Comments from readers are welcome to author@hjcb.nl.

Contents

PART I

Data and error analysis

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

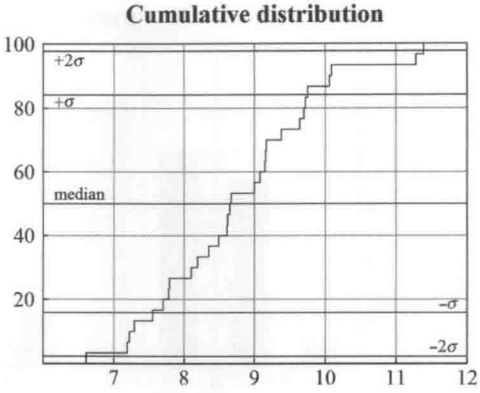


Figure 2.1 The cumulative distribution function of thirty observations. The vertical scale represents the cumulative percentage of the total.

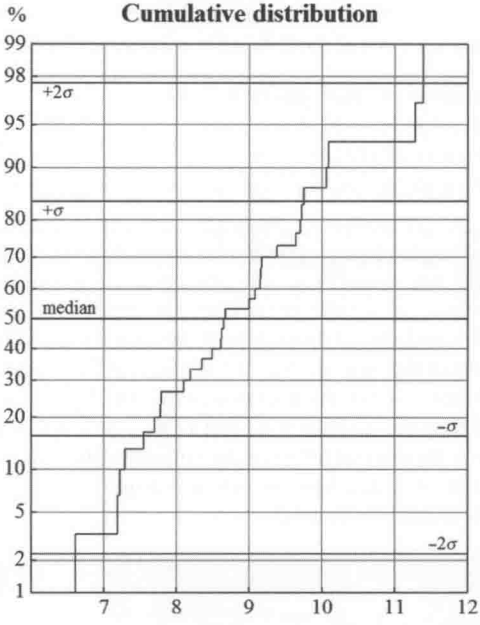


Figure 2.2 The cumulative distribution function of thirty observations. The vertical scale represents the cumulative percentage of the total on a probability scale, designed to produce straight lines for normal distributions.