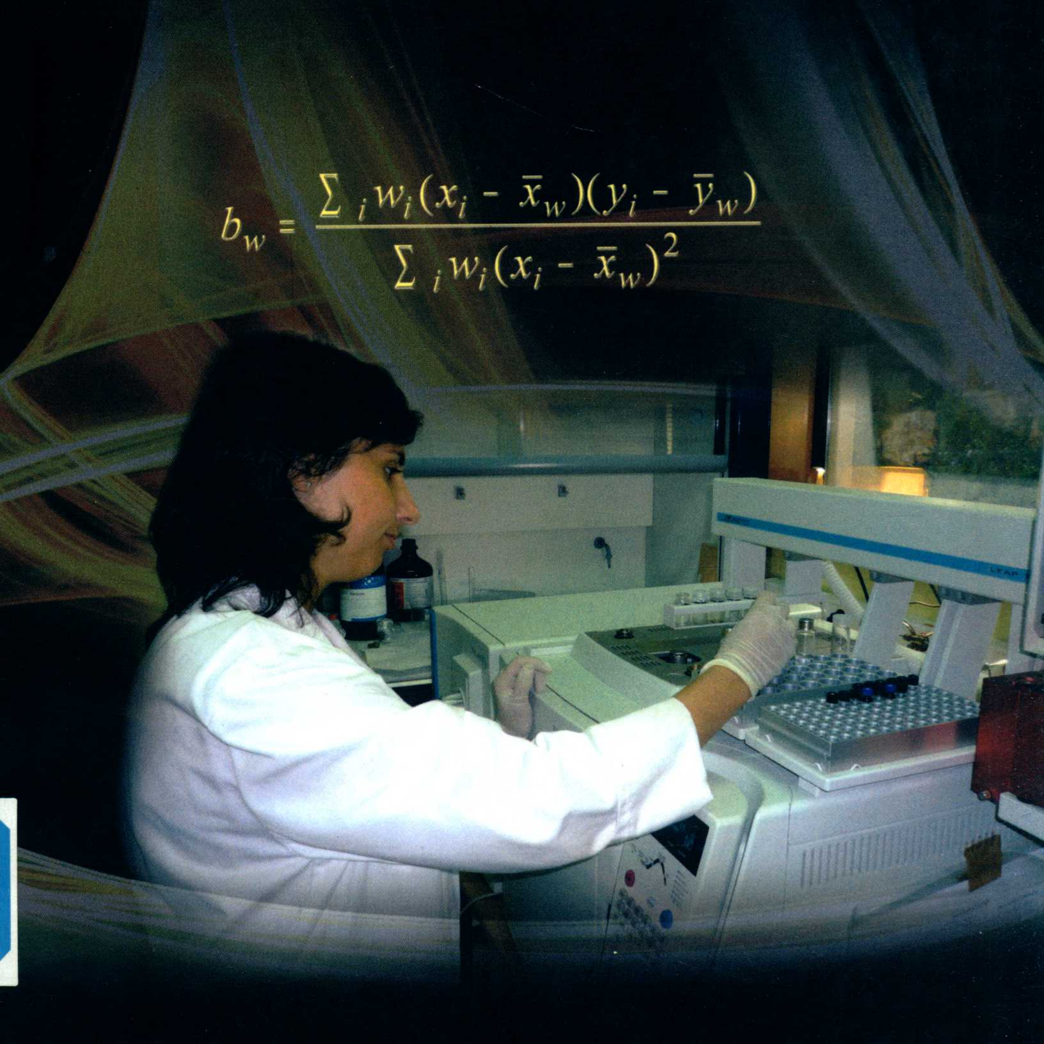
Notes on Statistics and Data Quality for Analytical Chemists

Michael Thompson • Philip J Lowthian



Imperial College Press

Notes on Statistics and Data Quality for Analytical Chemists

Michael Thompson, Philip J Lowthian
Birkbeck University of London, UK

ICP
Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

NOTES ON STATISTICS AND DATA QUALITY FOR ANALYTICAL CHEMISTS

ISBN-13 978-1-84816-616-5
ISBN-10 1-84816-616-8
ISBN-13 978-1-84816-617-2(pbk)
ISBN-10 1-84816-617-6(pbk)

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore by B & Jo Enterprise Pte Ltd

Notes on Statistics and Data Quality for Analytical Chemists

Preface

This book is based on the experience of teaching statistics for many years to analytical chemists and the specific difficulties that they encounter. Many analytical chemists find statistics difficult and burdensome. That is a perception that we hope to dispel. Statistics is straightforward and it adds a fascinating extra dimension to the science of chemical measurement. In fact it is hardly possible to conceive of a measurement science such as analytical chemistry that does not have statistics as both its conceptual foundation and its everyday tool. Measurement results necessary have uncertainty and statistics shows us how to make valid inferences in the face of this uncertainty. But, over the years, it has become apparent to us that statistics is much more interesting when it makes full use of the computer revolution.

It would be hard to overstate the effect that easily-available computing has had on the practice of statistics. It is now possible to undertake, often in milliseconds and with perfect accuracy, calculations that previously would have been impracticably longwinded and error-prone. A simple example is the calculation of the probabilities associated with density functions. Moreover, we can now produce in seconds several accurate graphical images of our datasets and select the most informative. These capabilities have transformed the applicability of both standard statistical methods and more recent computer-intensive methods. Textbooks for the most part have not caught up with this revolution and, to an unnecessary degree, are still stressing pencil-and-paper methods of calculation. Of course, a small number of pencil-and-paper examples of some elementary examples can assist learning, but they are too prone to mistakes for 'real-life' application. Many textbooks place a heavy stress on the mathematical basis of statistics. We regard this as inappropriate in an applied text. Analytical chemists do not need too many details of statistical theory, so we have kept these to a minimum. Drivers don't need to know exactly how every part of a car works in order to drive competently.

With ease of computation, there is, of course, a concomitant danger that people are tempted to use one of the many excellent computer statistics packages (or perhaps one of the not-so-excellent ones) without understanding what the output means or whether an appropriate method has been used. Analytical chemists have to guard against that serious shortcoming by exercising a proper scientific attitude. There are several ways of developing that faculty in relation to statistics. A key practice is the habitual careful consideration of the data before any statistics is undertaken. A visual appraisal of a graphical image is of paramount importance here, and the book is profusely illustrated with them: almost every dataset discussed or analysed is depicted. Another essential is developing an understanding of the exact meaning of the results of statistical operations. Finally, practitioners need the experience of both guided and unsupervised consideration of many examples of relevant datasets. Drivers don't need too many details of how the car works, but they do need the Highway Code, a road map, some driving lessons and as much practice as they can get.

The book is divided into quite short sections, each dealing with a single topic, headed by a 'Key points' box. Most sections are terminated by 'Notes and further reading' with useful references for those wishing to pursue topics in more detail. The sections are as far as possible self-contained, but are extensively cross-referenced. The book can therefore be used either in a systemic way by reading the sections sequentially, or as a quick reference by going directly to the topic of interest. Every statistical method and application covered has at least one example where the results are analysed in detail. This enables readers to emulate this analysis on their own examples. The statistical results on these examples have been cross-checked by at least two different statistics packages. All of the datasets used in examples are available for download, so that readers can compare the output of their own favourite statistical package with that shown in the book and thus verify that they are entering data correctly.

Statistics is a huge subject, and a problem with writing a book such as this is knowing where to stop. We have concentrated on providing a selection of techniques and applications that will satisfy the needs of most analytical chemists most of the time, and we make no apology for omitting any mention of the numerous other interesting methods and applications. Statisticians may be surprised at the relative emphasis placed on different topics. We have simply used heavier weighting on the topics that experience has told us that analytical chemists have most difficulty with. The book is cast in two parts, a technique-based approach followed by an application-based approach. This engenders a certain amount of overlap and duplication. Analytical chemists are thereby encouraged to create their own overview of the subject and see for themselves the relationship between tasks and techniques.

Statisticians will also notice that we use the 'p-value' approach to significance testing. This was adopted after careful consideration of the needs of analytical chemists. It greatly improves the transparency of significance testing so long as the exact meaning of the p-value is borne in mind, and we stress that meaning repeatedly. The alternative approaches tend to cause more difficulty. Experience has shown that analytical chemists find the somewhat convoluted logic of using statistical tables confusing and hard to remember. The confidence interval approach is simple but almost universally misunderstood among non-statisticians. Both of these methods also have the disadvantage of engendering the idea that the significance test can validly dichotomise reality — once you have set a level of confidence the test tells you 'yes' or 'no'. This tempts practitioners to use statistics to replace judgement rather than to assist it.

Finally, here are ten basic rules for analytical chemists undertaking a statistical analysis.

1. If you can, plan the experiment or data collection before you start the practical work. Make sure that it will have sufficient statistical power for your needs. Ensure that the data collection is randomised appropriately, so that any inferences drawn will be valid.
2. Make sure that you know how to enter the data correctly into your statistical software. After you have entered it, print it out for checking.
3. Examine the data as one or more graphical displays. This will often tell you all that you need to know. In addition it will tell you if your dataset is unlikely to conform to the statistical model that underlies the statistical test that you are proposing to use. Important features to look out for are: suspect data; lack of fit to linear calibrations; and uneven variance in regression and analysis of variance.
4. Select the correct statistical test, e.g., one-tailed or two-tailed, one sample, two sample or paired.
5. Make sure that you know exactly what the statistical output means, especially the $p-$value associated with a test of significance.
6. Be careful how you express the outcome in words. Avoid attributing probabilities to hypotheses (unless you are making a Bayesian analysis — not within the scope of this book).

7. Report the magnitude of an effect as well as its significance. Distinguish between effects that are statistically significant and those that are practically important.
8. After a regression always make plots of the residuals against the predictor variable. This will give you valuable information about lack of fit and uneven variance. It is sometimes useful to make other plots of the residuals, e.g., as a time series to detect drift in the measurement process.
9. If in doubt, ask a statistician.
10. Have fun!

Michael Thompson & Philip J Lowthian
Birkbeck University of London, UK
May 2010

* * * * * *

Data files used in the book can be downloaded from http://www.icpress.co.uk/chemistry/p739.html

Contents

PART 1
Statistics

Chapter 1

Preliminaries

This chapter sets the scene for statistical thinking, covering variation in measurement results and the properties of objects, and its graphical representation. The basis for statistical inference derived from analytical data is associated with the probability of obtaining the observed results under the assumption of appropriate hypotheses.

1.1 Measurement Variation

Key points

— Variation is inherent in results of measurements.

— We must avoid excessive rounding to draw valid conclusions about the magnitude of variation.

The results of replicated measurements vary. If we measure the same thing repeatedly, we get a different result each time. For example, if we measured the proportion of sodium in a finely powdered rock, we might get results such as 2.335, 2.281, 2.327, 2.308, 2.311, 2.264, 2.299, 2.295 per cent mass fraction (%). This variation is not the outcome of carelessness, but simply caused by the uncontrolled variation in the activities that comprise the measurement, which is often a complex multistage procedure in chemical measurement. Sometimes it may appear that results of repeated measurements are identical, but this is always a false impression, brought about by a limited digit resolution available in the instruments used to make the measurement or by excessive rounding by the person recording or reporting the data. If the above data are rounded to two significant figures, they all turn into an identical 2.3%, which tells us nothing about the magnitude of