

0056782

Design and Analysis in Chemical Research

Edited by

ROY L. TRANTER

Statistics and Data Evaluation Manager
QA Compliance Group
Glaxo Wellcome Operations
County Durham, UK



Sheffield
Academic Press



CRC Press

First published 2000
Copyright © 2000 Sheffield Academic Press

Published by
Sheffield Academic Press Ltd
Mansion House, 19 Kingfield Road
Sheffield S11 9AS, England

ISBN 1-85075-994-4

Published in the U.S.A. and Canada (only) by
CRC Press LLC
2000 Corporate Blvd., N.W.
Boca Raton, FL 33431, U.S.A.
Orders from the U.S.A. and Canada (only) to CRC Press LLC

U.S.A. and Canada only:
ISBN 0-8493-9746-4

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying or otherwise, without the prior permission of the copyright owner.

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

Printed on acid-free paper in Great Britain by
Bookcraft Ltd, Midsomer Norton, Bath

British Library Cataloguing-in-Publication Data:

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data:

Design and analysis in chemical research / edited by Roy Tranter.

p. cm. -- (Sheffield analytical chemistry : v. 3)

Includes bibliographical references and index.

ISBN 0-8493-9746-4 (alk. paper)

1. Chemistry, Analytic --Statistical methods. 2. Experimental design. I. Tranter, Roy. II. Series.

QD75.4.S8D48 1999

543'.07'2--dc21

99-28562
CIP

Design and Analysis in Chemical Research

Sheffield Analytical Chemistry

Series Editors: J.M. Chalmers and R.N. Ibbett

A series which presents the current state of the art of chosen sectors of analytical chemistry. Written at professional and reference level, it is directed at analytical chemists, environmental scientists, food scientists, pharmaceutical scientists, earth scientists, petrochemists and polymer chemists. Each volume in the series provides an accessible source of information on the essential principles, instrumentation, methodology and applications of a particular analytical technique.

Titles in the Series:

Inductively Coupled Plasma Spectrometry and its Applications

Edited by S.J. Hill

Extraction Methods in Organic Analysis

Edited by A.J. Handley

Design and Analysis in Chemical Research

Edited by R.L. Tranter

Spectroscopy in Process Analysis

Edited by J.M. Chalmers

Preface

Within the chemical sciences, statistics has the reputation of being hard and usable only by mathematicians or masochists. It also has the reputation of either telling us what we already know or making predictions that are wrong. So why should we bother with it?

Both of these reputations are undeserved and often stem from a dry theoretical statistics course or from experience of the statistics used publicly for essentially political aims. But 'real' statistics is quite different. It is the aim of this book to show that it is essentially an extension of the logical processes used by chemists every day, and that its use can, and does, bring greater understanding of problems more quickly and easily than the purely intuitive or "let's try and see" approaches.

For this we must be careful to distinguish between the tools we use to make the statistical calculations—the equations, algorithms and software—and the thought processes that allow us to decide which is the best tool to use and which is the best method for interpreting the results of its use. The latter is best described as *statistical thinking*. It encompasses the tools but extends the context to include awareness and appreciation of the sciences of measurement, experimentation and logic. It is the philosophy of rational data analysis and interpretation.

Statistics is a mathematical subject and it does involve equations—we cannot get away from this—but the equations are generally no more difficult than those routinely used in spectroscopy, kinetics, structure-activity relationships, molecular modelling or any other chemical system requiring computation. Many of the statistics equations (and concepts) are much simpler! What is different about statistics is that it is all about handling variability and uncertainty. Most chemists are more comfortable with certainty and, for many, the concept of error is associated with mistakes and poor work.

Simply accepting that all measurements have some uncertainty associated with them is the first, major step in coming to terms with statistics. The second is accepting that uncertainty can be measured and handled in a quantitative way. Once we can get a handle on uncertainty, we can control it. Although it is not possible, in any practical sense, to remove uncertainty, we can certainly reduce it and its effects, and so increase our confidence in the chemical truths that we discover.

Statistics is a broad subject and it has very wide application in the sciences, engineering, financial and behavioural arenas. It can be presented in many ways, all good in their own contexts. Here we have chosen to concen-

trate on principles and interpretation rather than on formal derivation and proof. References are given for those wanting to get into the latter. All we need to be aware of here is that all the methods we describe are well established, well documented and widely accepted. This does not mean that they give the "truth"—only that their properties are well understood and that they will give consistent and interpretable results when used appropriately.

You are not entering an equation-free zone. Far from it. But the equations needed to understand and/or to implement a particular tool are given with explanation and interpretation to help you come to terms with them.

One of the best ways of understanding how or where to use statistics is through applications. We have included many examples of actual or possible use from a wide area. The context of research chemistry is used throughout, but many of the good, easy-to-explain examples come from analytical and process chemistry, where quantitative measurement and interpretation are the norm. However, remember our objective of establishing principles. The principles in these examples apply to all branches of research chemistry, including organic and inorganic synthesis and molecular design, as well as the more obvious topics of physical chemistry and chemical physics.

In a book of this size, it is impossible to cover the whole of the vast subjects of statistics and chemometrics. So we have chosen to cover the basic statistical methods that underpin the *statistical thinking* approach. These are chapters 1-8. Chapters 9-13 describe the tools that are frequently used in chemical situations where a quantitative model is needed to describe or test a relationship between variables. Chapter 2 focuses on data quality, as we can have no confidence in statistical results if we have no confidence in the data, no matter how impressive the calculations might be.

These days, the computation of statistics is a trivial task. However, the easy availability of good hardware and software does mean that it is very much easier to do statistical computation without actually understanding what is being done. The corollary is that you have more time to develop that understanding and to consider the interpretation of the results that you have calculated. If you do get stuck or want advice on how to get away from the 'black box' approach, there are many quite friendly and helpful statisticians out there.

Finally to the authors. They are all people well known and well respected in their areas. Some are professionally trained statisticians and some are chemists with a deep understanding and appreciation of statistics. All are practical users of statistics and have wide experience of using statistics and the principles of *statistical thinking* in many areas of chemistry. I am greatly indebted to them for the time and effort they have put in to writing their chapters, and for the patience they have shown to me as editor of this volume. You will find much to appreciate in their work. Enjoy it and apply it!

Roy Tranter

Contributors

- Dr M. Robert Alecio** Director, Positive Probability Limited, 9 Church Street, Isleham, Ely, Cambridgeshire CB7 5RX, UK
Email: Robert@Alecio.freeserve.co.uk
- Mr Anthony G. Ferrige** Director, Positive Probability Limited, 9 Church Street, Isleham, Ely, Cambridgeshire CB7 5RX, UK
Email: gfv63@dial.pipex.com
- Marion Gerson** Director, Centre for Quality Engineering, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK
Email: m.e.gerson@ncl.ac.uk
- Mrs Sonya Godbert** Senior Consultant Statistician, Statistical Services, Glaxo Wellcome Research and Development, Park Road, Ware SG12 0DP, UK
- Professor Theodora Kourti** Department of Chemical Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada
Email: kourtit@mcmaster.ca
- Professor Olav M. Kvalheim** Department of Chemistry, The University of Bergen, Allegate 41, N-5007 Bergen, Norway
Email: olav.kvalheim@kj.uib.no
- Dr Ivan Langhans** CQ Consultancy, Kapeldreef 60, B-3001 Heverlee, Belgium
Email: CQ@CQConsultancy.be
- Dr Willem Melssen** Department of Analytical Chemistry, Katholieke Universiteit Nijmegen, Postbus 9010, 6500 GL Nijmegen, The Netherlands
Email: willem@sci.kun.nl

Dr Max A. Porter

Statistics Manager, Glaxo Wellcome UK
International Actives Supply, North Lonsdale
Road, Ulverston, Cumbria LA12 9DR, UK
Email: map41247@GlaxoWellcome.co.uk

Dr John M. Thompson

School of Mathematics and Statistics, Uni-
versity of Birmingham, Edgbaston, Birming-
ham B15 2TT, UK
Email: dr.jmthompson@cwcom.net
jmt@for.mat.bham.ac.uk

Dr Roy L. Tranter

Statistics and Data Evaluation Manager, QA
Compliance, Glaxo Wellcome UK Interna-
tional Product Supply, Harmire Road, Bar-
nard Castle, County Durham DL12 8DT, UK
Email: rlt48033@GlaxoWellcome.co.uk

Contents

1 Statistical thinking—The benefits and problems of a statistical approach	1
M. PORTER	
1.1 Introduction	1
1.2 Where to go	1
1.3 What is statistical thinking?	2
1.3.1 Statistics	2
1.3.2 Building in variability—Why all processes are subject to variability	3
1.3.3 A model of the inductive/deductive processes	4
1.3.4 Mathematical and statistical models	7
1.4 Types and causes of variability	7
1.4.1 Common and special causes	7
1.4.2 Bias, systematic and random variation	8
1.4.3 Example—Sampling from and assaying a bulk chemical	9
1.4.4 Impacts of variability on decision making and the design of investigations	10
1.4.5 Sampling—Why and how?	11
1.5 Probability	12
1.5.1 Measuring risk and uncertainty	12
1.5.2 Rules of probability	12
1.5.3 Random variables, distributions and statistics	14
1.5.4 Practical and statistical significance	16
1.6 Carrying out a statistical investigation	19
1.6.1 Steps in an investigation	19
1.6.2 Prospective and retrospective studies	23
1.6.3 Designing investigations	24
1.6.4 Pilot study	25
1.6.5 Validation and good statistical practice	26
1.6.6 Benefits and disadvantages	26
1.7 Skills, experts and systems	29
1.7.1 When and how to get help	29
1.7.2 What you can do for yourself	29
1.7.3 Local experts	30
1.7.4 What a statistician can contribute	31
1.7.5 Using statistical software	32
References	32
2 Essentials of data gathering and data description	34
R. TRANTER	
2.1 Introduction	34
2.2 Where to go	35
2.3 The data cycle	35

2.4	Data planning and design	36
2.4.1	Measurement systems	36
2.4.2	Experiment design	38
2.4.3	Randomisation	39
2.4.4	Data collection	41
2.4.5	Digitisation	42
2.4.6	Recording and reporting numbers	44
2.4.7	Rounding	46
2.4.8	Significant figures	47
2.5	Data description	48
2.5.1	Simple numerical checks	48
2.5.2	Trend plots and control charts	49
2.5.3	Scatter plots	52
2.5.4	Box and whisker plots	53
2.5.5	Cusum plots	54
2.5.6	Histograms and data distributions	58
2.5.7	Normality testing	60
2.5.8	Outliers and discordant values	61
2.5.9	Studentised range test for a single discordant value	64
2.5.10	Grubb's test for a single discordant value	65
2.5.11	Dixon's test for a single discordant value	65
2.6	Data preprocessing	66
2.6.1	Smoothing—moving average, Savitsky-Golay, EWMA	66
2.6.2	Integration and differentiation	70
2.6.3	Principal component analysis	75
2.6.4	Other transformations	81
	Bibliography	82

3 Sampling 85

J. THOMPSON

3.1	Introduction	85
3.2	Where to go	85
3.3	What is sampling?	86
3.4	Sampling—The Pandora's box of chemical/biological research/development	87
3.4.1	The peanut problem	88
3.4.2	Risks in diagnosis	89
3.4.3	Whose point of view?	90
3.5	Aims and objectives of sampling	91
3.5.1	Observational studies	91
3.5.2	Invasive, noninvasive, remote and indirect sampling	92
3.5.3	Sampling for process/quality control/improvement and for environmental regulation	93
3.5.4	What kinds of physical sampling can be done?	95
3.5.5	Sampling strategies	98
3.5.6	Statistical/chemometric aspects	99
3.6	Statistical sampling strategies	100
3.6.1	Random sampling	100
3.6.2	Systematic sampling	102
3.6.3	Stratified sampling	102
3.6.4	Sequential sampling	102

3.7	Sample size estimation using the concept of the power of a statistical test	103
3.7.1	Power and risks	103
3.7.2	Limit of detection and limit of quantitation	104
3.7.3	Other points	106
3.7.4	Software implementations	106
3.8	Sampling in the context of deterministic versus probabilistic assessment of compliance with a standard or threshold—use of the receiver operating characteristic (ROC) curve	107
3.9	Problems associated with behaviour of granular and other materials—their effects on designing sampling schemes and on estimating sampling reliability	109
3.10	Sampling for process control and quality management	109
3.11	Economic aspects of sampling designs	109
	Bibliography and References	110
4	Interpreting results	113
	M. GERSON	
4.1	Introduction	113
4.2	Where to go	113
4.3	The objectives of experimentation and data collection	114
4.4	Setting up a known system on which to experiment	116
4.4.1	Important assumptions about the ϵ_j term	117
4.5	Experimenting on the known system	119
4.5.1	Estimating the difference between two measurements	119
4.5.2	A confidence interval for the effect of changing the temperature	121
4.5.3	A revised confidence interval for temperature effect	122
4.5.4	Confidence intervals for variability—Standard deviation	125
4.5.5	Confidence intervals for variability—Ratio of two variances	127
4.5.6	Who needs confidence intervals?	130
4.5.7	Single-sided and double-sided intervals	131
4.6	Deciding on the size of a simple comparative experiment	132
4.6.1	A simple approach	132
4.6.2	A more general method	133
4.7	Reducing the amount of work to be done	135
4.8	Hypothesis testing and significance levels	136
4.9	Some more applications of confidence intervals	138
4.9.1	Equivalence studies	138
4.9.2	Paired comparisons	140
4.9.3	Analysis of variance for estimating different sources of variability	142
4.10	Appendix	144
4.10.1	Calculation of approximately Normally distributed random numbers	144
	References	144
5	Robust, resistant and nonparametric methods	145
	J. THOMPSON	
5.1	Introduction	145
5.2	Where to go	146
5.3	Some simple and useful concepts	147
5.4	Looking at continuous measurements variables	148

5.4.1	Some initial comments about the shapes of data distributions for continuous variables	148
5.4.2	Estimating the location of a single set of data using arithmetic and geometric means	149
5.4.3	Estimating the spread of a single set of continuous data	153
5.4.4	Estimating confidence intervals for location estimators for a single set of data	157
5.4.5	Estimating confidence intervals around spread estimates for a single set of data	158
5.5	Outlier tests for single sets of data	161
5.5.1	Exploratory data analysis methods for outlier detection based on the fourth spread	161
5.5.2	Exploring the shape of a single set of data	161
5.6	Robust and resistant methods in evaluating data transformations and the value of transformation	166
5.7	Randomness in a data set	168
5.8	Nonparametric methods for comparing locations of paired data sets	169
5.9	Nonparametric methods for comparison of two sets of unpaired data	173
5.10	Nonparametric goodness-of-fit tests to specific distributions	174
5.11	Nonparametric comparisons of the effects of one factor on more than two sets of data	175
5.11.1	Kruskal–Wallis one-way analysis of variance (ANOVA) by ranks, including multiple comparisons methods	175
5.11.2	Exploratory one-way ANOVA	176
5.11.3	Cross-classified ANOVA designs—nonparametric and resistant/robust methods for two-way ANOVA and more complex designs	177
5.12	Estimating functional relationships or making paired comparisons with robust and nonparametric regression methods for two variables	182
5.12.1	Tukey's three-group resistant line regression	183
5.12.2	Theil–Kendall regression using the median of pairwise slopes	184
5.12.3	Hettmansperger's rank regression methods	184
5.12.4	Rousseeuw's least median of squares (LMS) and least trimmed squares (LTS) regression methods	185
5.12.5	Tukey's biweight regression method	185
	References	186
6	Experiment design—Identifying factors that affect responses	188
	S. GODBERT	
6.1	What is design of experiments?	188
6.2	Where to go	190
6.3	Terminology	190
6.4	Getting started	194
6.5	Two-level designs	197
6.5.1	Full factorial	197
6.5.2	Example: Determining the best storage conditions using a full factorial	197
6.5.3	Blocking	203
6.5.4	Fractional factorial	206
6.6	Confounding	209
6.6.1	Resolution	209

6.6.2	Example: Crystallisation study to determine the factors affecting particle size uniformity using a resolution III screening design	210
6.6.3	Irregular fraction designs	213
6.6.4	Plackett–Burman designs	213
6.6.5	Taguchi designs	214
6.6.6	D-Optimal designs	215
6.6.7	Robustness designs	216
6.6.8	Centre points	216
6.6.9	Example: Determining robustness using a fractional factorial	216
6.6.10	Example: Assessing process deviation ranges using a highly fractionated factorial design	221
6.6.11	Other designs	225
6.6.12	Mixed-level designs	225
6.7	Data analysis and interpretation	225
6.8	Choosing a design	227
6.9	Effect of not following the design exactly	230
6.10	Other potential problems	231
6.11	Screening designs in context	232
6.11.1	Example: Summary of DOE performed during the development of a beta-lactam	233
	Bibliography and References	236
7	Designs for response surface modelling—Quantifying the relation between factors and response	237
	I. LANGHANS	
7.1	Introduction	237
7.2	Where to go	237
7.3	The basics	238
7.3.1	What is response surface modelling?	238
7.3.2	What has this got to do with experimental design?	240
7.4	Soft modelling using polynomials	240
7.4.1	Of true models and their approximations	240
7.4.2	Relation between the objective of a study and the complexity of the polynomial	244
7.5	Designs for response surface modelling	248
7.5.1	General considerations	248
7.5.2	Central composite design	249
7.5.3	Box–Behnken designs	252
7.5.4	Choosing a design	252
7.5.5	Case study	253
7.6	Doing the experiments	253
7.7	Analysing the data	255
7.7.1	General considerations	255
7.7.2	A step-by-step look at the analysis of designed data	255
7.7.3	Case study	260
7.8	A closer look at the properties of RSM designs	265
7.8.1	Prediction error	265
7.8.2	Maximum prediction error	265
7.8.3	Average prediction error	266
7.8.4	Uniform precision	266

7.8.5	Rotatability	267
7.8.6	Estimation of the individual effects	268
7.8.7	Robustness towards missing or 'wild' responses—replicated axial designs	269
7.8.8	Blocking	269
7.9	A glimpse of what else is out there	270
7.9.1	Optimal designs	271
7.9.2	Small composite designs	272
7.9.3	Constrained regions	272
7.9.4	Mixture problems	272
7.9.5	Categorical variables	274
7.9.6	Principal properties of multivariate designs	274
7.9.7	Space-filling designs	275
7.9.8	Power considerations—what are the smallest effects you can estimate?	275
7.9.9	Robustness modelling	276
7.9.10	Multiresponse optimisation	276
7.10	Further reading	277
7.11	Appendix: a case study in the use of response surface modelling	277
	Bibliography	278
8	Analysis of Variance. Understanding and modelling variability	279
	M. PORTER	
8.1	Introduction	279
8.2	Where to go	280
8.3	Preliminaries	281
8.3.1	Variables and factors	281
8.3.2	The types and causes of variability	281
8.3.3	Properties of estimates	282
8.4	Variability in data	283
8.4.1	A model for total variability in a data set	283
8.4.2	Models, estimates and the analysis of variance	285
8.4.3	Alternatives to least squares and the analysis of variance	288
8.5	Modelling variability in simple linear regression	291
8.5.1	Residuals and estimation	292
8.5.2	Analysis of variance	292
8.5.3	Assessing the appropriateness of the model	294
8.5.4	Variability in the predictor	295
8.6	One-way or fully randomised ANOVA	296
8.6.1	Fixed and random effects	296
8.6.2	Estimation of fixed effects	297
8.6.3	Analysis of variance	298
8.6.4	Estimation of random effects	299
8.6.5	Allocation of experimental material	301
8.7	Two-way, two-factor or randomised blocks ANOVA	301
8.7.1	Blocking	301
8.7.2	Crossed classifications	302
8.7.3	Nested classifications	304
8.8	A general approach	305
8.8.1	A general model and analysis	306
8.8.2	Classification of model components	307
8.8.3	Generalised linear models	308

8.8.4	Software for general linear models	308
8.9	Examples of two-factor analyses of variance	308
8.9.1	Randomised blocks design	309
8.9.2	Replicated two-factor design	310
8.9.3	Nested Factors	311
	References	313
9	Optimisation and control	314
	T. KOURTI	
9.1	Introduction	314
9.2	Terminology	314
9.3	Where to go	315
9.4	Why process control and optimisation?	316
9.4.1	Process control	316
9.4.2	Optimisation	318
9.5	Statistical process control (SPC)	318
9.5.1	Univariate Shewhart charts	319
9.5.2	A general model of Shewhart charts	321
9.5.3	Univariate Cusum charts	323
9.5.4	Univariate EWMA charts	325
9.5.5	Multivariate charts for statistical quality control	327
9.5.6	Hotelling's T^2 and chi-squared multivariate charts	328
9.5.7	Multivariate Cusum charts	330
9.5.8	Multivariate EWMA	332
9.5.9	Multivariate control charts based on latent variables	333
9.5.10	Principal component analysis (PCA) for multivariate monitoring	334
9.5.11	Partial least squares (PLS) for multivariate monitoring	336
9.5.12	Control charts based on latent variables	337
9.5.13	Fault diagnosis	339
9.5.14	Multiway data	340
9.5.15	Multiblock data	340
9.5.16	Issues in latent variable analysis and SPC	341
9.5.17	Other applications of multivariate charts	344
9.6	Optimisation of processes	345
9.6.1	General ideas	345
9.6.2	Optimise functions of many variables by changing one variable at a time	348
9.6.3	Multiresponse optimisation	349
9.6.4	Procedure for optimisation with empirical models	350
9.6.5	Sequential methods	354
9.6.6	Process optimisation using historical data	359
9.6.7	Product design with latent variables	360
	References	361
10	Grouping data together—Cluster analysis and pattern recognition	365
	W. MELSEN	
10.1	Introduction	365
10.2	Where to go	365
10.2.1	Visualisation of data	365
10.2.2	Similarity and finding clusters and groups in the data	366

10.2.3 Using known groupings to predict membership of new samples	367
10.2.4 Transforming data for better models	367
10.2.5 A brief route through the chapter	368
10.3 Visualisation and mapping	369
10.3.1 Principal component analysis	369
10.3.2 Nonlinear mapping (NLM)	373
10.3.3 Kohonen self-organising feature map neural network	376
10.3.4 Parallel coordinates	379
10.4 Clustering of multivariate measurements	381
10.4.1 Single, average and complete linkage	383
10.4.2 Ward's clustering method	386
10.4.3 Forgy's clustering method	388
10.5 Classification	391
10.5.1 Linear discriminant analysis	391
10.5.2 Soft independent modelling of class analogy	396
10.5.3 Multilayer feed-forward neural networks	400
10.5.4 Validation of the classification model	405
10.6 Appendix A. Transformation and scaling of the data	408
10.6.1 No transformation or scaling	408
10.6.2 Range scaling	409
10.6.3 Mean centring	409
10.6.4 Autoscaling	409
10.6.5 Other transformations	410
10.6.6 Example	410
10.7 Appendix B. Measures of (dis)similarity	413
10.7.1 Minkowski distance	413
10.7.2 Mahalanobis distance	415
10.7.3 Correlation coefficient	416
10.7.4 Which to use?	416
Bibliography	418
11 Linear regression	421
R. TRANTER	
11.1 Introduction	421
11.2 Where to go	422
11.3 Some terminology	422
11.4 Cause and effect	424
11.4.1 General	424
11.4.2 Relationships between variables	426
11.5 Correlation, covariance, r and R^2	428
11.6 Regression	432
11.7 Simple linear regression	433
11.7.1 Linear models	433
11.7.2 Linear least squares	434
11.7.3 Assumptions	435
11.7.4 Checking the results	436
11.7.5 An example	437
11.7.6 Standard errors of estimates	439
11.7.7 The ANOVA table	441
11.7.8 Lack of fit	443