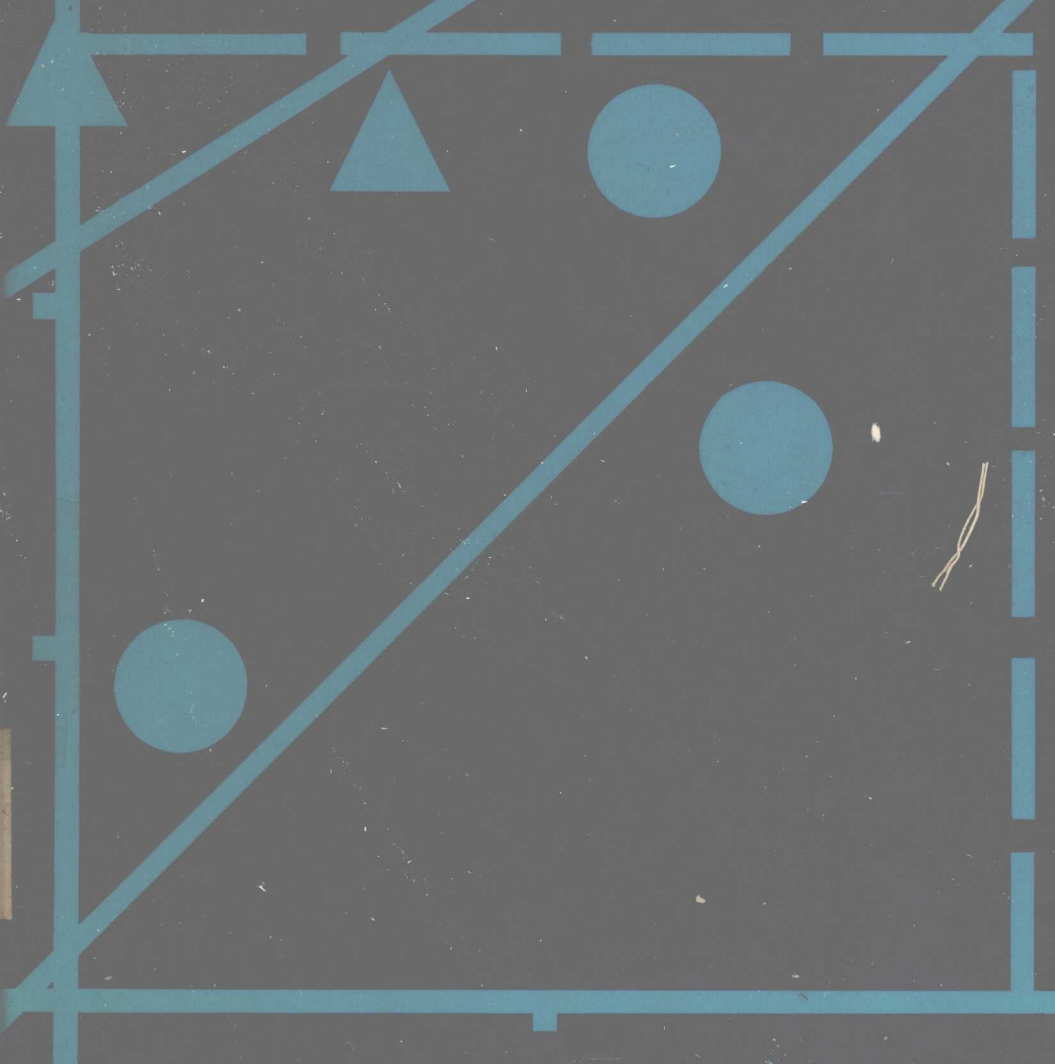
STATISTICS FOR ANALYTICAL CHEMISTS

ROLAND CAULCUTT
RICHARD BODDY



STATISTICS FOR ANALYTICAL CHEMISTS

Roland Caulcutt
and
Richard Boddy
Statistics for Industry (UK) Ltd

LONDON NEW YORK
CHAPMAN AND HALL

First published 1983 by
Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE
Published in the USA by
Chapman and Hall
733 Third Avenue, New York NY10017

Printed in Great Britain by
J. W. Arrowsmith Ltd, Bristol

ISBN 0 412 23730 X

British Library Cataloguing in Publication Data

Caulcutt, Roland
Statistics for analytical chemists.
1. Mathematical statistics
I. Title II. Boddy, Richard
519.5′0246541 QA276

ISBN 0-412-23730-X

Library of Congress Cataloging in Publication Data

Caulcutt, R. (Roland)
Statistics for analytical chemists.

Bibliography: p.
Includes index.
1. Chemistry, Analytic—Statistical methods.
I. Boddy, Richard, 1939- . II. Title.
QD75.4.S8C38 1983 519.5′024541 82-23637
ISBN 0-412-23730-X

Preface

This book is based upon material originally prepared for courses run by Statistics for Industry (UK) Ltd. Over a number of years written material was repeatedly extended and refined as the authors developed a clearer image of the statistical needs of the course members. This development was retarded, however, by the adoption at the outset of an assumption that proved to be false. We assumed that the majority of those scientists sharing the title 'analytical chemist', would also share common problems and have very similar needs. We were mistaken.

Experience revealed, for example, that many analytical chemists were primarily concerned with errors of calibration, whilst others had no interest whatsoever in this area, but were seeking advice on the problems of interlaboratory trials. We also found that this latter interest was shared by scientists and technologists of other disciplines, ranging from civil engineers to nuclear physicists.

It would be foolish, therefore, to suggest that this book offers ready solutions to *all* the statistical problems of *every* analytical chemist. We could reasonably claim, however, that *any* analyst with a statistical problem would benefit from adopting this text as a foundation for his introductory reading. The approach adopted throughout the book has two distinctive features, which also characterize the courses on which the book is based. This approach is both 'problem centred' and 'non-mathematical', thus enabling the reader to concentrate upon three essential elements:

(a) how to use statistical techniques,
(b) how to interpret the results of statistical analysis,
(c) how to check the assumptions underlying the statistical techniques.

Many of the statisticians and scientists who lecture on Statistics for Industry courses have made numerous suggestions during the preparation of this book. These suggestions have undoubtedly led to an improvement in the readability and the usefulness of the book. We are deeply grateful to all concerned, who are too numerous to list individually. We are particularly grateful to Tony Wilson (now retired but formerly with the Water Research Association) for helping us to appreciate the diverse problems of the analytical chemist and for offering guidance in our search for solutions.

The policy of continuous improvement adopted by Statistics for Industry has hopefully, resulted in a more readable book. On the other hand this same policy does makes great demands upon our typist, Christine Robinson, who has managed to produce a beautiful typescript from the mis-spelt jottings of one author and the illegible notes of the other.

Statistics for Industry (UK) Ltd runs many courses in applied statistics at a variety of venues throughout each year. All of these courses, or *workshops* as they are more correctly described, are intended for scientists and technologists in the chemical and allied industries. The current list of workshops includes:

Introduction and Significance Testing
Statistics in Research and Development
Statistics for Analytical Chemists
Design of Experiments
Statistical Quality Control

In response to the needs of our customers new workshops are introduced every year. Full details of all courses and consultancy services can be obtained from:

The Conference Secretary,
Statistics for Industry (UK) Ltd,
14 Kirkgate,
Knaresborough,
North Yorkshire HG5 8AD.

Tel: (0423) 865955

Contents

1
What is statistics?

'Statistics' is a word with many meanings. We are doubtful that a simple definition can be found which will be entirely satisfactory, even in such a specialized text. Let us therefore explore several popular definitions of statistics before selecting one for further consideration. These definitions will overlap to some extent but each emphasizes a point of some importance.

Definition 1: 'Statistics is a branch of mathematics'.

Definition 2: 'Statistics is a set of techniques which can be used to prove almost anything'.

Definition 3: 'Statistics is an extremely dull subject and its application involves very tedious calculations'.

Definition 4: 'Statistics is a body of knowledge which can be of use to anyone who has taken a sample'.

The first definition contains some truth. Many mathematicians study certain aspects of statistics and in doing so consider that they are exploring just another branch of mathematics. It is also true that a statistician needs a good grounding in mathematics especially if he or she is to engage in basic research. For the *applied statistician* however, an ability to communicate with his client (e.g. the analytical chemist) is just as important as mathematical expertise whilst the client for his part may be able to avoid *all* contact with mathematics. You may be very sceptical of this last statement. Perhaps you are a chemist who does not have access to an experienced applied statistician. Perhaps you have attempted to teach yourself statistics from a book which was very mathematical. Perhaps you were taught statistics alongside chemistry without being shown the connections between the two subjects. We assert nonetheless that the scientist or technologist can make use of statistics even if he has very little knowledge of mathematics. It is certainly true that you will find very little mathematics in this book.

The second definition also contains some truth. Clearly it is possible to deceive the unwary by presenting only a selected part of a set of data, or by reporting only those conclusions which support one's prejudices. It is also possible to arrive at invalid conclusions by using statistical techniques inappropriately. Because of this possibility we include in later chapters a discussion of the *assumptions* underlying the recommended statistical techniques. Armed with a knowledge of these assumptions the scientist can safely make use of statistics and will be in a strong position to detect the invalid conclusions of others.

The third definition may have been partly true of some applications of statistics many years ago. With the advent of electronic digital computers, and especially with the recent proliferation of microcomputers, the drudgery of calculation has been eliminated. As for dullness of statistics, the reader must judge for himself, but the authors will be very disappointed if they fail to communicate the excitement of adapting statistical techniques to the particular problems of the analytical chemist.

The last of the four definitions is more useful as a foundation for introducing the basic ideas of applied statistics. Clearly the analytical chemist is well aware of the need to take *samples*. If however he is prepared to view the sampling process through the eyes of the statistician then the statistical techniques may seem more reasonable. We will explore two situations in which an analytical chemist has taken a sample. The first is not in a laboratory environment but it will serve to illustrate an important point.

EXAMPLE 1.1

Circulation of *The Analyst* in 1980 was approximately 9000 copies per month. Many copies are read by more than one person, of course, and the editor of *The Analyst* would like to *estimate* the average readership per copy. From the computer file of subscribers he selects 100 addresses at random and despatches a questionnaire to each. All of the questionnaires are returned and the editor calculates the average readership of the 100 copies to be 2.18. Is it safe for the editor to claim that the average readership is more than 2 readers per copy?

■■■

This very simple example illustrates several important points which would not be revealed so clearly in a more complex situation. The sample consists of 100 copies of *The Analyst* and the sample average is 2.18 readers per copy. The editor wishes to make a statement about all 9000 copies and a statistician would refer to this larger group as the *population*.

> A sample is simply a small group taken from a larger group about which we wish to draw a conclusion. This larger group is known as the population.

Clearly the validity of the conclusion will depend upon *how* the sample is drawn from the population. In this example the editor selected the 100 addresses at *random*. He was using a method known as *random sampling* which ensures that each member of the population has the same chance of being included in the sample. In many situations it is not possible to take a random sample as we shall see later.

Concerning the relationship between the sample and the population, several points need to be stressed:

(a) If the editor were to repeat the exercise he would almost certainly get a different sample and a different value for the sample average.
(b) Since average readership varies from sample to sample, it is rather unlikely that the editor's sample average (2.18) will be exactly equal to the population average (i.e. the average readership of all 9000 copies).
(c) Presumably the population average will be *close to* the sample average, but can the editor be sure that the population average will be greater than 2 readers per copy? (This question will be answered in a later chapter.)

EXAMPLE 1.2

An analytical chemist wishing to evaluate a new method carries out a preliminary investigation in which he makes six replicate determinations of the copper content of a solution which is known to have a copper content of 60.0 p.p.m. Each determination requires the preparation of a 10 ml sample and the determinations are:

58.2 61.0 56.6 61.5 53.8 56.9

Can the analyst draw any conclusions from this limited amount of data?

The analyst would have been delighted if all six determinations had been exactly equal to 60.0. He expects, however, to find error in any determination and is not surprised that there is variability amongst the results. He is a little disappointed that the average determination (58.00) is not closer to the true concentration of 60.0 p.p.m. but he suspects that the difference might have been less if he had taken a larger number of determinations.

■ ■ ■

Clearly the analyst is interested in the *bias* and *precision* of the new method but before we discuss these important concepts we will draw attention to several features of the sample and the population. (You will recall that we earlier defined a sample as being a small group taken from a larger group known as the population.)

(a) In this situation the chemist might say that he had prepared *six samples* whereas the statistician would say that there is *one sample* containing six observations. Clearly there is some scope for confusion with the chemist focusing on distinct quantities of material and the statistician concentrating on sets of numbers. Throughout this book both meanings of the word 'sample' will be used but we hope that the meaning will be obvious from the context.

(b) Perhaps the benefit to be gained by taking a statistician's view of the sample will be clearer when we talk about the population. If the sample consists of six determinations then the population must also contain determinations, but more than six of them. We could define the population as being *all* determinations that the analyst might have made on the solution using the new method. If we accept this definition then the population is infinitely large and, furthermore, only six members of the population actually exist in a material sense. With such a population how are we to ensure that every member has the same chance of being included in the sample? It would appear to be impossible.

(c) Perhaps you would prefer not to talk about the population. It is intuitively obvious, however, that we are not likely to draw a valid conclusion unless our sample is representative of the population. Furthermore many statistical techniques do refer to a population and have an underlying assumption that the sample has been selected at random from this population.

(d) The analyst is more interested in the precision of the test method than in the hypothetical population put forward by the statistician. We will see in a later chapter how an estimate of precision can be obtained by measuring the variability in repeat determinations. We will also discuss techniques for comparing the precision of two test methods. All such procedures are, however, based on certain assumptions about the population from which the sample was taken.

(e) The analyst is also interested in the possibility that the test method might be biased. The average of the six determinations (58.00 p.p.m.) is certainly not equal to the true concentration (60.0 p.p.m.) but this does not prove that the new method is biased. Perhaps the sample average would have been closer to 60.0 p.p.m. if a larger number of determinations had been made. Is it possible that the sample average would have been *equal* to 60.0 p.p.m. if *all* possible determinations had been made on the solution? Clearly we can never answer such questions with absolute certainty but we will, in a later chapter, explore a technique which helps us to decide beyond reasonable doubt whether a test method is biased in any particular situation.

2

Describing a set of data

2.1 Introduction

Before we can explore the many areas in which the practising analytical chemist makes use of statistics, we must first examine some simple techniques which can be used to summarize a set of data. If we have two or more determinations we will naturally be concerned with the scatter of these measurements around some average value. To convey an impression of this scatter, to ourselves or others, a diagram can be very helpful and we will make frequent use of simple graphical techniques throughout this book.

Scatter and average can also be expressed in quantitative terms if we are prepared to carry out certain calculations on the data. These calculations are greatly facilitated by the use of a modern pocket calculator, which reduces the drudgery and increases our confidence in the results.

We will also introduce in this chapter the *normal distribution curve* which has proved useful in analytical chemistry for describing the scatter of errors. The use of this curve will always be based on an assumption that it is applicable to the situation under investigation.

2.2 Describing a small set of data

Let us take another look at the copper content determinations from the previous chapter. You will recall that the analytical chemist had made six repeat determinations on a solution which was known to have a concentration of 60.0 p.p.m. The results were:

58.2 61.0 56.6 61.5 53.8 56.9

Even with such a small set of data a graphical representation can be useful. A suitable graph can be obtained very easily if we represent each determination by a point (or a blob) on a line. The *blob chart* (Fig. 2.1) is a plot of the six determinations.

Marked on the blob chart are the true concentration (60.00) and the mean determination (58.00) *Mean* is a word very frequently used in statistics as an

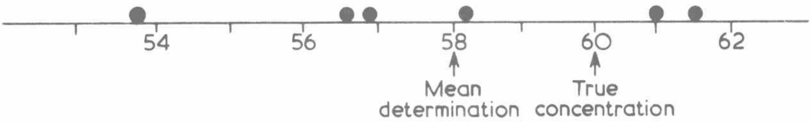


Figure 2.1 Six determinations of copper content – a blob chart

alternative to *average*. The mean determination is calculated by adding up the six determinations and then dividing the total by six. A formula for this operation is:

$$\boxed{\text{Sample mean } (\bar{x}) = \sum x/n}$$

In this formula $\sum$ is a Greek capital letter *sigma* and $\sum x$ represents the sum of the determinations, whilst n represents the number of determinations. The symbol $\bar{x}$ (pronounced x bar) is often used to represent a sample mean (if you are not familiar with sigma notation see Appendix 1. The symbols and formulae used are given in Appendix 2). It may seem pedantic to introduce a formula with Greek letters to describe the calculation of a simple average but this sigma notation will prove useful when we meet more complex calculations.

As an alternative to calculating the mean we could put the determinations into ascending order and select the middle one. This would be known as the *median* determination. With an even number of determinations we find that there isn't a middle *one* and we take the average of the middle *two*. In ascending order the six determinations are:

$$53.8 \qquad 56.6 \qquad \underline{56.9} \qquad \underline{58.2} \qquad 61.0 \qquad 61.5$$

The median determination is $(56.9 + 58.2)/2$ which is 57.55 p.p.m.

Let us return to the blob chart in Fig. 2.1. It is probably true that the main value of such a diagram is that it can give an impression of the scatter or spread or variability of the determinations. The variability of a set of repeat determinations gives an indication of the precision (or lack of precision) of the test method. It is useful, therefore, to be able to quantify the spread of a set of measurements. This can be achieved by subtracting the smallest measurement from the largest to get the *range* of the sample.

$$\boxed{\text{Sample range} = \text{highest value} - \text{lowest value}}$$

For our six determinations the sample range is 61.5 minus 53.8 which is 7.7. If the new test method were less precise the six determinations would probably have been more widely spread and the sample range would then have been greater than 7.7. If all six determinations had been identical then the sample range would have been zero.

Though the sample range is very easily calculated it is not a very reliable measure of spread because it is over-dependent on the two most extreme values in the sample. A much more useful measure of spread is the sample standard deviation or its close relative the sample variance.

Sample variance $=\sum(x-\bar{x})^2/(n-1)$ OR $(\sum x^2-n\bar{x}^2)/n-1)$

Sample standard deviation $=\sqrt{}$(sample variance)

Use of these formulae will be illustrated by calculating the variance, and then the standard deviation, of the six determinations which have been listed in the left hand column of Table 2.1.

Table 2.1 Calculation of the variance of six determinations

	Determination x	*Deviation* *from mean* $(x-\bar{x})$	*Squared* *deviation* $(x-\bar{x})^2$	
	58.2	0.2	0.04	
	61.0	3.0	9.00	
	56.6	−1.4	1.96	
	61.5	3.5	12.25	
	53.8	−4.2	17.64	
	56.9	−1.1	1.21	
Total	348.0	0.0	42.10	$=\sum(x-\bar{x})^2$
Mean	58.00			

$$\begin{aligned}\text{Sample variance} &= \textstyle\sum(x-\bar{x})^2/(n-1)\\ &= 42.10/5\\ &= 8.42\\ \text{Sample standard deviation} &= \sqrt{8.42}\\ &= 2.902\end{aligned}$$

In Table 2.1 the mean determination (58.00) has been subtracted from each individual determination to obtain the *deviations from the mean* in the second column. Note that some of the deviations are positive and some are negative whilst the column adds up to zero. Squaring the deviations gives the numbers in the third column and the total of this column (42.10) is known as the *sum of squares*. To get the sample variance we divide the sum of squares by $(n-1)$. This divisor is known as the *degrees of freedom* and we say that the sum of squares has $(n-1)$ degrees of freedom.

Variance = (sum of squares)/degrees of freedom)

The question is often asked 'Why do we divide by $(n-1)$ and not by n?' In other words 'Why has the sum of squares got $(n-1)$ degrees of freedom?' Three points need to be stressed in answer to this question:

(a) *Occasionally* one might wish to divide by n when calculating a sample variance or standard deviation. Many pocket calculators will calculate the standard deviation very rapidly offering a choice of either n or $(n-1)$ as a divisor. The buttons are usually labelled σ_n and σ_{n-1}.
(b) To avoid confusion we will *always* divide by $(n-1)$ throughout this book when calculating a sample standard deviation.
(c) As we will see later, the purpose of calculating a sample standard deviation is almost invariably to estimate a population standard deviation. It can be shown mathematically (or by repeated experiments) that the use of n as a divisor will give a sample standard deviation which tends to *underestimate* the population standard deviation, whereas the use of $(n-1)$ gives what is known as an 'unbiased estimator'.

Both the variance and the standard deviation are measures of spread. Had the determinations been more widely scattered then both the standard deviation and the variance would have had larger values. Had all six determinations been equal then both measures of spread would have been equal to zero. When, you might wonder, would we use a standard deviation rather than a variance? An important point to note is that the sample standard deviation, like the sample mean, is expressed in the *same units* as the original data. If, for example, we were interested in the weights of analytical chemists and we recorded the weights of a sample of ten such chemists in pounds, then the mean weight would be in pounds and the standard deviation would also be in pounds, whilst the variance would be in pounds squared. The standard deviation is, therefore, of more direct use to the chemist, but variances will arise from time to time throughout this book and we will even use the sum of squares as a measure of variability in certain circumstances.

Quite often it is convenient to express a standard deviation as a percentage of a known value or of an average value. If we express the sample standard deviation as a percentage of the sample mean we obtain a dimensionless measure of spread known as the coefficient of variation (CV) or relative standard deviation.

Coefficient of variation = (standard deviation/mean) × 100