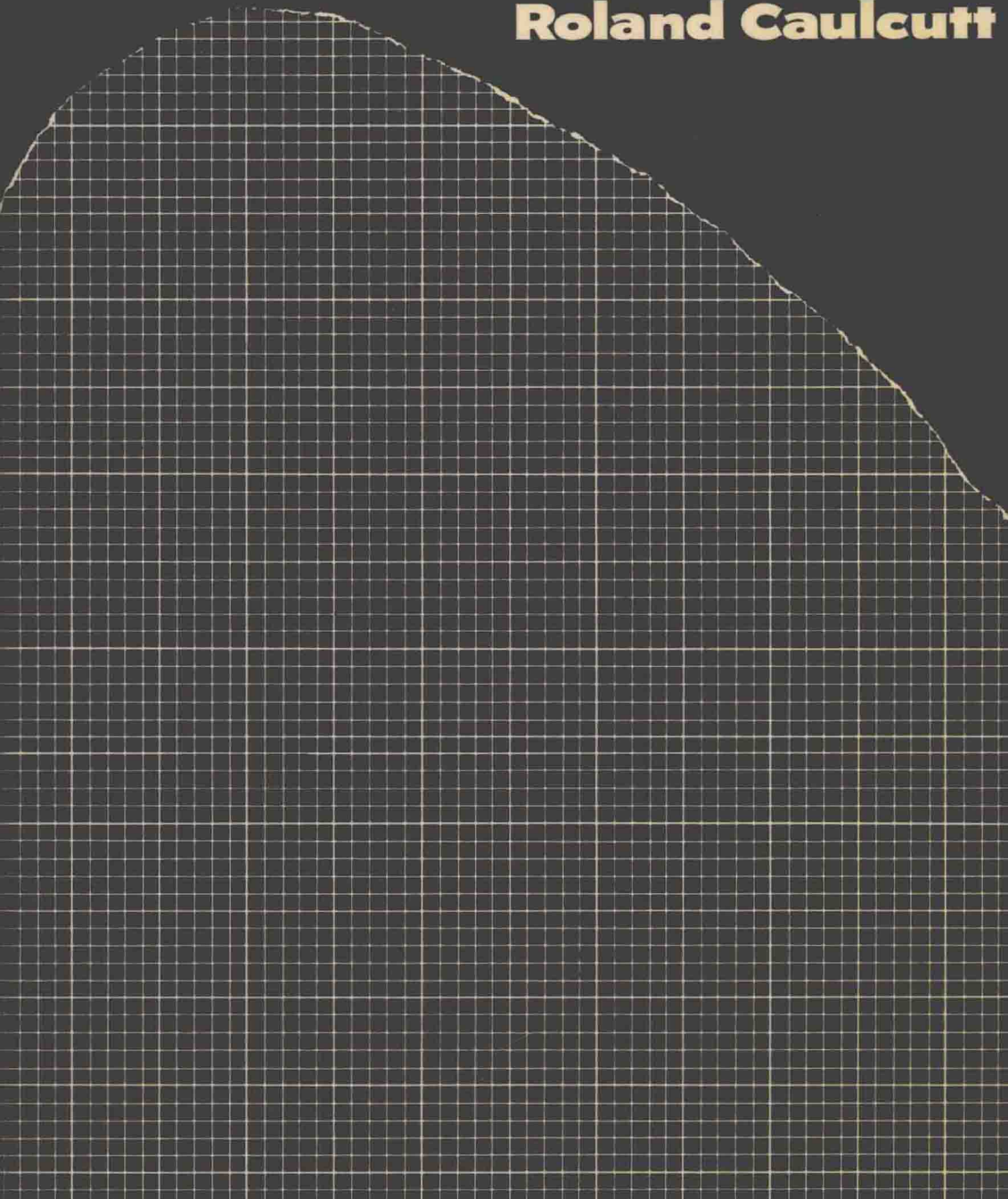
STATISTICS IN RESEARCH AND DEVELOPMENT

Roland Caulcutt



STATISTICS IN RESEARCH AND DEVELOPMENT

Roland Caulcutt
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 Richard Clay
(The Chaucer Press) Ltd, Bungay, Suffolk

Typeset in Great Britain by Keyset Composition
Colchester, Essex

ISBN 0 412 23720 2

British Library Cataloguing in Publication Data

Caulcutt, R.

Statistics in research and development.
1. Mathematical statistics
I. Title
519.5 QA276

ISBN 0-412-23720-2

Library of Congress Cataloging in Publication Data

Caulcutt, R.

Statistics in research and development.
Bibliography: p.
Includes index.
1. Statistics. I. Title.
QA276.12.C387 1982 001.4'22 82-12800

ISBN 0-412-23720-2

Preface

This book is based upon material originally prepared for courses run by Statistics for Industry (UK) Ltd. Over a period of some four or five years the notes were repeatedly extended and refined as the author developed a clearer image of the needs of the rather heterogeneous course members. It is hoped that this effort will not have been in vain and that this final version will be of use to engineers, physicists and biologists, in addition to the chemists for whom the original notes were intended.

The approach adopted throughout this book has two distinctive features, which also characterize the statistical courses on which the book is based. This approach is both 'problem centred' and 'non-mathematical', thus enabling the reader to concentrate on 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.

The spread of microcomputers into laboratories and the increasing availability of statistical computer programs has created the possibility of scientists and technologists carrying out their own statistical analysis, without reference to a statistician. It is hoped that this book will be very helpful to any scientist who attempts to 'go it alone', but the book will certainly *not* convert a scientist or technologist into a statistician. In fact the reader would be well advised to regard this as an 'introductory' book, despite the great statistical heights reached in certain chapters of Part Two. Several of the texts listed in the bibliography could be used to consolidate and to broaden any understanding gained from this book.

Many of the statisticians and scientists who lecture on Statistics for Industry courses have made numerous suggestions during the preparation of this book. I am particularly grateful to Dick Boddy, Derrick Chamberlain, John Sykes and David Arthur for their advice and support. Dr A. F. Bissell offered many constructive comments on an earlier draft of this book, for which I am very grateful. This policy of continuous improvement has, hopefully, resulted in a more readable book. It has certainly made great demands on my typist, Christine Robinson, who has managed to produce a beautiful manuscript from

a collection of mis-spelt and unpunctuated jottings. This does not absolve myself from the responsibility for any errors that remain.

Statistics for Industry (UK) Ltd run many courses in applied statistics at a variety of venues throughout the year. The courses include:

Introduction and significance testing
Statistics in research and development
Statistics for analytical chemists
Design of experiments
Statistical quality control

New courses are introduced every year and full details of all courses can be obtained from:

The Conference Secretary,
Statistics for Industry (UK) Ltd,
14 Kirkgate,
Knaresborough,
N. Yorkshire,
HG5 8AD.
Tel: (0423) 865955

Contents

Statistical Tables

PART ONE

1
What is statistics?

The purpose of this chapter is to single out the prime elements of statistics so that you will not lose sight of what is most important as you work through the detail of subsequent chapters. Certain words will be introduced which will recur throughout the book; they are written in bold letters in this chapter and each one is underlined when it first appears. Let us start with a simple definition:

> Statistics is a body of knowledge which can be of use to anyone who has taken a **sample**.

It is appropriate that the first word which has been presented in bold letters is **'sample'**, since its importance cannot be overemphasized. Many people in the chemical and allied industries occasionally find themselves in the unenviable position of needing to investigate a **population** but are constrained by the available resources to examine only a **sample** taken from the **population**. In essence a **population** is simply a large group. In some situations the **population** is a large group of people, in other situations it is a large group of inanimate objects, though it can in some cases be more abstract. Let us consider some examples of situations in which someone has taken a **sample**.

Example 1.1
A works manager wishes to **estimate** the average number of days absence due to sickness for his employees during 1977. From the files which contain details of absences of the 2300 employees his secretary selects the record cards of 50 employees. She calculates that the average number of days of sickness absence of the 50 employees during 1977 is 3.61 days.
■■■

In this example the **population** consists of 2300 employees whilst the **sample** consists of 50 employees. After the **sample** was taken there was no doubt about the *average sickness absence* of the employees in the **sample**. It is the **population** average about which we are uncertain. The works manager intends to use the **sample** average (3.61 days) as an **estimate** of the **population** average. Common sense would suggest that there are dangers to be heeded. Perhaps it is worthwhile to spell out several points, some of which may be obvious:

(a) If the secretary had selected a different set of 50 employees to include in the **sample** the sample average would almost certainly have had a different value.

(b) The **sample** average (3.61) is very unlikely to be equal to the **population** average. (More formally we could say that the **probability** of the **sample** average being equal to the **population** average is very small.)

(c) Whether or not the **sample** average is close to the **population** average will depend on whether or not the **sample** is **representative** of the **population**.

(d) No one can guarantee that a **sample** will be **representative** of the **population** from which it came. We can only hope that, by following a reputable procedure for taking the **sample**, we will end up with a **sample** which is **representative** of the **population**. It is unfortunately true, as Confucius may have pointed out, that 'He who takes a **sample** takes a risk'.

(e) One reputable procedure for taking a **sample** is known as **random sampling**. The essential feature of this method is that every member of the **population** has the same chance (or **probability**) of being included in the **sample**. The end product of **random sampling** is known as a **random sample**, and all the statistical techniques which are introduced in this book are based upon the assumption that a **random sample** has been taken.

Example 1.2

Nicoprone is manufactured by a batch production process. The plant manager is worried about the percentage impurity, which appears to be higher in recent batches than it was in batches produced some months ago. He suspects that the impurity of the final product may depend on the presence of polystyline in the resin which is one of the raw materials of the process. The supplier of the resin has agreed that the polystyline content will not exceed 1% on average and that no single bag will contain more than 2.5%. Approximately 900 bags of resin are in the warehouse at the present time. The warehouse manager takes a **sample** of 18 bags by selecting every 50th bag on the shelf. From each of the selected bags a 20 gram **sample*** of resin is taken. The determinations of polystyline content are:

1.6%	0.5%	3.1%	0.7%	0.8%	1.7%	1.4%	0.8%	1.1%
0.9%	2.4%	0.6%	2.2%	2.9%	0.3%	0.5%	1.0%	1.3%

We can easily calculate that the average polystyline content is equal to 1.32% and we notice that two of the determinations exceed 2.5%. So the **sample** average is certainly greater than the specified limit of 1.0%, but we need to consider what this average *might* have been if the warehouse manager had selected a different sample of bags, or perhaps we should ask ourselves what the average polystyline content would have been if he had sampled *all* the bags in the warehouse.

■■■

*Note the different usage of the word **sample** by the chemist and the statistician. In this example the chemist speaks of *18 samples* whilst the statistician speaks of *one sample* containing 18 items.

Questions like these will be answered in subsequent chapters. At this point we will probe the questions more deeply by translating into the language of the statistician. In this more abstract language we would ask 'What conclusions can we draw concerning the **population**, based upon the **sample** that has been examined?' This in turn prompts two further, very important questions, 'What exactly is the **population** about which we wish to draw conclusions?' and perhaps surprisingly 'What exactly is the **sample** on which the conclusions will be based?'

It is easier to answer these questions in reverse order. The **sample** can be looked upon as either:

(a) 18 bags of resin,
(b) 18 quantities of resin, each containing 20 g, or
(c) 18 measurements of polystyline content.

Whether it is better to take (a), (b) or (c) will depend upon such chemical/physical considerations as the dispersion of the polystyline within the resin and how the **variability** between bags compares with the **variability** within bags. Let us assume that each measurement gives a true indication of the polystyline content of the bag it represents and we will take (a) as our definition of the **sample**.

If our **sample** consists of 18 bags of resin then our **population** must also consist of bags of resin. The 18 bags were chosen from those in the warehouse, so it might seem reasonable to define our **population** as 'the 900 bags of resin in the warehouse', but does this cover *all* the resin about which the manager wishes to draw conclusions? He may wish to define the **population** in such a way as to include all bags of resin received in the past from this supplier and all bags to be received in the future. Before taking this bold step he would need to ask himself, 'Is the **sample** I have taken **representative** of the **population** I wish to define?'

It would obviously not be possible to take a **random sample** from such a **population** since the batches to be received in the future do not yet exist. Whenever we attempt to predict the future from our knowledge of the past we are talking about a **population** from which we cannot take a **random sample**. We may, nonetheless, be confident that our **sample** is **representative**. (The statistician prefers to discuss **random samples** rather than **representative samples** since the former are easier to define and are amenable to the tools of **probability** theory. The statistician does not, however, wish to see the statistical tail wagging the scientific/technical dog.)

Even if the plant manager defines his **population** as 'the 900 bags of resin in the warehouse' he still hasn't got a **random sample**. When the warehouse manager selected every 50th bag from the shelf he was practising what is known as **systematic sampling**. This is a procedure which is often used in the inspection of manufactured products and there is a good chance that **systematic sampling** will give a **representative sample** provided there are no hidden patterns in the **population**.

It has already been stated that the statistical techniques in this book are built upon the mathematical basis of **random sampling**, but this is only *one* of the many assumptions used by statisticians. An awareness of these assumptions is just as important to the scientist or technologist as the ability to select the appropriate statistical technique. For this reason a substantial part of Chapter 7 is devoted to the assumptions underlying the important techniques presented in Chapters 4 to 6.

2
Describing the sample

2.1 Introduction

In the previous chapter we focused attention on three words which are very important in statistics:

Sample
Population
Variability.

When a scientist or technologist is using statistical techniques he is probably attempting to make a generalization, based upon what he has found in one or more samples. In arguing from the particular to the general he will also be inferring from the *sample* to the *population*. Whilst doing so it is essential that he takes account of the *variability* within the sample(s).

It is the variability in the sample that alerts the scientist to the presence of random variation. Only by taking account of this random variation can we have an objective procedure for distinguishing between real and chance effects. Thus a prerequisite of using many statistical techniques is that we should be able to *measure* or describe the variability in a set of data.

In this chapter we will examine simple methods of describing variability and in doing so we will confine our attention to the *sample*.

2.2 Variability in plant performance

Higson Industrial Chemicals manufacture a range of pigments for use in the textile industry. One particular pigment, digozo blue, is made by a well established process on a plant which has recently been renovated. During the renovation various modifications were incorporated, one of which made the agitation system fully automatic. Though this program of work was very successful in reducing the number of operators needed to run the plant, production of digozo blue has not been completely trouble free since the work was completed. Firstly, the anticipated increase in yield does not appear to have materialized, and secondly, several batches have been found to contain a disturbingly large percentage of a particular impurity.