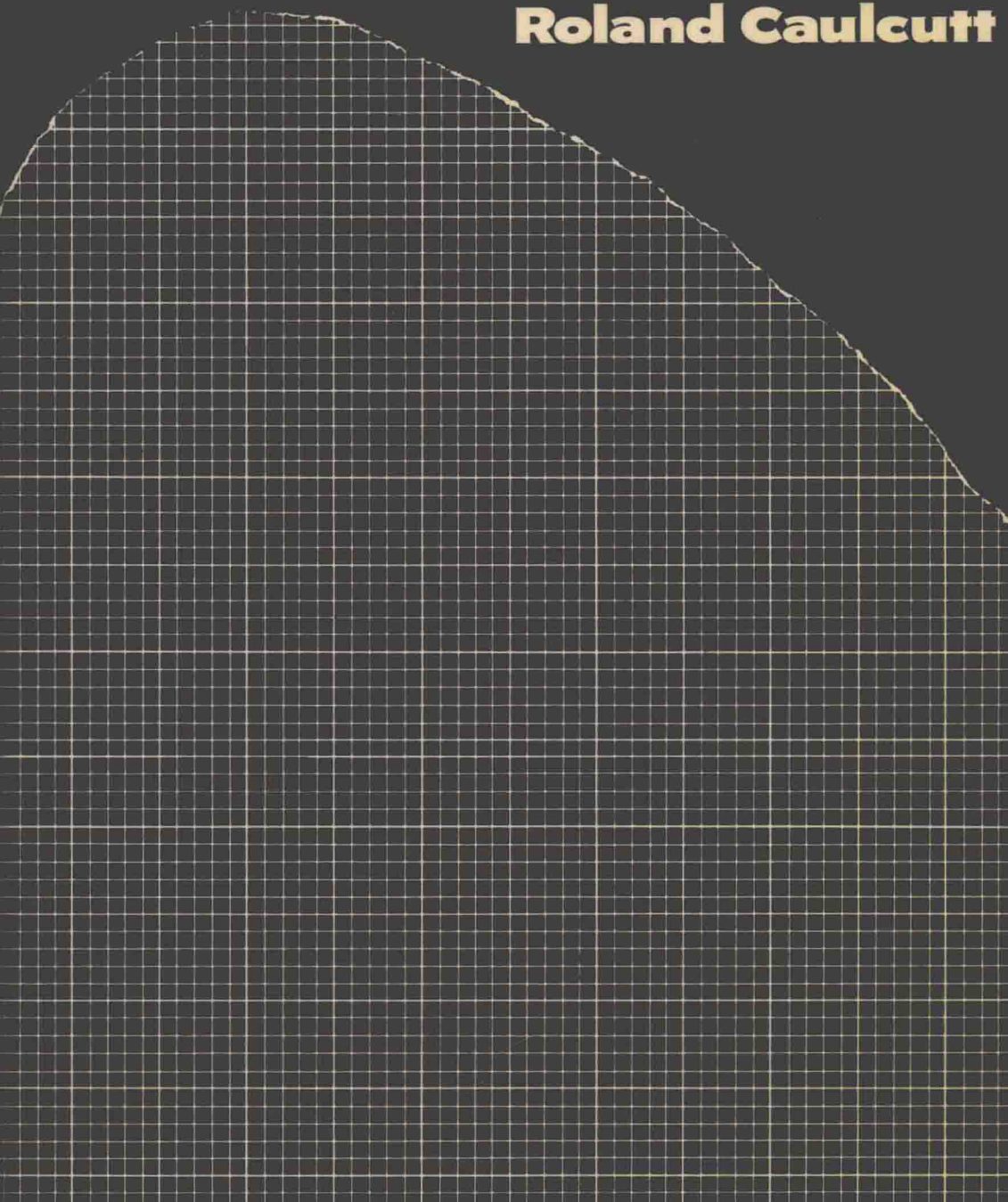


# STATISTICS IN RESEARCH AND DEVELOPMENT

**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

©1983 R. Caulcutt

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

All rights reserved. No part of this book may be reprinted, or reproduced or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or in any information storage and retrieval system, without permission in writing from the publisher.

---

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 Cataloguing 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

## *Preface*

xi

## **PART ONE**

<b>1</b>	<b>What is statistics?</b>	<b>3</b>
<b>2</b>	<b>Describing the sample</b>	<b>7</b>
2.1	Introduction	7
2.2	Variability in plant performance	7
2.3	Frequency distributions	9
2.4	Measures of location and spread	14
2.5	Cumulative frequency distributions	20
2.6	Summary	21
	Problems	22
<b>3</b>	<b>Describing the population</b>	<b>25</b>
3.1	Introduction	25
3.2	Probability distributions	26
3.3	The Poisson distribution	27
3.4	The normal distribution	30
3.5	Normal probability graph paper	37
3.6	Summary	40
	Problems	40
<b>4</b>	<b>Testing and estimation: one sample</b>	<b>43</b>
4.1	Introduction	43
4.2	Has the yield increased?	43
4.3	The significance testing procedure	48
4.4	What is the new mean yield?	50
4.5	Is such a large sample really necessary?	53
4.6	Statistical significance and practical importance	54
4.7	Summary	55
	Problems	56

<b>5</b>	<b>Testing and estimation: two samples</b>	57
5.1	Introduction	57
5.2	Comparing the precision of two test methods	57
5.3	Comparing the bias of two test methods	60
5.4	Does the product deteriorate in storage	63
5.5	A much better experiment	67
5.6	Summary	70
	Problems	70
<b>6</b>	<b>Testing and estimation: proportions</b>	72
6.1	Introduction	72
6.2	Does the additive affect the taste?	72
6.3	Where do statistical tables come from?	74
6.4	What percentage of users will detect the additive?	77
6.5	Is the additive equally detectable in all flavours?	81
6.6	The proportion of defective items in a batch	85
6.7	Summary	87
	Problems	88
<b>7</b>	<b>Testing and estimation: assumptions</b>	90
7.1	Introduction	90
7.2	Estimating variability	90
7.3	Estimating variability for a <i>t</i> -test	92
7.4	Assumptions underlying significance tests – randomness	96
7.5	Assumptions underlying significance tests – the normal distribution	97
7.6	Assumptions underlying significance tests – sample size	98
7.7	Outliers	99
7.8	Summary	103
	Problems	103

## PART TWO

<b>8</b>	<b>Investigations and statistics</b>	
8.1	Introduction	109
8.2	Some definitions	109
<b>9</b>	<b>Detecting process changes</b>	110
9.1	Introduction	112
9.2	Examining the previous 50 batches	112
9.3	Significance testing	117
9.4	Interpretation	121
9.5	Summary	124
	Problem	124

<b>10</b>	<b>Investigating the process – an experiment</b>	126
10.1	Introduction	126
10.2	The plant manager's experiment	126
10.3	Selecting the first independent variable	126
10.4	Is the correlation due to chance?	128
10.5	Fitting the best straight line	133
10.6	Goodness of fit	135
10.7	The 'true' regression equation	138
10.8	Accuracy of prediction	140
10.9	An alternative equation	142
10.10	Summary	143
	Problems	143
<b>11</b>	<b>Why was the experiment not successful?</b>	149
11.1	Introduction	149
11.2	An equation with two independent variables	149
11.3	Multiple regression analysis on a computer	153
11.4	An alternative multiple regression equation	158
11.5	Graphical representation of a multiple regression equation	159
11.6	Non-linear relationships	161
11.7	Interactions between independent variables	162
11.8	Summary	164
	Problems	165
<b>12</b>	<b>Some simple but effective experiments</b>	169
12.1	Introduction	169
12.2	The classical experiment (one variable at a time)	169
12.3	Factorial experiments	172
12.4	Estimation of main effects and interaction ( $2^2$ experiment)	174
12.5	Distinguishing between real and chance effects	176
12.6	The use of multiple regression with a $2^2$ experiment	181
12.7	A $2^3$ factorial experiment	185
12.8	The use of multiple regression with a $2^3$ experiment	189
12.9	Two replicates of a $2^3$ factorial experiment	191
12.10	More regression analysis	195
12.11	Summary	198
	Problems	198
<b>13</b>	<b>Reducing the size of an experiment</b>	201
13.1	Introduction	201
13.2	The design matrix	201
13.3	Half replicates of a $2^n$ factorial design	203
13.4	Quarter replicates of a $2^n$ factorial design	208
13.5	A useful method for selecting a fraction of a $2^n$ factorial experiment	209
13.6	A hill-climbing approach to optimization	211



13.7	Blocking and confounding	212
13.8	Summary	215
	Problems	216
<b>14</b>	<b>Improving a bad experiment</b>	<b>220</b>
14.1	Introduction	220
14.2	An alternative first experiment	220
14.3	How good is the alternative experiment?	224
14.4	Extending the original experiment	226
14.5	Final analysis	229
14.6	Inspection of residuals	235
14.7	Summary	237
	Problems	238
<b>15</b>	<b>Analysis of variance</b>	<b>241</b>
15.1	Introduction	241
15.2	Variation between samples and within samples	241
15.3	Analysis of variance and simple regression	250
15.4	Analysis of variance with multiple regression	253
15.5	Analysis of variance and factorial experiments	255
15.6	Summary	259
	Problems	260
<b>Appendix A</b>	<b>The sigma (<math>\Sigma</math>) notation</b>	<b>262</b>
<b>Appendix B</b>	<b>Notation and formulae</b>	<b>263</b>
<b>Appendix C</b>	<b>Sampling distributions</b>	<b>267</b>
<b>Appendix D</b>	<b>Copy of computer print-out from a multiple regression program</b>	<b>272</b>
<b>Appendix E</b>	<b>Partial correlation</b>	<b>275</b>
<b>Appendix F</b>	<b>Significance tests on effect estimates from a <math>p2^n</math> factorial experiment</b>	<b>279</b>
	<b>Solutions to problems</b>	<b>283</b>
	<b>Bibliography and further reading</b>	<b>335</b>
	<b>Index</b>	<b>349</b>

# —————Statistical Tables—————

Table A	Normal distribution	336
Table B	Critical values for the $t$ -test	338
Table C	Critical values for the $F$ -test	339
Table D	Critical values for the triangular test	343
Table E	Critical values for the chi-squared test	344
Table F	Confidence interval for a population standard deviation	345
Table G	Critical values for Dixon's test	346
Table H	Critical values for the product moment correlation coefficient	347
Table I	Critical values for the cusum test	347

# **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 polystyrene content.

Whether it is better to take (a), (b) or (c) will depend upon such chemical/physical considerations as the dispersion of the polystyrene 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 polystyrene 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.