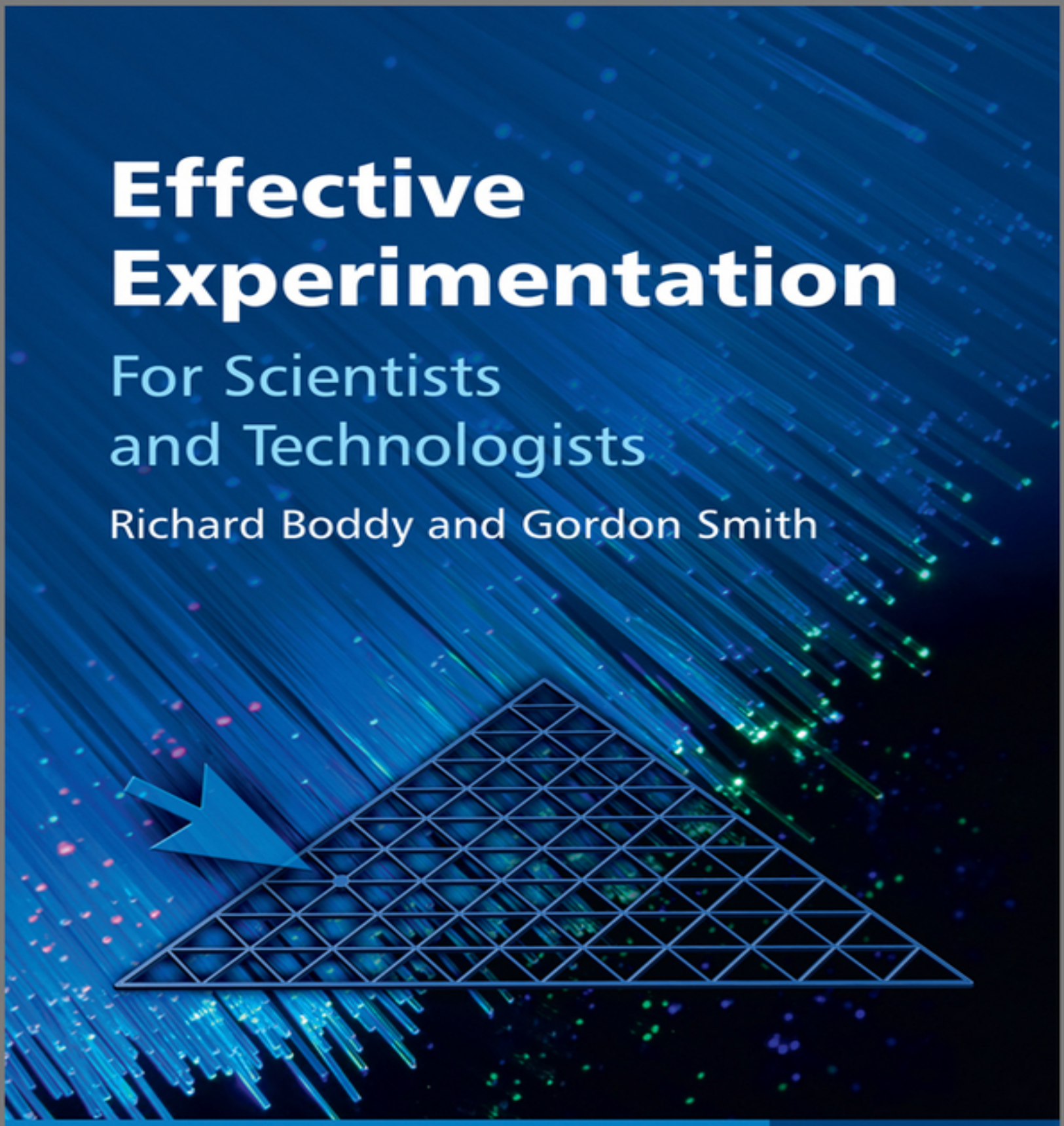
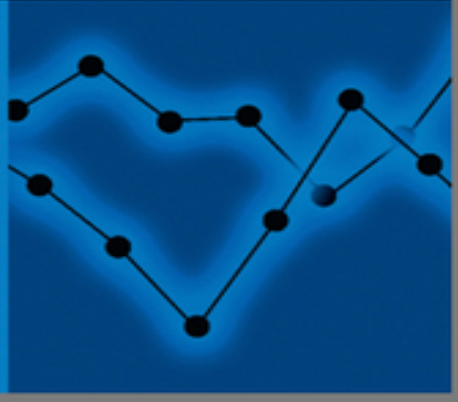
Effective Experimentation

For Scientists and Technologists

Richard Boddy and Gordon Smith



WILEY



Contents

Effective Experimentation

For Scientists and Technologists

Richard Boddy
Gordon Smith

Statistics for Industry, UK

WILEY
A John Wiley and Sons, Ltd, Publication

This edition first published 2010

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Library of Congress Cataloguing-in-Publication Data

Boddy, Richard, 1939-

Effective experimentation : for scientists and technologists / Richard Boddy, Gordon Laird Smith.

p. cm.

Includes index.

ISBN 978-0-470-68460-3 (hardback)

1. Science—Experiments—Statistical methods. 2. Technology—Experiments—Statistical methods.

I. Smith, Gordon (Gordon Laird) II. Title. Q182.3.B635 2010

507.2’7—dc22

2010010298

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

ISBN: 978-0-470-68460-3

Preface

This is a practical book for those engaged in research within industry. It is concerned with the design and analysis of experiments and covers a large repertoire of designs. But in the authors' experience this is not enough – the experiment must be effective. For this the researcher must bring his knowledge to the situations to which he needs to apply experimentation. For example, how can the design be formulated so that the conclusions are unbiased and the experimental results are as precise as needed?

Each chapter starts with a situation obtained from our experience or from that of our fellow lecturers. A design is then introduced and data analysed using an appropriate method. The chapter then finishes with a critique of the experiment – the good points and the limitations.

The book has been developed from the courses run by Statistics for Industry Limited for over 30 years, during which time more than 10,000 scientists and technologists have gained the knowledge and confidence to apply statistics to their own data. We hope that you will benefit similarly from our book. Every design in the book has been applied successfully.

The examples have been chosen from many industries – chemicals, plastics, oils, nuclear, food, drink, lighting, water and pharmaceuticals. We hope this indicates to you how widely statistics can be applied. It would be surprising if statistics could not be applied successfully by you to your work.

The book is supported by a number of specially designed computer programs and Excel Macros. These can be downloaded from Wiley's website. Although the reader can gain much by just reading the text, he/she will benefit even more by downloading the software and using it to carry out the problems given at the end of each chapter.

The book gives a brief overview of introductory statistics. For those who feel they need a more comprehensive view before tackling this book can refer to *Statistical Methods in Practice* (2009) by the same authors.

Statistics for Industry Limited was founded by Richard Boddy in 1977. He was joined by Gordon Smith as a Director in 1989. They have run a wide variety of courses worldwide, including Statistical Methods in Practice, Statistics for Analytical Chemists, Statistics for Microbiologists, Design of Experiments, Statistical Process Control, Statistics in Sensory Evaluation and Multivariate Analysis. This book is based on material from their Design of Experiments course.

Our courses and our course material have greatly benefited from the knowledge and experience of our lecturers: Derrick Chamberlain (ex ICI), Frits Quadt (ex Unilever), Martin Minett (MJM Consultants), Alan Moxon (ex Cadbury), Ian Peacock (ex ICI), Malcolm Tillotson (ex Huddersfield Polytechnic), Stan Townson (ex ICI), Sam Turner (ex Pedigree Petfoods) and Bob Woodward (ex ICI). In particular we would like to acknowledge Dave Hudson (ex Tioxide) who wrote the Visual-Basic-based software, John Henderson (exChemdal) whowrote the Excel-based software and Michelle Hughes who so painstakingly turned our notes into practical pages.

Supporting software is available on the book companion website www.wiley.com/go_effective.

Richard Boddy

Gordon Smith

Email: s4i@aol.com

April 2010

1

Why bother to design an experiment?

1.1 Introduction

There are many aspects involved in successful experimentation. This book concentrates mainly on designing and analysing experiments but there is much more required from you, the experimenter. You must research the subject well and include prior knowledge available from previous experiments within your organization. You should also consider a strategy for the investigation such as considering a series of small investigations. You must plan the experiment operationally so it can be successfully undertaken and, lastly, having analysed the experiment you must be able to interpret the analysis and draw valid conclusions.

If you follow that path, then you should have completed a successful project.

If not, then you may have wasted resources, had insufficient trials or data to be able to make conclusions that will stand up to scrutiny, or end up by making invalid claims.

There is no guarantee of finding all the answers, but you will have been well informed and will have made the most efficient use of the information and data available.

Let us consider some situations that illustrate the benefits of using designed experiments.

1.2 Examples and benefits

1.2.1 Develop a better product

An oil formulator has been charged with the task of improving the formulation of a lubricating oil in order to improve the fuel economy of motor engines. There are two important components of the formulation-type of base oil and level of friction modifier. Without knowledge of experimental design, he does not wish to change both variables at once. He keeps to the current level of friction modifier and changes the base oil, gaining an improvement. The next trial therefore uses the new base oil and he changes the level of friction modifier. It also gains a small improvement. He reports that the new oil should be made with new base oil and changed level of friction modifier and can be called ‘New Improved’.

This is an inefficient way of exploring the experimental space, even assuming that only two levels of each variable are possible. He has assumed that a change resulting from changing the level of friction modifier when using one base oil will be repeated with the other, but this does not often happen with manufacturing processes. There are often interactions. He should have tested all four combinations of the two levels of both variables in a factorial experiment. It may be that the best combination is none of those that he examined, as in the example in Chapter 4.

Such an experiment, if replicated (more than one trial at each set of conditions) would give him the following benefits:-

(i) determination of the effects of each variable and knowledge of whether or not there is an interaction between them;

(ii) determination of whether an improvement can be made;

(iii) a measure of the batch-to-batch variability that enables him to test differences for significance.

1.2.2 Which antiperspirant is best?

A toiletries company has developed some formulations of an antiperspirant and wishes to determine which one is most effective. After chemical and microbiological tests the only realistic way is to test them out on volunteers in a carefully controlled environment. Perhaps at first thought a large number of volunteers should be assembled, and formulations allocated at random to the volunteers, each person testing one formulation. The trouble with this approach is that there is a lot of variation between one person and another in amounts of perspiration and the effectiveness of an antiperspirant, which would obscure any differences that there might be between formulations.

An experimental design is needed so that person-to-person differences ('nuisance' variation) can be identified but their effect removed when comparing the formulations. Thus, a panel of volunteers is gathered, and each one tests every formulation. The person-to-person variation would be there but would affect all the results, but differences between formulations should be more consistent. This design is known as a randomized block design, introduced in Chapter 18.

The benefits of this design are:-

(i) formulations can be directly compared;

(ii) person-to-person variability can be quantified but its effect eliminated in the comparison of formulations;

(iii) the best formulation can be identified.

1.2.3 A complex project

Bungitallin Spices are developing a new spice for lightly flavoured cheeses. They have identified 30 ingredients,

decided on a composition and produced a trial sample. The taste seems reasonable and they decide to proceed with a marketing campaign to launch their new product.

Now clearly Bungitallin have a great knowledge of spices and it is perhaps not surprising in the spice industry that 30 ingredients have been included. However, there are many questions that are readily brought to mind.

i. How did they decide on the composition?

ii. Could they have done better if they used experimental design?

iii. Do all 30 ingredients contribute to taste? How many can be discerned and at what levels? How many can be removed without any discernible effect in the taste?

iv. How many of the 30 ingredients are necessary for texture or other parameters and at what levels?

v. How do the ingredients interact with each other?

vi. How is the spice to be produced? What are the process conditions? How robust are the conditions?

vii. How much variation can we expect from batch-to-batch? Is this acceptable or does it need reducing?

Clearly there area lot of questions to be answered. If we attempt to answer all the questions in an unstructured manner the cost may be far greater than the profit from launching the spice. On the other hand, if we do nothing Bungitall in may be left with a failure at great cost. Thus,we must investigate, but in an efficient way.

Experimental design offers an approach that will enable us to achieve our objective in an efficient manner and give us unbiased results, thus enabling us to have confidence in our conclusions.

Different chapters of this book will help you to answer these questions.

Questions iii), iv) and v) can be investigated using *factorial* or *fractional factorial designs* followed by *response surface methods* to achieve the best formulation. If the experiment

is too large to carry out in one trial a *central composite design* may be employed.

Questions vi) and vii) can be investigated using *saturated designs or computer-aided experimental designs (CAED)*.

Question vii) can be investigated using *randomized block* or *Latin Square designs*.

1.3 Good design and good analysis

Of course, it is not only necessary to carry out a good design but it must be followed by a good analysis-in fact, when designing an experiment we should also consider how it is to be analysed.

This book starts with a chapter that covers *summary statistics, the normal distribution, confidence intervals and significance testing.* Later it refers to multiple regression, a necessary tool when the design has an imbalance which can occur for many reasons such as ‘lost’ data.

All these designs and methods of analysis will greatly enhance your experiments but we must not forget the most important aspect of experimental design-the researcher’s knowledge. The design is aimed at increasing this knowledge and making it more rigorous so that we have a high degree of certainty that actions resulting from the design will prove to be successful.

2

A change for the better – significance testing

2.1 Introduction

‘Have we improved the process?’ This is a frequently asked question. A change has been made to the process on the basis that it will improve it, but do subsequent results confirm this? It is perhaps usual for people to say, for example, that before the change the process gave a mean reading of 29.3, but now it is 29.6, therefore the change has worked. However, all processes are subject to noise (error, variation) and thus the change to 29.6 could be due to noise rather than the change. How are we to know? As well as quantifying the mean of a process we must also quantify the variability. Then we shall be in a position to know whether 29.6 could have been caused by noise or whether it must have been due to the process change because it is well beyond the value that could have been attributed to noise.

As well as looking at the above situation we shall take the opportunity in this chapter to introduce some of the building blocks of statistics used in analysing experiments – measure of average and variability, blob charts, histograms, normal distribution, confidence intervals and significance tests. However, these are very brief expositions of these topics and for a more in depth treatment readers are referred to ‘*Statistical Methods in Practice’ by Boddy & Smith*.

2.2 Towards a darker stout

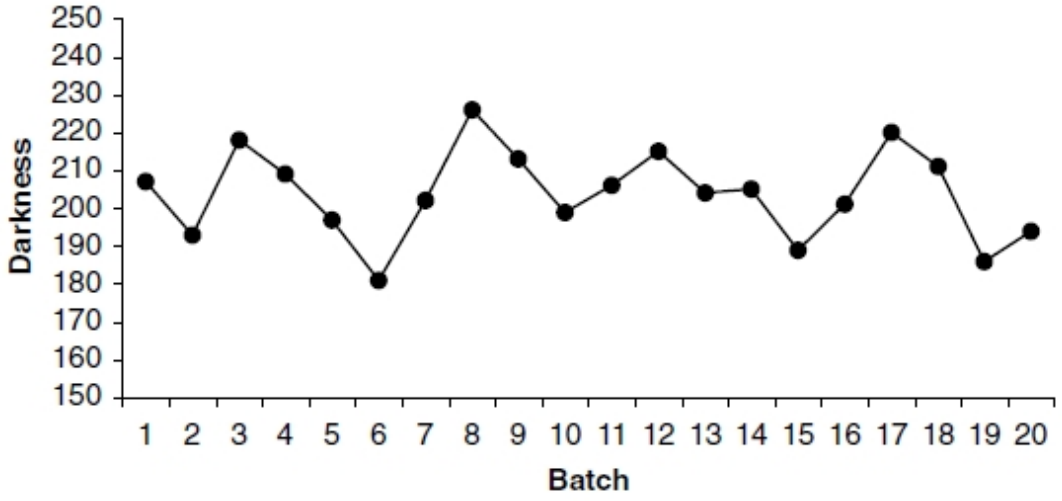
Stout is a drink that traditionally is dark - in fact the darker the better. Trentside Ales use adjunct AX751 to give a dark colour but a rival product BZ529 has been trialled on a pilot. The trials indicated that BZ529 gave a darker colour so Rob Whetham, the Development Manager of Trentside, has decided to put it into production in one of their four vats and then compare it with batches using AX751. The past 20 batches gave the results shown in Table 2.1 for darkness using AX751 as measured on a spectrometer:

Table 2.1 Darkness of 20 batches using AX751

Batch	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Darkness	207	193	218	209	197	181	202	226	213	19
Batch	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Darkness	206	215	204	205	189	201	220	211	186	19

Rob's first task is to look for trends since the presence of a trend would indicate the process was not stable and it would be difficult to judge whether the new additive had been effective. A runs chart for the data is shown in Figure 2.1 that indicates that the data has no trend. A better analysis would be using a cusum technique as outlined in '*Statistical Methods in Practice*'.

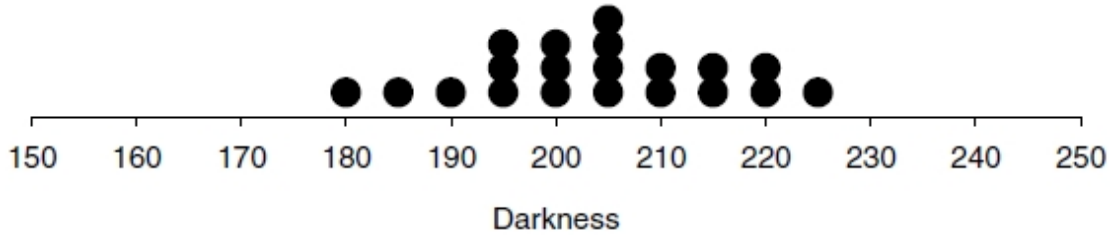
Figure 2.1 Runs chart.



Rob next looks at the distribution of data using a blob diagram as shown in Figure 2.2.

The distribution is typical of a process under control with values spread fairly symmetrically with more in the centre.

Figure 2.2 Blob diagram.



2.3 Summary statistics

The distribution of the data, as shown in a blob diagram or histogram, provides a valuable visual way of making judgements about the data. As well as the plots it is usually valuable to summarize data in terms of average and variability using the mean and standard deviation.

The mean is the sum of all observations divided by the number of observations.

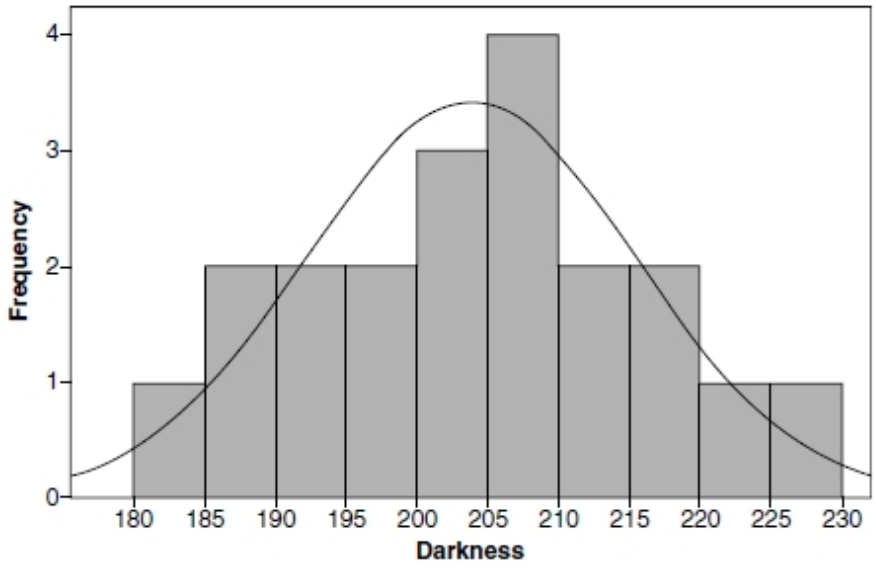
The sum was 4076, there were 20 observations, so the mean in this case equals 203.8. It is usually referred to as the sample mean ($\bar{x}$).

The standard deviation can be defined as the average of the **deviations** from the mean. It is a strange sort of average but it does indicate the amount of variation. It is usually referred to as the sample standard deviation (s). Its value is 11.7. If we look at the 20 values we see that 14 are within one standard deviation of the mean, i.e. 192.1 to 215.5 and six are outside these limits which is to be expected when we use a standard deviation.

2.4 The normal distribution

Figure 2.3 shows the data presented in an alternative way to a blob diagram.

Figure 2.3 Histogram plus normal-distribution curve.



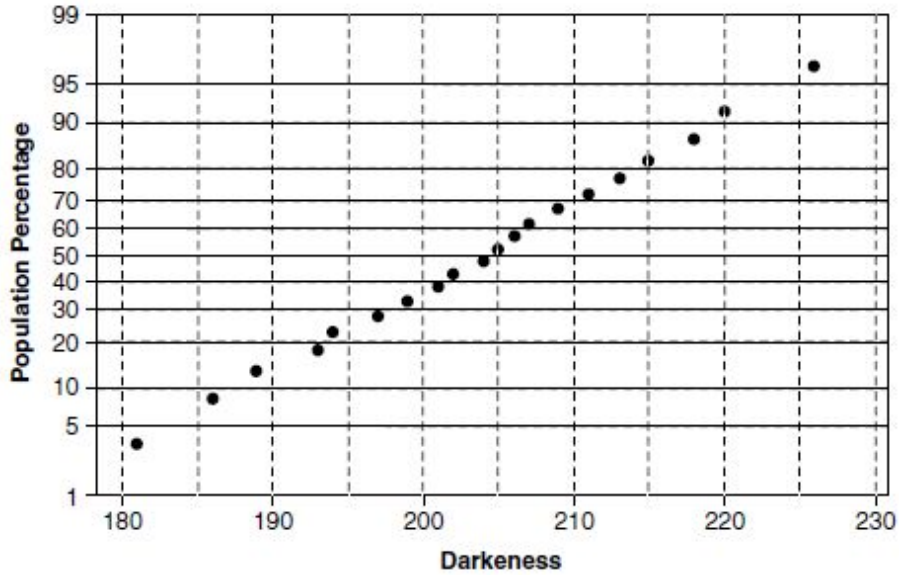
In a histogram, the height of each bar is proportional to the number of batches within the specified range. Superimposed onto the histogram is a normal-distribution curve. The normal distribution occurs when a process is subject to many additive errors, none of which are dominant in magnitude. It is found in abundance in processes where effort has been made to reduce the effect of any dominant source of error. It is also found in many natural processes.

The normal distribution is defined by two parameters, the population mean and population standard deviation. However, the population statistics are unknown and Rob has only sample data so it is these values – i.e. mean of 203.8 and a standard deviation of 11.7 that have been used to draw the normal distribution.

The normal distribution is often important in analyses of experiments since it provides a check on the validity of the conclusions. It is, however, very difficult to assess using histograms. A far better check is a probability plot shown in Figure 2.4. In this, the 20 observations have been ranked in order and then plotted against a nonlinear scale that

represents the position of the observation if it was obtained from a perfect normal distribution.

Figure 2.4 Normal probability plot.



The judgement we must now make, allowing for sampling or other errors, is: 'Is the shape of the plot sufficiently linear?' This takes judgement but Rob is assured that a straight line is a very good fit.

2.5 How accurate is my mean?

Rob has 20 batches with darkness values. Assuming these are representative samples from a stable process, how accurate is the mean? In order to calculate its accuracy we use a 95% confidence interval for the population mean (μ) that is given by the formula:

$$\bar{x} \pm \frac{ts}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean

s is the sample standard deviation

n is the number of observations

and t is the coefficient obtained from Table A.2 using a 95% confidence level and n – l degrees of freedom.

Degrees of freedom are the number of independent deviations from the mean used to calculate the standard deviation, one less than the number of observations. To illustrate this, if we have two results, say 240 and 260, the mean is 250 and both results must give the same deviation of 10. Thus, there is only one degree of freedom.

In our example:

$\bar{x} = 203.8$

$s = 11.7$

$n = 20$

$t = 2.09$ with 19 degrees of freedom at a 95% confidence level

$$95\%\ \text{confidence interval} = 203.8 \pm \frac{2.09 \times 11.7}{\sqrt{20}}$$
$$= 198.3\ \text{to}\ 209.3.$$

This means we are 95% certain that this stable process will, in the long run, produce stout using AX751 with a mean darkness between 198.3 and 209.3.

Rob is pleased with the narrowness of the interval. This should allow him to make a reasonable judgement about whether the new additive is better.

2.6 Is the new additive an improvement?

Rob is now in a position to design and carry out an experiment. He decides to use six trials, not based on design considerations but on product considerations. If the adjunct BZ529 is detrimental to the product he would wish to abandon it fairly quickly.

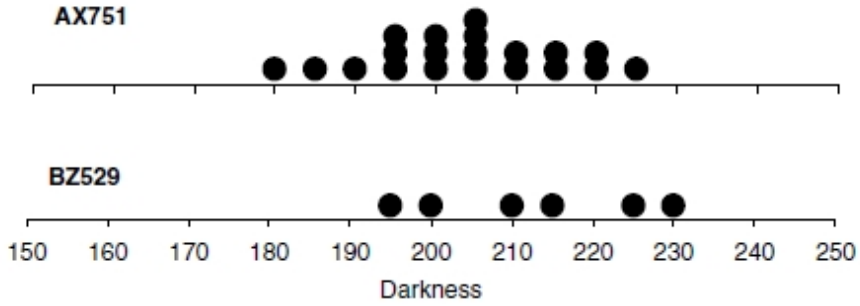
The next six batches give darkness values as shown in Table 2.2.

Table 2.2 Darkness values of 6 batches from BZ529

Batch	21	22	23	24	25	26
Darkness	215	231	208	195	225	198

Rob is both encouraged and discouraged by the results. The result of 231 is higher than darkness readings obtained with AX751 but 195 and 198 are below the mean of the last 20 batches. On the other hand, the mean of BZ529(212.0) is higher than obtained with AX751. A plot of the data for both adjuncts is shown in Figure 2.5.

Figure 2.5 Blob diagram for the two adjuncts.



Because of variation any two means are always likely to be different. What we need to know is whether this difference is due to chance or is it due to the adjuncts? We can answer this by carrying out a two-sample t-test, but before doing so we need to combine the two standard deviations from the two adjuncts. The summary statistics for the two adjuncts are shown in Table 2.3.

Table 2.3 Summary statistics

	Mean	Standard deviation	Number of results	Degrees of freedom
Adjunct AX751	203.8	11.7	20	19
Adjunct BZ529	212.0	14.4	6	5

$$\text{Combined SD(s)} = \sqrt{\frac{(df_A \times SD_A^2) + (df_B \times SD_B^2)}{df_A + df_B}}$$

where SD_A, SD_B are the sample standard deviations for AX751 and BZ529 with df_A and df_B degrees of freedom, respectively

$$= \sqrt{\frac{(19 \times 11.7^2) + (5 \times 14.4^2)}{19+5}}$$

$$= 12.3 \text{ with } 19 + 5 = 24 \text{ degrees of freedom.}$$

We use the following procedure to carry out the significance test:

Null hypothesis: In the long run the two adjuncts will give the same mean, i.e. $\mu_A = \mu_B$

Alternative hypothesis: In the long run the two adjuncts will give different means, i.e.

$$\mu_A \neq \mu_B$$

$$\text{Test value:} = \frac{|\bar{x}_A - \bar{x}_B|}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

where $\bar{x}_A$, $\bar{x}_B$ are the sample means for AX751 and BZ529, respectively; $|\bar{x}_A - \bar{x}_B|$ is the magnitude of the difference between $\bar{x}_A$ and $\bar{x}_B$; n_A, n_B the number of observations, and s the combined SD.

$$= \frac{|203.8 - 212.0|}{12.3\sqrt{\frac{1}{20} + \frac{1}{6}}}$$

$$= 1.43.$$

Table value: t = 2.06 from Table A.2 with (n_A + n_B 2) = 24 degrees of freedom at a 5% significance level

Decision: If the test value is greater than the table value we reject the null hypothesis and accept the alternative.

However, in this case the test value is less than the table value. We cannot reject the null hypothesis.

Conclusion: There is insufficient evidence to conclude that there is a difference between the means for AX751 and BZ529.

Rob is disappointed. He has failed to establish that the new adjunct is significantly better but there are grounds for being optimistic – the sample mean is higher, the test value is approaching the tablevalue and he only produced six batches with the new adjunct. How many batches should he have produced?

2.7 How many trials are needed for an experiment?

The number of trials required in a two-sample experiment is estimated by the formula:

$$n_A = n_B = 2\left(\frac{ts}{c}\right)^2$$

where n_A, n_B are the number of trials for each adjunct, c the difference that is required to be significant, s is a measure of the combined standard deviation, and t is a value from Table A.2 with the same number of degrees of freedom as the combined standard deviation.

In this example:

s = 12.3 and t = 2.06 from Table A.2 based on 24 degrees of freedom at a significance level of 5%

The difficult decision for Rob is the size of the difference (c) that needs to be found significant by the experiment. It is decided that improving the darkness on average by 8 will lead to a noticeable better stout.

$$n_A = n_B = 2\left(\frac{ts}{c}\right)^2 = 20.$$

We should note that this is a very useful formula, not only for a two-sample t-test but for factorial experiments given in later chapters. With two-level factorial designs this formula can be used to give an indication of the size of experiment required.