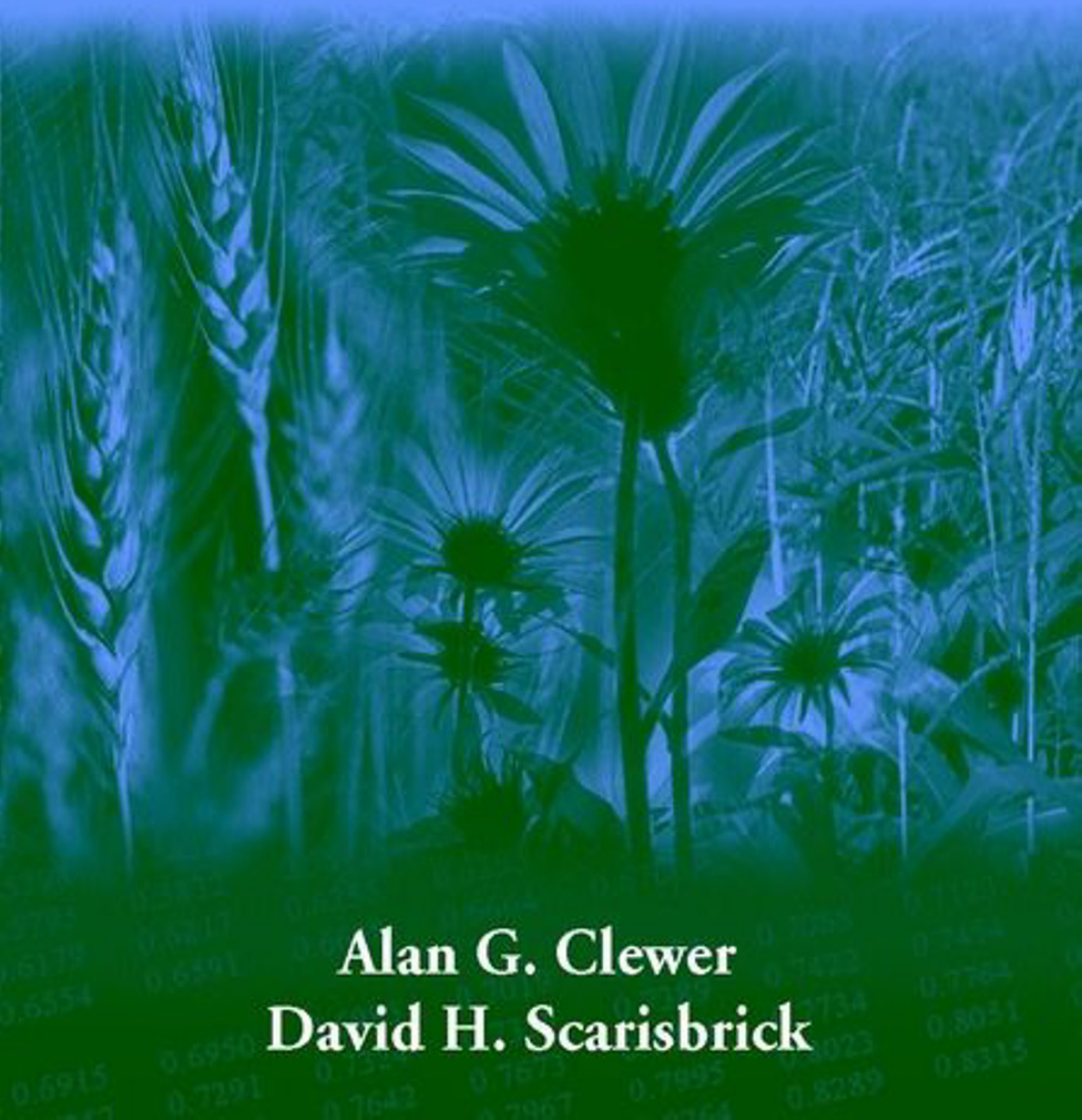
WILEY

Practical Statistics and Experimental Design for Plant and Crop Science



Alan G. Clewer
David H. Scarisbrick

PRACTICAL STATISTICS AND EXPERIMENTAL DESIGN FOR PLANT AND CROP SCIENCE

PRACTICAL STATISTICS AND EXPERIMENTAL DESIGN FOR PLANT AND CROP SCIENCE

Alan G. Clewer and David H. Scarisbrick

T. H. Huxley School of Environment,
Earth Sciences and Engineering
Imperial College at Wye,
Ashford, Kent, UK

JOHN WILEY & SONS, LTD

Chichester • New York • Weinheim • Brisbane • Singapore • Toronto

Baffins Lane, Chichester,
West Sussex PO19 1UD, England

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries):
cs-books@wiley.co.uk
Visit our Home Page on: http://www.wiley.co.uk
or http://www.wiley.com

Templates for many of the figures in this book were created using Microsoft Excel software

Other Wiley Editorial Offices

John Wiley & Sons Inc., 605 Third Avenue,
New York, NY 10158-0012, USA

WILEY-VCH Verlag GmbH, Pappelallee 3,
D-69469 Weinheim, Germany

Jacaranda Wiley Ltd, 33 Park Road, Milton,
Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons (Canada) Ltd, 22 Worcester Road,
Rexdale, Ontario M9W 1L1, Canada

Library of Congress Cataloging-in-Publication Data

Clewer, Alan G.
Practical statistics and experimental design for plant and crop science / Alan G. Clewer and David H. Scarisbrick.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-89908-9 (hbk : alk. paper) – ISBN 0-471-89909-7 (pbk. : alk. paper)
1. Botany–Research–Statistical methods. 2. Botany, Economic–Research–Statistical methods. 3. Biometry. I. Scarisbrick, D. H. II. Title.
QK51. C58 2001
580'.7'27–dc21 00-047313

British Library Cataloguing in Publication Data

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

ISBN 0 471 89908 9 (PPC)
ISBN 0 471 89909 7 (Pbk)

Typeset in 10/12pt Times by Dobbie Typesetting Ltd, Tavistock, Devon.

Contents

Preface

The references at the end of this book confirm that there are many textbooks on statistics for students who are interested in applied biology. Most of these cover the same subject material, although they vary quite widely in styles of presentation. Thus, it is important to answer the question—why has another book been written? It would be pointless to review the same topics on the design and analysis of experiments unless some original features can be detected by the reader in the 19 chapters that make up this book. One claim to originality is that this text is closely linked to the computer outputs from three commonly used statistical packages (Genstat 5—release 4.1, Minitab—release 12.1 and SAS—release 6.12). However, this in itself may not be sufficient to justify the vast amount of time required to produce another text, and provide a satisfactory answer to the above question.

The answer is more closely linked to our concern about the misuse of statistics by many students and their lack of understanding of the basic principles that underlie the statistical techniques they refer to in their dissertations. Because it is now all too easy to carry out inappropriate analyses by computer, the advent of statistical packages has diluted many students' understanding and interest in the basic principles which are the foundation stone of good design. They often look for an analytical method that seems to 'fit' their data. This frequently results in problems of interpretation; the design of experiments and the related data analysis should not be treated as separate components of the experimental process. Design is more important than analysis, because without a good design, the analysis is meaningless. Matching experimental data with a program given in a textbook on computing and statistics after the experiment has been completed is rarely successful and should be avoided.

As external examiners we frequently find computer summaries of data analyses in the appendices of dissertations, although when faced with simple questions the majority of students seem to have very little understanding of the terms and figures given in their printouts. For example, the ANOVA summary usually includes a mean square (MS) column. It is now quite rare for students to give the correct translation of the error (residual) mean square and even more rare for them to know that this is also an estimate of variance. There is always muddle in relation to the constants quoted for simple regression equations, and discussion of the usefulness of slope and intercept is usually missing mainly because the role of the constants and equations is rarely understood.

Thus, the book has mainly been written to demonstrate both the usefulness of statistics (like previous texts), and also to provide a clear explanation of terms, figures and symbols given in computer printouts. We decided not to link these discussions to one statistical package in order to illustrate the diversity of layouts that are used to provide a summary of the same results. For most statistical procedures the three packages have much output in common, but some are better than others for certain options. For the examples for which we give computer outputs, we use the package that we consider illustrates the main points we are trying to make. The book also encourages students to review the underlying principles of many statistical tests before using them in their research work; this point should be noted by supervisors! When interviewing students in relation to their data-handling methods it is sometimes the supervision that can be faulted, especially when students confess that they simply browsed a recommended text in order to find a statistical example which had similarities to their own design and experimental results.

Students reading this book should initially work through the text and exercise examples using a hand calculator. This technique assists understanding and interpretation. When interpretive skills have been achieved, results can then be confirmed by studying the computer outputs provided. Future data can then be immediately analysed using computer packages.

The classical textbooks describe how to do calculations by hand using the *correction factor method.* This method is no longer needed due to the widespread use of computers and hand calculators. However, we do explain it in the chapter on one-way analysis of variance. In subsequent chapters we show how the various sums of squares can be calculated using the standard deviation function provided by most calculators.

Biological statistics (biometry) is relevant to all areas encompassed by the general subject title Applied Plant Biology. It is an essential tool which is used to uncover or discover scientific information contained within raw (new) data. Like applied plant sciences, biometry is a broad-based subject. It is concerned with all aspects of experimentation (design, sampling methods, data analysis, interpretation and discussion) which are relevant to the research objectives. Experiments should always be designed in such a way that questions posed by the objectives have a good chance of being answered. This implies that the method of analysis and statistical tests are determined before the experiment is carried out.

This book introduces students to the important role of biometry in applied plant science research. Because many readers will have little prior knowledge of biological statistics, the first five chapters are used to summarise the basic principles which underpin simple statistical concepts. Although a more advanced treatise is provided in later chapters, the mathematical theory underpinning the techniques described is mainly ignored. Instead, the text provides some description and justification of the most important calculations which are widely used in statistics testing, and a detailed review of computer output which is now commonly presented by students in their dissertations. For each output an interpretation is given, and for many, how most entries can be found from a hand calculator. A novel feature of the book is the inclusion of examples showing how the sums of squares for the various terms in the most popular analysis of variance models are partitioned without using algebra.

Many biology students feel uneasy with statistical calculations and equations, and are rarely concerned with acquiring an understanding of the mathematical theory of statistics. They are mainly interested in learning how to apply a range of statistical tools to design, analyse and then interpret their results. Although this approach to biological statistics is commonplace, it is the authors' view that effort should still be made to understand some of the background calculations associated with many tests used to compare treatment means. This is an important part of comprehending statistical methods; comprehension is still important even in an era when computers usually undertake the time-consuming labour of arithmetic calculations.

When statistical analyses were carried out using hand-operated calculators, the amount of data collected and analysed was mainly controlled by the sluggishness of desk machinery. Because there were no mountains of computer output to file and review, raw data and results were usually pondered and discussed in great detail. Packaged computer programs have revolutionised data-handling techniques. They have removed most of the drudgery from the analysis of large experiments which compare a range of treatments, and eliminated arithmetic errors. In addition, because most programs provide facilities for reviewing raw data, it is now much easier to check that assumptions which underpin many statistical tests are really true before proceeding with an analysis. An initial study using graphs, scatter diagrams and tables is helpful in deciding whether a particular statistical method is really appropriate for the new data being examined.

It could be argued that statistical tables are no longer required to carry out statistical tests when the *P*-value is given in the computer output. Nevertheless, to understand the *P*-value, a familiarity with the underlying distribution of the test statistic is required, so we include statistical tables in Appendices 1 to 10 inclusive. They are also required for those readers without a computer or relevant software.

The authors hope that this book will assist students and researchers in crop and plant sciences to explain in simple language the objectives of simple statistical tests, and achieve an understanding of the principles of experimentation. This book can be read at several levels and so will be useful for a wide range of readers. It will be useful for teachers. It will be useful for users of statistical packages who want to interpret the output. It will be useful for researchers who want to know how to design a simple experiment and analyse and present the results. It will also be useful for those who want to know a little of the background theory needed to justify the procedures and how the calculations are performed.

ACKNOWLEDGEMENTS

We wish to thank BASF for permission to reproduce some diagrams and tables. We also thank MINITAB, SAS and Genstat for permission to include computer output. SAS is a registered trade mark of the SAS Institute Inc.

Chapter 1

Basic Principles of Experimentation

1.1 INTRODUCTION

The principles of experimentation can be studied by students who enjoy pure mathematics, and by those who wish to use mathematical principles as tools only for the design and analysis of their experiments. When carrying out research work at an experimental station or university, the applied biologist may be able to discuss the layout and analysis of experimental work with a professional statistician. However, when working in isolated rural areas (especially in developing countries), the experimenter must demonstrate a basic understanding of statistics and also have the confidence to solve design and data-handling problems. Even when professional support is available, it is still essential that the researcher is aware of the wide range of statistical procedures used by applied biologists. More importantly, he or she must have acquired sufficient statistical skills to interpret and discuss experimental results which are analysed using computer packages such as Genstat, Minitab and SAS.

Experimental objectives must be clearly and concisely stated at the outset of an investigational programme. Before starting a field or greenhouse experiment, it is wise to purchase a diary. The first page should be used to give a clear exposition of research objectives. The diary should contain a summary of previous cropping, a description of the treatments, detailed site plans and daily observations. When an experiment is written up, these observations may help to explain results that may at first seem to be anomalous.

1.2 FIELD AND GLASSHOUSE EXPERIMENTS

There are two main systems of experimentation used to explore the effect(s) of experimental treatments on plant development and yield (Figure 1.1). In both, treatments may be applied before, at, as well as after sowing. For example, the objectives of an experiment may be to compare different seed dressings, fertiliser

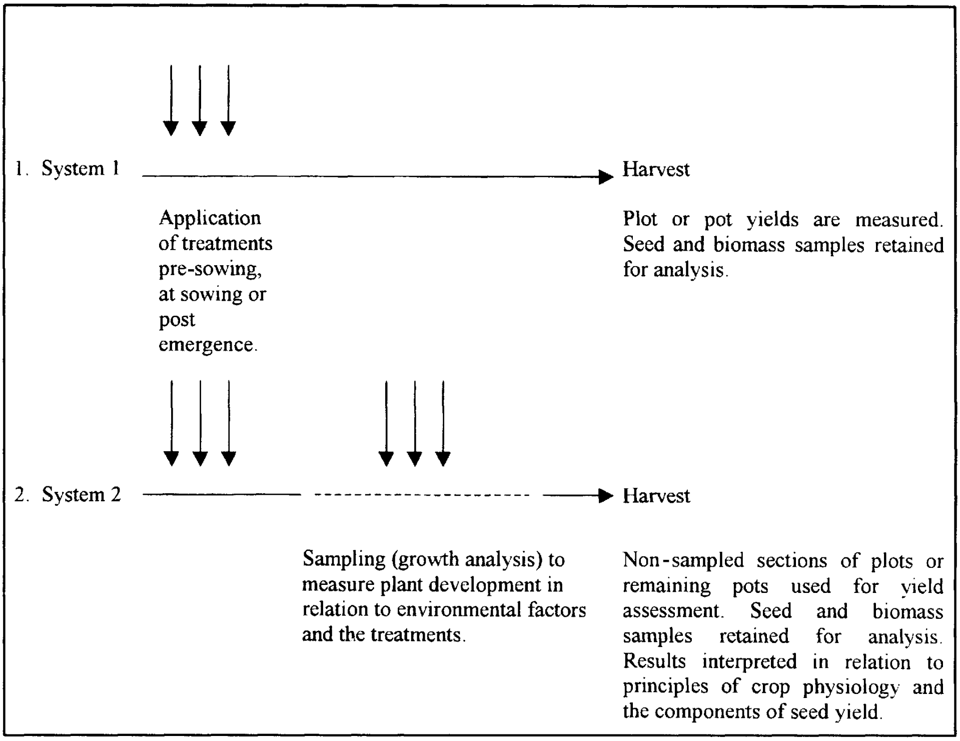

1. System 1
Harvest
Application of treatments pre-sowing, at sowing or post emergence.
Plot or pot yields are measured. Seed and biomass samples retained for analysis.
2. System 2
Harvest
Sampling (growth analysis) to measure plant development in relation to environmental factors and the treatments.
Non-sampled sections of plots or remaining pots used for yield assessment. Seed and biomass samples retained for analysis. Results interpreted in relation to principles of crop physiology and the components of seed yield.

Figure 1.1. Systems of experimentation

placement at drilling, or the post-emergence application of a plant growth regulator at defined growth stages. The first descriptive system shown in Figure 1.1 is widely used by agronomists at arable research centres and farm demonstration sites. However, if the chosen treatments result in a significant increase in yield, the agronomist may be unable to fully explain the field results. It is impossible to provide an in-depth discussion of morphological and physiological factors affected by the treatments unless some additional measurements such as light interception, leaf area index, or crop growth rate are taken during the growing season (System 2).

System 1 is also rather risky because the results of a season's technical work (site mapping, cultivations, sowing, pot and plot maintenance), are solely dependent on data collected at final harvest. It is frustrating when treatment effects which were clearly visible during mid-season are obscured by lodging and seed loss due to thunderstorm damage during the ripening period. Similar end-of-season losses can also occur in glasshouse trials using System 1. In a hydroponic study on the response of wheat to varying concentrations of sulphur, the ears were destroyed by field mice which invaded the glasshouse cubicle just two days before the planned final harvest date. A trial which needed a great deal of maintenance time in relation to monitoring and topping up the different nutrient solutions only provided advice on rodent control! If measurements such as number of tillers per plant and sulphur concentrations in leaves and stems had been taken during vegetative development (System 2), some useful discussions of the effects of sulphur depletion on wheat development may have resulted.

The second system provides useful background information which may be used to interpret and discuss the final results. For example, when regular descriptive samples are taken the effects of treatments on plant morphology and the reproductive components of yield can be determined. If more sophisticated facilities are available, it may also be helpful to carry out physiological measurements such as chlorophyll concentration, light interception and photosynthesis. Subsequently, yield data can be analysed and interpreted in much more detail; additional confidence may then result in the reliability of a new product or variety. If final yield data are lost from System 2, the researcher may still have sufficient information in his or her diary to create a scientific report on vegetative development, flowering, and the early stages of reproductive development. Data collected during the experimental growing season may also be used to look for possible relationships between measurements such as soil temperature, rainfall, emergence and plant development using simple correlation techniques (Chapter 7).

As a disadvantage, the introduction of crop sampling increases experimental costs. Growth analysis is especially tedious and time consuming, and it is essential to ponder original experimental objectives before investing time in measuring components such as leaf area and numbers of seeds in pods or spikelets. The applied plant scientist may enjoy philately as a hobby, but must avoid the temptation to file vast quantities of descriptive data unless the scientific reasons are clearly defined. Computer packages will analyse data and produce means and standard errors. However, they are unable to guide the researcher on interpretative skills and sampling procedures or indicate whether the time invested in measuring a particular parameter is really worthwhile.

1.3 CHOICE OF SITE

The choice of field and the siting of a trial are probably the most important decisions made by the researcher. Incorrect choice of site may mean that the experimental results are difficult to interpret or are even meaningless. There are several common-sense starting points. It is essential to choose a uniform site as trials are designed to detect differences between small plot areas. Thus any variations in soil texture or pH within a site may partly or completely mask treatment effects. Local knowledge is clearly important, and it may be wise to consult historical records of drainage systems and former field boundaries. Previous site management must be researched because many legume and oilseed crops should be sown only once in a 5-year rotation — the principles of crop rotation must be adhered to when the experimental objectives are being defined.

An accessible site makes for ease of sampling, although it is worth remembering that trials situated near parking areas and footpaths may be subjected to damage by trespassers. It is best to avoid compacted headlands and wooded areas as downdraughts from trees can cause severe lodging. In order to minimise edge effects it is usually advisable to surround the experimental site with the crop species being studied. Frequently, plots will be positioned so that they can be sprayed with protective agrochemicals using the tramlines of a surrounding commercial crop.

When undertaking an agronomic research programme, it is valuable to have growth room and glasshouse facilities in addition to field sites. This enables detailed physiological work to be continued during the winter months. In theory, research glasshouses should provide a reasonably uniform environment in which light intensity, temperature and daylength can be controlled by computer technology. In practice, the glasshouse can be a more variable environment to work in than the field. There are frequently wide and sudden fluctuations in temperature between overcast and sunny periods even in the winter months, and variation in light profiles especially when neighbouring research cubicles are using a different daylength regime. In order to minimise these problems, it is important to re-randomise the position of the growing containers on the greenhouse benches from time to time, and surround the experimental unit with guard pots to reduce edge effects. It is essential to monitor pests and diseases — mildew, botrytis, aphids and white fly are common problems in the glasshouse, and while biological systems for insect control can be used, it is wise to have insurance supplies of agrochemicals in store, especially for the control of fungal diseases.

1.4 SOIL TESTING

Large experimental sites are more likely to include variations in soil texture, thus for most arable crop investigations the total experimental area rarely exceeds 1 ha. Larger plots are usually necessary in grassland or grazing studies in order to accommodate fencing, gangways, weighing pens and sufficient replication of the livestock assigned to individual grazing treatments.

Regardless of size it is essential to provide soil analyses before the treatments are applied. The purpose is to assess the adequacy, surplus or deficiency of available nutrients for crop growth. A standard soil analysis package measures soil pH and estimates available concentrations of phosphorus, potassium and magnesium. Some minor elements such as boron and copper can also be measured using soil samples, while others, for example manganese, are usually assessed from plant samples. For nutrient and pH assessment, soil is usually removed from the top 10–20 cm. Although it may be valuable to study nutrient levels in deeper profiles, it is often backbreaking work if traditional soil augers are used. Mechanical soil sampling equipment is available, but this is more expensive than hand-operated augers and rarely available at Experimental Stations in developing countries.

The number of soil samples must be sufficiently large to be representative of the experimental site. Within each site samples should be taken across a W pattern; 6–10 separate samples being taken along each arm of the W. Samples are usually bulked into one bag. The bulked sample comprises around 1 kg of soil which is taken to represent an entire site containing approximately 2000 t soil/ha to a ploughing depth of 20 cm. Clearly, using a small bulked sample is not a precise measure of nutrient or acidity levels and the problem of accuracy is increased by the use of very small soil subsamples for nutrient analysis. Although the researcher may have invested a great deal of time and physical effort in order to obtain a representative soil sample, the amount of soil used by the analytical chemist is often tiny. For example, when determining ammonium nitrate and nitrite nitrogen levels from fresh soil, a subsample of 20 g soil is used. For potassium, magnesium, sodium and manganese only 5 g of dried soil is required when using ammonium acetate extraction techniques. As a result it is important to assess the number of subsamples in relation to the variation between replicates analyses. If the level of variation between three or four replicates is high additional subsamples must be analysed, although for some analyses such as sulphate-sulphur this decision will be expensive. Unfortunately, all soil laboratory procedures are time consuming and costly, but this must not deter the researcher from clearly defining the pH, soil texture and nutrient status of the experimental site prior to drilling.

1.5 SATELLITE MAPPING

The careful control of inputs or precision farming is not new, and many farmers have selectively applied some agrochemicals for many years using their local knowledge of individual fields. For example, patch spraying with graminicides or the application of lime to sections of fields or headlands can often be cost effective.

New technology has been developed which may supersede conventional soil-sampling procedures for assessing the causes of yield variations that occur within most large arable fields. It is based on satellite-controlled navigation systems or GPS (Global Positioning Satellites). GP yield monitoring systems can be fitted to combine harvesters, with up to 500 grain weight checks/ha during harvest. The computer system can create a colour-coded yield map of each field. This shows yield variations across large areas, which would have been difficult to detect from auger samples. It has resulted in the development of management systems for the precise application of nutrients, and as a result manufacturers have now launched computer-controlled fertiliser spreaders.

The success of satellite precision farming systems will ultimately be controlled by their cost effectiveness, and the reliability of microchips linked to machinery which is designed to work under field conditions. For experimentation, it must be remembered that the most common reason for yield variations within a field is soil type. Boundaries between different soil textures were formerly defined by hedgerows, thus it may be easier to consult historical farm maps before using satellite technology. At present this exciting technology is of little value for choosing the position of an experimental site within a field. Background variation is still best defined by conventional soil analysis and historical research, and ultimately controlled by the correct choice of experimental design.

1.6 SAMPLING

A crop sample is a small portion of a population taken for detailed study. It may be a length of row, quadrat area, a number of plants or pots taken at random. Hopefully, it will be representative (large) enough to inform the researcher what he or she needs to know about the whole population of a particular treatment.

Some crop-sampling procedures provide excellent exercises for improving physical fitness! Hand lifting, bagging and labelling samples of potatoes and sugar beet requires stamina on hot summer days. Locating and then removing quadrat areas of oilseed rape from dense, tangled and lodged crops is a challenge of patience and technical skill for the research agronomist. It is important to prepare bags and labels before going into the field, and to check that each plot has been allocated a sample bag (using the field plan) before starting work. Luggage labels made from cardboard are ideal — if felt tip pens are used check that the ink is waterproof. If cold storage facilities are not available, it is best to process samples block by block — assuming a block design has been used (Chapter 10). This minimises problems associated with wilting and loss of dry matter.

As statistical packages are now readily available, it is tempting to carry out data analyses before examining the form of distribution that is associated with the sampled data. This examination is highly recommended. Crop experimentation encompasses both discrete and continuous distributions. Variables such as number of branches, number of pods per plant and number of seeds per pod have discrete distributions, while those such as plant height and dry weight have continuous distributions (Chapters 2 and 3).

Sampling schemes should only be agreed after a careful assessment has been made of plant establishment. They must avoid bias, but at the same time take into account variability in establishment possibly caused by poor drill calibration. The number of plants in individual rows of adjacent plots of winter wheat is shown in Table 1.1.

There are large differences between rows within each plot with rows 5 and 3 having the lowest plant populations. The position of these two rows varied according to the direction of drilling, the problem having mainly arisen due to shallower coulter depth. In this experiment the background variation in establishment was high, and as a result a random sampling system based on a length of one row would have been inappropriate. Instead, a quadrat area encompassing eight rows was randomly removed on each sampling occasion following establishment. It is always important to determine the reliability of the sampling system by examining the level of variation in plant establishment between replicate samples taken from the same treatment.

The morphology of individual crop plants varies widely in an unevenly established crop. Between-plant variation (plot background variation) is especially problematic to sampling procedures in poorly established precision-sown or transplanted crops such as maize and tobacco. Sampling difficulties encountered in many field experiments are clearly illustrated by data obtained from two experimental plots of winter oilseed rape. Quadrat samples (0.33 m^2) were randomly taken when the lowermost terminal raceme buds were yellowing. The individual dry weights of all plants were recorded and the background variation is summarised in Table 1.2.

Table 1.1. The number of winter wheat plants in six plots and individual rows (replicate 1) of 0.5 m prior to the application of husbandry treatments

	Plot Number					
	↓	↑	↓	↑	↓	↑
Row	1	2	3	4	5	6
1	45	73	46	73	55	72
2	35	47	31	54	40	41
3	37	**23**	27	**24**	28	**19**
4	28	18	32	29	29	25
5	**14**	52	**10**	57	**7**	45
6	28	70	25	71	23	56
7	37	71	38	79	41	79
8	39	76	54	90	56	84

Notes: Each plot measured 12 × 1.8 m and contained 10 rows—the inner 8 rows were used for sampling. Plants in 0.5 m row lengths were counted.
↑ ↓ = direction of drilling.

Table 1.2. Ranges of plant dry weights (g)

	Quadrat plant number	Mean plant weight (g)	Range of plant weight
Treatment 1 (+128 kg N/ha)	22	7.81	0.36–16.65
	45	5.95	0.30–16.00
	63	4.36	0.31–13.12
	39	5.21	0.39–20.09
	31	6.73	0.57–17.80
	32	8.35	0.76–27.51
	36	5.11	0.49–14.87
	37	6.34	0.24–19.53
Means	38.1	6.23	
Treatment 2 (+236 kg N/ha)	50	7.79	0.60–26.47
	33	7.34	0.48–19.71
	39	6.34	0.83–34.41
	28	9.13	0.73–31.85
	48	9.26	0.55–26.29
	26	6.56	1.35–29.84
	26	8.58	0.40–17.46
	28	12.22	0.74–36.30
Means	34.8	8.40	

For both nitrogen fertiliser treatments the numbers of plants varied widely, and each quadrat contained a number of very small and large specimens. Possible treatment differences may not have been detected if only small plants had been subsampled from each quadrat for growth analysis. It is essential to use a subsampling system which eliminates bias because in practice there is always a tendency to avoid large plants as their subsequent analysis in the crop laboratory is time consuming.

The wide range of plant-to-plant variation shown in Table 1.2 cannot be quantified accurately if only small numbers of plants are studied. Yet subsample sizes in many research papers on rapeseed agronomy and physiology have only consisted of 3–20 plants per treatment. A detailed statistical study of data given in Table 1.2 indicated that for the sample mean plant weight to be within 1 g of the true weight (with 95% confidence), a random sample of approximately 600 plants would be required! As this is time consuming it again raises the question of the value of overcollecting descriptive data unless the accuracy of a sampling system is known. In System 2 it may be better to minimise the time invested in crop description (traditional growth analysis), in order to study in more detail environmental and physiological parameters affecting canopy development and crop yield. The latter approach (crop modelling), has been made more accessible with the availability of field recording equipment which is directly linked to computers.

Although statisticians will always recommend random sampling, this may not be practical in tall, high-density plots of crops such as oilseed rape without causing damage to surrounding areas. In this situation it may be wiser to adopt a 'step-ladder' sampling system in which a quadrat enclosing inner plot rows is first removed from a uniform area at either the top or bottom of each experimental plot. After leaving a discard distance a second quadrat can be removed on a later date using the first sampled area as a working base. This process can then be repeated. An unsampled, undisturbed area of plot must be left for commercial yield assessment. However, the importance of selecting the sampling units in such a way that they shall be as representative as possible of the entire population cannot be overemphasised.

Chapter 2

Basic Statistical Calculations

2.1 INTRODUCTION

When all statistical analyses were carried out using hand-operated calculators, the amount of experimental data collected and analysed was partly controlled by the sluggishness of early desk machinery. Because each analysis may have taken many minutes or even hours to complete, raw data were pondered in detail before calculations were attempted.

Package computer programs have revolutionised attitudes to data-collection and data-handling systems in applied biological research. Sadly, they also seem to have diluted many students' understanding of basic statistics. It is now far too easy to rely uncritically on computer output and to carry out sophisticated analyses which may be inappropriate and lead to misleading conclusions.

Computer technology has greatly improved presentational but not interpretative skills. For example, during oral examinations many students are unable to explain basic statistical terms such as standard error and variance, even when exquisitely presented tables and figures created by computer technology include summaries of statistical tests. Their attitude now seems to be 'don't think, use the computer', and if the output looks good then include it in the dissertation to impress the examiner!

The main objective of this chapter is to provide a definition and clear understanding of some basic statistical terms which are commonly used when analysing data collected from field and glasshouse experiments. For simplicity, only a small number of observations is included in the analyses so that the reader can check the results using a hand calculator.

2.2 MEASUREMENTS AND TYPE OF VARIABLE

The unit on which measurements are made may be a whole plot, a small area of a plot, a single plant, a stem, a leaf, etc. Suppose the experimental unit is an individual plant. For each, measurements may be made on several variables, such as height, weight, leaf area or number of internodes. Variables may be discrete, continuous or categorical.

A **continuous** variable is one that can take any value in a certain range. For instance, plant height is a continuous variable. If one plant has a height of 20 cm and another a height of 21 cm, it is possible to find a third plant with a height of between 20 and 21 cm. For continuous variables, measurements are approximate because they have to be rounded off to a whole number or to a fixed number of decimal places.

A **discrete** variable is one which can only take certain values. An example is the number of seeds in a pod. This number must be an integer such as 0, 1, 2, 3, etc. We cannot have a pod with 2.1 seeds.

A **categorical** variable is formed when data are classified into categories. For example, each plant measured could be classified according to variety. In this case variety is a category, sometimes called a **classification variable**. If the varieties are given names, there is no natural order. If there were three varieties, we could assign the numbers 1, 2 and 3 to them, but it would be meaningless to do any calculations on these numbers. However, it may be meaningful to count the number of plants of each variety. These data may be summarised in a table or a bar chart.

2.3 SAMPLES AND POPULATIONS

One of the main objectives of statistical analysis is to find out as much as possible about a population. Most populations are far too large to be measured. For example, suppose the population under study is a field of wheat. You may want to know the average yield per plant. As resources are not available to measure every wheat plant, a random sample can be taken. The average yield of these plants is an estimate of the mean yield of all plants in the field. The estimate calculated could be ‘a long way’ from the true value. A statement is needed of how close the sample mean is likely to be to the population mean. For example, it would be helpful to state with 95% confidence that the mean yield per plant lies between 25 and 30 g. The calculation of a **confidence interval** requires some background theory, and in the following discussion a population of field plants is assumed.

N = population size which may be the total number of wheat plants in the field. It is likely to be very large and unknown. For example, a farm crop of wheat in the UK will consist of approximately 250 plants/m^2, or 2.5 million plants per hectare.

μ = the **population mean**. This is rarely known. It may be the mean yield per plant of all the plants in the field. If all the plants were measured, μ would be calculated by adding up all the yields and dividing by N. The formula for μ is

$$\mu = \frac{\Sigma x}{N}$$

x is the symbol used for yield, and Σ is the summation sign. It means add up all the xs (the yields).

As it would be impractical to assess the yield of 2.5 million individual plants, an estimate of μ can be found by taking a sample from the population. To be unbiased and representative of the population, the plants to be chosen for inclusion in the sample must be selected at random. In this way all individuals in the population have an equal chance of being included in the sample.

n = sample size (this may be the number of plants included in the sample)
x_1 = yield of first plant in the sample
x_2 = yield of second plant in the sample
x_n = yield of nth plant in the sample.

Using the sigma notation, $\Sigma x = x_1 + x_2 + \ldots x_n$

$\bar{x}$ is the symbol for **sample mean** and its formula is

$$\bar{x} = \frac{\Sigma x}{n}$$

Example 2.1

The heights of a random sample of 5 plants are 14.8, 15.2, 17.4, 11.6 and 12.5 cm. The mean height is

$$\bar{x} = \frac{14.8 + 15.2 + 17.4 + 11.6 + 12.5}{5} = 14.30 \text{ cm}$$

The mean is a measure of location or central tendency. Another measure of location is the median.

2.3.1 Median

If there are n numbers, the median is the $(n + 1)/2$ ranked number. If n is odd, this is the middle number after sorting them in order of magnitude, and if n is even it is the average of the middle two.

The data of the last example after sorting are: 11.6, 12.5, 14.8, 15.2 and 17.4. The median is therefore 14.8. If 13.1 is added, the median is the average of 13.1 and 14.8, namely 13.95.

The median is preferred to the mean when the distribution is very *skew* (non-symmetrical). For instance, if 17.4 is replaced by 37.4 in the original data set, the median is still 14.8, but the mean is 18.3.

The distribution is *positively skewed* when there is a small proportion of unusually high values which normally results in the mean being larger than the median. The distribution is *negatively skewed* when there is a small proportion of unusually low values which normally results in the mean being smaller than the median.

2.3.2 Population Variance

The population variance is denoted by σ^2 and it is the average of the squared deviations from the population mean. It is a measure of the variation in the values and the formula is

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

It cannot be calculated because it is impossible to measure all the x values. It is estimated by calculating the sample variance.

2.3.3 Sample Variance

An unbiased estimator of the population variance is the sample variance, denoted by s^2. Its formula is

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

where $\bar{x}$ is the sample mean and n is the number of sample observations.

To calculate s^2, we find the sum of the squares of the deviations from the sample mean and divide by the **degrees of freedom** ($n - 1$). Division by n would give a biased estimate of σ^2. The sample variance is used in hypothesis testing and in calculating confidence intervals (Chapters 4 and 5).

Example 2.2 Sample variance for plant heights in Example 2.1
Table 2.1 shows the original heights, their deviations from the sample mean and the squares of these deviations. Recall that the mean is 14.30 cm.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{71.5}{5} = 14.3 \quad \text{and} \quad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{21.20}{4} = 5.30$$

Notice that $\Sigma(x - \bar{x}) = 0$. It is always true that the sum of the deviations from the sample mean add to zero.

2.3.4 Degrees of Freedom

Only $n - 1$ of the deviations are free to vary. Once the sample is taken, the sample mean is fixed. If $n - 1$ of the deviations from the sample mean are calculated, the nth deviation is fixed, as all n deviations must add to zero. As a result, there are $n - 1$ degrees of freedom (df) associated with this estimator of population variance. If the population mean was known and substituted for $\bar{x}$ in the formula for s^2, there would be n df because in this situation where $n - 1$ of the deviations from μ are known, the nth deviation from μ cannot be predicted. The sum of the deviations of the n sample values from the sample mean is zero, but the sum of the deviations of the n sample values from the population mean is not zero.

Exercise 2.1
Use the above method to find the sample variance of the numbers 13.1, 16.4, 19.5, 22.0, 25.5, 18.7. The answer is 18.58.

Table 2.1. Corrected sums of squares for Example 2.2

	x	$(x - \bar{x})$	$(x - \bar{x})^2$
	14.8	0.5	0.25
	15.2	0.9	0.81
	17.4	3.1	9.61
	11.6	−2.7	7.29
	12.5	−1.8	3.24
Total	71.5	0	21.20

2.3.5 Corrected Sum of Squares

A measure of variation which is frequently used in later chapters is the corrected sum of squares. This is the sum of the squares of the deviations from the sample mean: it is denoted by Sxx and its formula is

$$Sxx = \Sigma(x - \bar{x})^2$$

The sample variance can thus be written as $s^2 = Sxx/(n - 1)$ and Sxx can be written as $Sxx = (n - 1)s^2$.

For Example 2.2, $Sxx = 21.20$

2.3.6 The Computational Formula for *Sxx*

When calculating the deviations from the sample mean by hand, rounding-off errors will occur if $\bar{x}$ is not recorded to a sufficient number of decimal places. If a large number of decimal places are used, the calculations become tedious. This problem can be avoided by using the following alternative formula:

$$Sxx = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

Using this version, the corrected sum of squares is the uncorrected sum of squares minus (the square of the sum divided by n). $(\Sigma x)^2/n$ is called the correction factor and denoted by CF.

Example 2.3

Now recalculate s^2 for Example 2.2 using the correction factor method. Table 2.2 shows the details.

$$\begin{aligned} Sxx = \Sigma x^2 - \frac{(\Sigma x)^2}{n} &= 1043.65 - \frac{(71.5)^2}{5} \\ &= 1043.65 - 1022.45 \\ &= 21.20 \text{ as found earlier} \end{aligned}$$

Table 2.2. Sums of squares for Example 2.2

	x	x^2
	14.8	219.04
	15.2	231.04
	17.4	302.76
	11.6	134.56
	12.5	156.25
Total	71.5	1043.65

2.3.7 Standard Deviation

The standard deviation is the square root of the variance and is measured in the original units. If the x values are measured in cm, the variance is in cm^2. As this is a difficult term to work with, the problem is removed by taking the square root. Thus, the standard deviation is in cm.

Population standard deviation $= \sigma$
Sample standard deviation $= s = \sqrt{s^2}$

For Example 2.2, $s = \sqrt{5.30} = 2.302$ cm

For most distributions which are fairly symmetrical, about 95% of the population lies within two standard deviations of the mean.

2.3.8 The Coefficient of Variation (CV)

The CV is the standard deviation expressed as a percentage of the mean. It is independent of the units of measurement.

For the population $CV = \frac{\sigma}{\mu} \times 100\%$

For a sample $CV = \frac{s}{\bar{x}} \times 100\%$

For Example 2.2, $CV = \frac{2.302}{14.3} \times 100\% = 16.10\%$

The concept of coefficient of variation can be better understood by considering the following two data sets:

I: 2.1 3.5 4.7 5.2 6.4
II: 102.1 103.5 104.7 105.2 106.4

They both have the same variation. You should verify that their sample standard deviations are both 1.645. However, the variation within set I is very large in relation