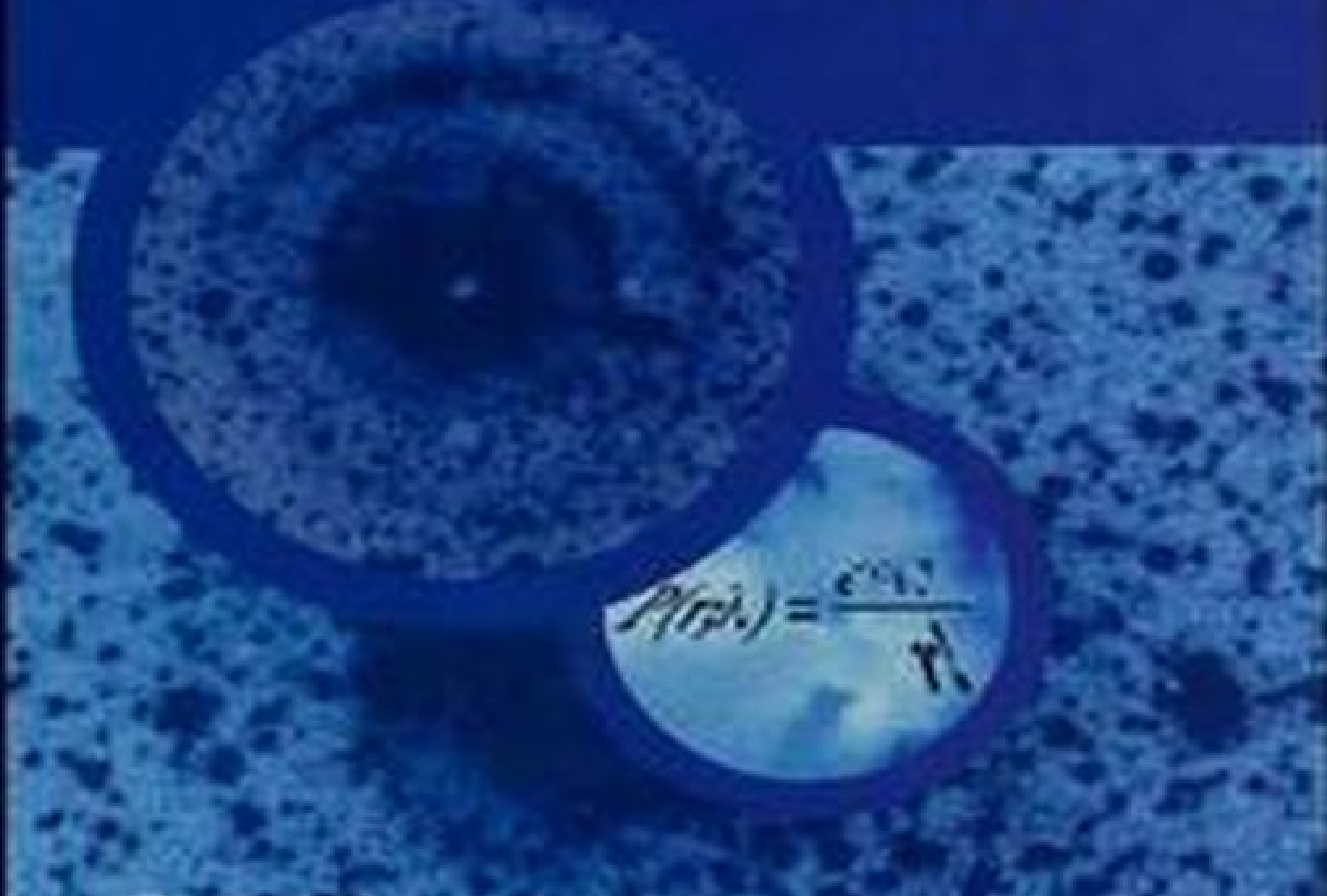
Statistics

A Guide to the Use of Statistical Methods in the Physical Sciences

R. J. Barlow



WILEY

Contents

The Manchester Physics Series

General Editors

D. J. SANDIFORD: F. MANDL: A. C. PHILLIPS

Department of Physics and Astronomy, University of Manchester

Properties of Matter:	B. H. Flowers and E. Mendoza
Optics: *Second Edition*	F. G. Smith and J. H. Thomson
Statistical Physics: *Second Edition*	F. Mandl
Electromagnetism: *Second Edition*	I. S. Grant and W. R. Phillips
Statistics:	R. J. Barlow
Solid State Physics: *Second Edition*	J. R. Hook and H. E. Hall
Quantum Mechanics:	F. Mandl
Particle Physics: *Second Edition*	B. R. Martin and G. Shaw
The Physics Of Stars: *Second Edition*	A. C. Phillips
Computing For Scientists:	R. J. Barlow and A. R. Barnett

STATISTICS

A Guide to the Use of Statistical Methods in the Physical Sciences

Roger Barlow

Department of Physics
Manchester University

John Wiley & Sons

CHICHESTER NEW YORK BRISBANE TORONTO SINGAPORE

Baffins Lane, Chichester
West Sussex PO19 1UD, England

Reprinted December 1991, February 1995, July 1996, October 1997, July 1999

Other Wiley Editorial Offices

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

Jacaranda Wiley Ltd, G.P.O. Box 859, Brisbane,
Queensland 4001, Australia

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

John Wiley & Sons (SEA) Pte Ltd, 37 Jalan Pemimpin 05-04,
Block B, Union Industrial Building, Singapore 129809

Library of Congress Cataloging-in-Publication Data

Barlow, Roger (Roger J.)
Statistics: a guide to the use of statistical methods in the physical sciences.

(The Manchester physics series)
Bibliography: p.
Includes index.
1. Statistical physics. I. Title. II. Series.

QC174.8.B365 1989 530.1′595 88-33908

ISBN 0 471 92294 3

ISBN 0 471 92295 1 (pbk.)

British Library Cataloguing in Publication Data

Barlow, Roger
Statistics: a guide to the use of statistical methods in the physical sciences.
1. Statistical mathematics
I. Title II. Series
519.5

ISBN 0 471 92294 3
ISBN 0 471 92295 1 pbk

To my father

Editors’ preface to the Manchester Physics Series

The Manchester Physics Series is a series of textbooks at first degree level. It grew out of our experience at the Department of Physics and Astronomy at Manchester University, widely shared elsewhere, that many textbooks contain much more material than can be accommodated in a typical undergraduate course; and that this material is only rarely so arranged as to allow the definition of a shorter self-contained course. In planning these books we have had two objectives. One was to produce short books: so that lecturers should find them attractive for undergraduate courses; so that students should not be frightened off by their encyclopaedic size or their price. To achieve this, we have been very selective in the choice of topics, with the emphasis on the basic physics together with some instructive, stimulating and useful applications. Our second objective was to produce books which allow courses of different lengths and difficulty to be selected, with emphasis on different applications. To achieve such flexibility we have encouraged authors to use flow diagrams showing the logical connections between different chapters and to put some topics in starred sections. These cover more advanced and alternative material which is not required for the understanding of latter parts of each volume.

Although these books were conceived as a series, each of them is self-contained and can be used independently of the others. Several of them are suitable for wider use in other sciences. Each Author’s Preface gives details about the level, prerequisites, etc., of his volume.

The Manchester Physics Series has been very successful with total sales of more than a quarter of a million copies.

We are extremely grateful to the many students and colleagues, at Manchester and elsewhere, for helpful criticisms and stimulating comments. Our particular thanks go to the authors for all the work they have done, for the many new ideas they have contributed, and for discussing patiently, and often accepting, the suggestions of the editors.

Finally, we would like to thank our publishers, John Wiley & Sons Ltd, for their enthusiastic and continued commitment to the Manchester Physics Series.

D. J. Sandiford
F. Mandl
A. C. Philips
February, 1997

The generall end therefore of all the book is to fashion a noble person in vertuous and gentle discipline

—Edmund Spencer

Author’s Preface

Many science students acquire a distinctly negative attitude towards the subject of statistics. The reasons for this are clear. The traditional first year concentrated statistics course of derivations and exhortations makes little impact on the young undergraduates, who want to get to grips with the basic truths of their chosen subject and have no interest in sordid details like error bars. The hapless students then go to laboratory classes, in which their enjoyment of the experiments is marred by the awful chore of the ‘error analysis’ at the end, where, whatever they do, they inevitably get harshly criticised for doing it wrong. Under such circumstances, ‘statistics’ can soon become a collection of meaningless ritual, to be gone through correctly if harsh words and bad marks are to be avoided.

As a student I was no different from any other in this respect. But later, in the real world, doing real experiments, statistics began to matter. Over the years I got to grips with the subject, by talking to colleagues and digging in reference books, and was agreeably surprised to discover that it had an internal logic and structure. Once one really got into it, it made sense. Eventually the time came when people started asking me questions, and I somehow acquired a reputation as the local statistics expert. On this basis I devised a course, which was given as a set of lectures to students at Manchester University. This has convinced me that statistics can be taught to students in such a way as to make it interesting for them, and give them a real grasp of the subject.

This book has grown out of the lecture notes given out with the course. Despite the shelves full of books on ‘statistics’ in any library or university bookshop, there is a desperate lack of any suitable textbook for the physical sciences beyond the very elementary level. The books available are mainly aimed at the biological and social sciences; for those of us in other fields they are inappropriate, both in content and treatment. They deal largely with samples and surveys, and the problems of hypothesis testing, whereas we are more concerned with the theory of measurements and errors, and with the problem of estimation. Furthermore they assume, usually correctly, that those for whom they are intended (geographers, psychologists, and suchlike) will fear and loathe anything at all mathematical. They therefore avoid anything beyond (or even, in some cases, including) the most elementary algebra. Now, although physicists and chemists may fight shy of high-powered abstract mathematics, they can happily differentiate and integrate simple functions and follow basic algebra. They are thus entitled to a reasonable explanation of the mathematics involved in statistical calculations, and able to benefit from it. This book thus assumes a reasonable degree of numeracy from the reader, but nothing outstanding—any real mathematician will find it hopelessly naive and unrigorous.

This book is thus the textbook I would like to have had available, both as a student and when teaching students, and for my own use with real problems. I hope that others will find it useful and interesting, and that it will eventually lead them not only to use and understand statistics, but to enjoy it.

I would like to record my acknowledgements to the many people who, by discussions and advice, have helped form my ideas on the subject, to the students on my course for acting as guinea-pigs for the material, to John Ellison for many helpful comments in preparing the manuscript for publication, and finally to my wife Ann for putting up with the trials of a traumatic author with patience and understanding.

4 October 1988

ROGER BARLOW
Manchester

‘It’s *not the figures themselves,’, she said finally,* ‘it’s *what you do with them that matters’*

—K.A.C. Manderville

CHAPTER 1

Using Statistics

Statistics is a tool. In experimental science you plan and carry out experiments, and then analyse and interpret the results. To do this you use statistical arguments and calculations. Like any other tool—an oscilloscope, for example, or a spectrometer, or even a humble spanner—you can use it delicately or clumsily, skilfully or ineptly. The more you know about it and understand how it works, the better you will be able to use it and the more useful it will be.

The fundamental laws of classical science do not deal with statistics or errors. Newton's law of gravitation, for example, reads in pure and beautiful simplicity. The figure in the denominator is given as 2—exactly 2, not 2.000 ± 0.012 or anything messy like that. This can lead people to the idea that statistics has nothing to do with 'real' scientific knowledge.

$$F = \frac{GMm}{r^2}$$

But where do the laws come from? Newton's justification came from the many detailed and accurate astronomical observations of Tycho Brahe and others. Likewise Ohm's law

$$V = IR$$

which appears so straightforward and elementary to us today, was based on Ohm's many careful measurements with primitive apparatus. When you are *studying* science you may find no use for statistics—until you meet quantum mechanics, but that is another story—but as soon as you begin *doing* science, and want to know what measurements really mean, it becomes a matter of vital importance.

This is a textbook on statistics for the physical sciences. It treats the subject from the basic level up to a point where it can be usefully applied in analysing real experiments. It aims to cover most situations that are likely to be met with, and also provide a grasp of statistical ideas, terminology, and language, so that more advanced works can be consulted and understood should the need arise. It is thus intended to be usable both as a textbook for students taking a course in the subject, and also as a handbook and reference manual for research workers and others when they need statistical tools to extract their experimental results.

These two modes of use give rise to requirements in the ordering of the material which are not always happily reconcilable. For reference use one wants to group all material on a given topic together, but for teaching purposes this would be like learning a language from a dictionary. The solution adopted is that the unstarred sections cover the material roughly appropriate to a first year undergraduate course. They can sensibly be taken in order, with no anticipation of later material. The starred sections fill in the gaps; they may require knowledge of material in later sections, but when this occurs it is explicitly pointed out. Most of the basic material is in the early chapters, and Chapters 7, 9, and 10 contain entirely higher-level material. First-time-through readers should not be scared or put off by any apparent mathematical complexity they observe in some of the starred sections: these can (and should) be skipped over with a clean conscience, as they are not needed for later unstarred sections of the course.

'Data! Data! Data!', he cried impatiently.
'I can't make bricks without clay'.

—Sir Arthur Conan Doyle

CHAPTER 2

Describing the Data

It all starts with the data. You may call them a *set of results,* or a *sample* or the *events*, but whatever the name, they consist of a set of basic measurements from which you're trying to extract some meaningful information.

To make your data mean something, particularly to an outside audience, you need to display them pictorially, or to extract one or two important numbers. There are many such numbers and ways of presenting the data in graphic form, and this chapter is devoted to methods of describing the data in a useful and meaningful way, without attempting any deeper analysis or inference. This is known as *descriptive statistics.*

2.1 TYPES OF DATA

Data[†] are called *quantitative* or *numeric* if they can be written down as numbers, and *qualitative* or *non-numeric* if they cannot. Qualitative data are rather hard to work with as they do not offer much scope for mathematical treatment, so most of the subject of statistics, and likewise most of this book, deals with quantitative, numerical measurements.

Quantitative measurements divide further into two types. Some, by their very nature, have to be integers and these are called *discrete* data. Others are not constrained in this way and their values are real numbers. These are called *continuous* data. Continuous data cannot be recorded exactly, as you cannot write down an infinite number of decimal places. Some sort of *rounding* and loss of precision has to occur.

For example, if you were to examine a sample of motor cars and record their colours, these would be qualitative data. The number of seats in each car has to be an integer, and would be discrete numeric data, as would the number of wheels. The lengths and the weights of the cars would be continuous numeric data.

Usually one of the first things to do in making sense of the data (which is just a pile of raw results) is to divide them into *bins* (also called *groups* or *classes* or *blocks).* For example, the results of tossing 20 coins, each of which comes down either heads (H) or tails (T)

$$\{\mathbf{H}, \mathbf{T}, \mathbf{H}, \mathbf{H}, \mathbf{T}, \mathbf{H}, \mathbf{T}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{T}, \mathbf{T}, \mathbf{H}, \mathbf{T}, \mathbf{T}, \mathbf{H}, \mathbf{T}, \mathbf{H}, \mathbf{H}, \mathbf{T}\}$$

can be written as {11H,9T}. This conveys the same information much more clearly and concisely.

For continuous numeric data it is not quite so simple, as your values are (almost certainly) all different, if you use enough decimal places. You have to group together adjacent numbers, using a range of values to define each bin. This means further rounding of values and throwing away precision information, which is the price you pay for rendering the data comprehensible. Usually the bins are chosen to be all the same uniform size, but in some cases it makes sense to use non-uniform bins of different sizes.

For discrete numeric data this grouping together of adjacent values is not compulsory, but it may be desirable when the numbers of data points with any particular value are small.

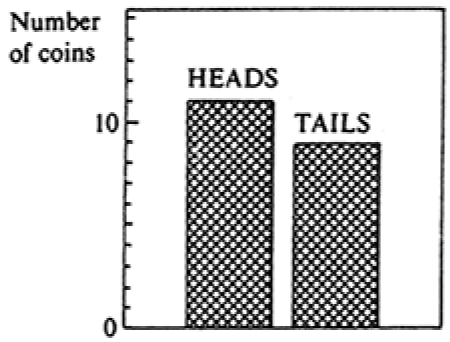
2.2 BAR CHARTS AND HISTOGRAMS

The numbers of events in the bins can be used to draw bar charts (see Figure 2.1) and histograms.

There is a technical difference between a bar chart and a histogram in that the number represented is proportional to the *length* of bar in the former and the *area* in the latter. This matters if non-uniform binning is used. Bar charts can be used for qualitative or quantitative data, whereas histograms can only be used for quantitative data, as no meaning can be attached to the width of the bins if the data are qualitative.

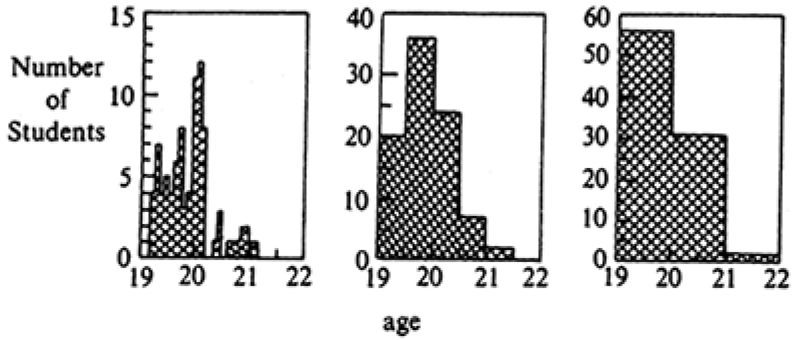
For quantitative, numeric, data, you have to choose the width of the bins to be used in the display (see Figures 2.2). This requires thought. If the bins chosen are too narrow, then there are very few events in each bin, and the numbers are dominated by fluctuations. Ideally there should be at least ten events in each bin, and the more the better. If they are too wide, then real detail can be obscured if the bin stretches over genuine variations in the distribution. Ideally, the difference between contents of adjacent bins should be small. The choice is yours—it is a matter of personal judgement. It may well be, particularly if the number of events is small, that there is no way of satisfying both ideal requirements. In this case you just have to do the best you can with the data available.

Fig. 2.1. A bar chart displaying the data in the previous section.



There are other ways of representing the data using pictures: ideographs, frequency polygons, pie charts, prismograms, scatter plots, and many more. However, it is not necessary to give you all the details. They are designed to be straightforward to understand, and are therefore straightforward to use. Some people become very excited about ‘right’ and ‘wrong’ ways of doing things, and come almost to blows over whether gaps between bars in a bar chart are compulsory or optional, and similar trivial matters. Such details are not really important. It is very much a matter of your own taste as to how you display your data, and the scales and axes you use. The object of any description is to convey an idea of your data to your audience in a way that is effective, easy to grasp, and honest. That is all that really matters. (For examples of dishonest methods, consult *How to Lie with Statistics* by Darrell Huff (see Bibliography) or any daily newspaper.)

Fig. 2.2. The ages (in years) of a group of second year students, showing the effects of choosing different bin sizes for the same data.



2.3 AVERAGES

2.3.1 The Arithmetic Mean

If you want to describe your data with just one number, the best and most meaningful one to use is almost certainly the arithmetic mean. This is denoted by an upper bar over the quantity concerned: thus if there are *N* elements in the set of data

$$\{x_1, x_2, x_3, \ldots, x_N\}$$

then the mean value of *x* is

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \quad (2.1)$$

In the same way you can calculate the mean value of any function *f(x):*

$$\bar{f} = \frac{1}{N}\sum_i f(x_i). \quad (2.2)$$

If the data have been binned, and bin *j* corresponds to a value X_j and contains n_j, data elements, then these means can be written

$$\bar{x} = \frac{1}{N}\sum_j n_j x_j \quad (2.3)$$

$$\bar{f} = \frac{1}{N}\sum_{j} n_j f(x_j).$$
(2.4)

Note carefully the apparent difference between these formulae and the ones above. Which you use depends on whether you are summing over *elements* of the data set or over *bins.* If rounding has occurred then the average from the bins is less accurate than the average over the unrounded elements, so it is better to use the unrounded data if you can.

Example Weights of foils

The weights of 5 metal foils are 25g, 24g, 27g, 29g, and 25g.

$$\text{The mean weight is therefore } \frac{25+24+27+29+25}{5} = 26\text{g}.$$

Example Occupancy of cars

In a survey of 100 cars passing a checkpoint, 72 contained only 1 occupant (the driver), 23 had 2, 2 had 3, and 3 had 4.

$$\text{Mean number of occupants per car} = \frac{72\times 1 + 23\times 2 + 2\times 3 + 3\times 4}{100} = 1.36.$$

2.3.2 Alternatives to the Arithmetic Mean

The *geometric mean* of two numbers is the length of the side of a square of area equal to the product of the two numbers. For *N* numbers it is defined as

$$\sqrt[N]{x_1 x_2 x_3 \cdots x_N}.$$

The *harmonic mean* is the reciprocal of the arithmetic mean of the reciprocals:

$$\frac{N}{1/x_1 + 1/x_2 + 1/x_3 + \cdots + 1/x_N}.$$

Pythagoras discovered that notes from strings whose lengths were in the ratio $1:\frac{1}{2}:\frac{1}{3}:\cdots$ were pleasing or ‘harmonic’. Hence for two numbers the harmonic mean is the intermediate value such that the three reciprocals are in arithmetic progression and the numbers belong to the sequence.

The *root mean square* is just what it says, i.e.

$$\sqrt{\frac{x_1^2 + x_2^2 + x_3^2 + \cdots + x_N^2}{N}}.$$

All of these are less common than the arithmetic mean, so that if the ‘mean’ is mentioned without a qualifying adjective, this refers to the arithmetic mean.

The *mode* is the most popular value in a set of data. It is easy to find, but it can be a misleadingly unrepresentative number to quote.

The *median* is the halfway point, in the sense that half the data elements fall below it, and half above. It is preferable to the mean in describing data where the order or *rank* in the variable is more important than the numerical value.

Actually, although the median is generally defined as the point with half the data below and half above, it is not really quite so simple. If the data have an odd number of elements, all with different values, then the median is taken as the middle one. This therefore has (N – l)/2 values above it and (N – l)/2 below, so there are in fact slightly less than half below and above. If there are several data points with the same value, perhaps because the data have been binned, then the best one can do to define a ‘central’ bin is to state that not more than half lie below it, and not more than half above, i.e. there will also be some at that value—and the numbers above and below may be different. If the number of elements is even there is a further complication if the two midmost points have different values, as then any number between them would satisfy the definition; the median is, by convention, taken as halfway between the two.

2.4 MEASURING THE SPREAD

2.4.1 The Variance

The mean $\bar{x}$ describes all your data with just one number. This can be useful, but it can also be misleading. Consider the two sets of (fictitious) data in Figure 2.3.

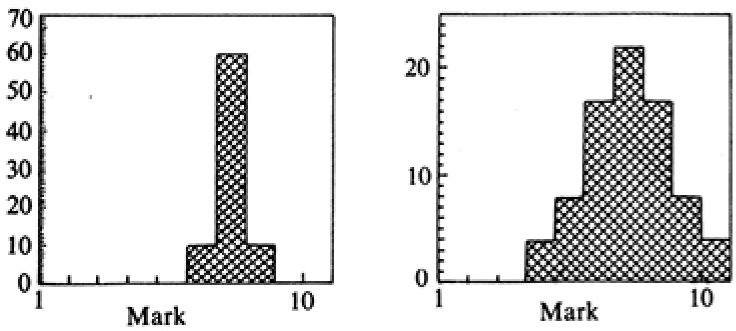
Both sets of marks have a mean of 7.0, but they differ greatly. The first assessor departs from the average only when the student is outstandingly good or bad, and then only by a small amount, but the second marks over a wider range. They are distributed very differently, and we need a number to express the spread or *dispersion* of the data about the mean.

The average deviation from the mean is not a useful quantity, as the positive and negative deviations cancel and the sum is automatically zero.

$$\begin{aligned}\frac{1}{N}\sum_i (x_i - \bar{x}) &= \frac{1}{N}\sum_i x_i - \frac{1}{N}\sum_i \bar{x} \\ &= \bar{x} - \bar{x} \\ &= 0.\end{aligned}$$

However, you can stop the contributions from different elements cancelling by squaring them, which forces them to be positive. Thus the average *squared* deviation from the mean is a sensible measure of the spread of the data. It is called the *variance* of *x* as it expresses how much x is liable to vary from its mean value $\bar{x}$, and is written *V(x):*

Fig. 2.3. Histograms showing the marks awarded by two demonstrators in assessing the performances of 80 students in the laboratory.



$$V(x)=\frac{1}{N}\sum_i (x_i-\bar{x})^2. \quad (2.5)$$

In the same way, any function of ***x*** has a variance too:

$$V(f)=\frac{1}{N}\sum_i (f(x_i)-\bar{f})^2. \quad (2.6)$$

The definition of *V(x)* can be manipulated to give a simpler formula for it. This type of manipulation is used so often that this time we will go through it in detail.

Starting from $V=\frac{1}{N}\sum_i (x_i-\bar{x})^2$

multiply out the square $=\frac{1}{N}\sum_i (x_i^2-2x_i\bar{x}+\bar{x}^2)$

separate into three sums $=\frac{1}{N}\sum_i x_i^2-\frac{1}{N}\sum_i 2x_i\bar{x}+\frac{1}{N}\sum_i \bar{x}^2$

extract some factors to get $=\frac{1}{N}\sum_i x_i^2-\frac{1}{N}2\bar{x}\sum_i x_i+\frac{1}{N}\bar{x}^2\sum_i 1$

which reduces to $=\overline{x^2}-2\bar{x}^2+\bar{x}^2$

and finally $=\overline{x^2}-\bar{x}^2.$

So the fundamental formula is obtained

$$V(x)=\overline{x^2}-\bar{x}^2 \quad (2.7a)$$

or, equivalently,

(2.7b) $$V(x) = \frac{1}{N}\sum x_i^2 - \left(\frac{1}{N}\sum x_i\right)^2$$

or, in words, the variance is the mean square minus the squared mean.

2.4.2 The Standard Deviation

The root mean squared deviation is called the *standard deviation* and given the symbol *σ*. It is just the square root of the variance (see previous section) and can be expressed in various equivalent forms (using equations 2.5 or 2.7):

(2.8a) $$\sigma = \sqrt{V(x)}$$

(2.8b) $$\sigma = \sqrt{\overline{x^2} - \bar{x}^2}$$

(2.8c) $$\sigma = \sqrt{\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2}$$

(2.8d) $$\sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2}.$$

σ represents a reasonable amount for a particular data point to differ from the mean. The exact numerical details depend on the case, but usually one is not surprised by data points one or two standard deviations from the mean, whereas a data point three or more *σ* away would cause a few raised eyebrows.

Broadly speaking, practical scientists like to work with *σ* rather than *V,* as it has the same units and dimensions as *x*. Statisticians, on the other hand, tend to use *V* as it is easier to manipulate. It does not really matter which you use, as it is trivial to translate from one to the other.

Example Laboratory marks

The narrow histogram in Figure 2.3 has 10 cases of 6 marks, 60 of 7 marks, and 10 of 8 marks. The gives a mean of 7, a variance of 0.25, and thus a standard deviation of

0.50 marks. The broad one has a standard deviation of 1.46 marks, nearly three times larger.

Example Monitoring

A company produces ball-bearings whose mean mass is 30 grams, with a standard deviation of 0.1 gram. Quality control inspectors check the production line by weighing a ball-bearing every morning. If its mass lies between 29.8 and 30.2 grams they assume all is well. If it is outside these 2σ limits—the 'warning level'—but within 29.7 and 30.3 grams they weigh some more ball-bearings. If it is outside the 3σ limits of 29.7 and 30.3 grams—the 'action level'—they halt the production line.

2.4.3 Different Definitions of the Standard Deviation

The definition of σ is a minefield of alternatives, and to call it the 'standard' deviation is something of a sick joke. It is important to face up to this, for when people are unaware of the differences between the definitions they get confused and dismayed by factors of $\sqrt{N/(N-1)}$ that appear apparently out of nowhere. This leads to a tendency to insert such factors at random and generally incorrect moments.

Equation 2.8 defined the standard deviation of a data sample as

$$\sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2}.$$

So far so good. However, our data are presumably taken as a sample from a parent distribution,[†] which has a mean and a standard deviation, denoted μ and σ. In terms of expectation values:

$$\mu = \langle x \rangle$$

(2.9) $$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}.$$

There is thus a clear distinction between $\bar{x}$, the mean of the sample, and μ, that of the parent, and complete confusion between σ, the standard deviation of the sample, and σ, that of the parent.

This is not really too bad, as it is generally clear which is meant. However, it gets worse. Some authors define the term 'standard deviation' as the r.m.s. deviation of the data points from the 'true' mean μ, rather than the sample mean $\bar{x}$:

$$\sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2}. \quad (2.10)$$

This is felt to be a more fundamental and 'truer' quantity than that defined in equation 2.8, but it is not much use if you do not know the value of μ. However, an estimate of this, which (when squared) gives an unbiased estimate of σ^2 of the parent, is given by

$$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}. \quad (2.11)$$

This is fair enough, and is considered in detail in Chapter 5, but we now have four definitions of 'standard deviation', three of which (equations 2.8, 2.10, and 2.11) are to some extent rivals.

The reason for all this regrettable mess is a chicken- and -egg argument as to which comes first (i.e. is more fundamental), equation 2.8 or 2.9. One school of thought says that there is a real, true, ideal distribution with a standard deviation defined by 2.9, which is best measured by 2.10 or, failing that, 2.11, and the standard deviation of your sample, as defined by 2.8, has no real significance. The opposing view, to which I incline myself, is that equation 2.8 is what you actually measure, and from a descriptive point of view that is that; any further developments towards properties of the parent distribution come under the