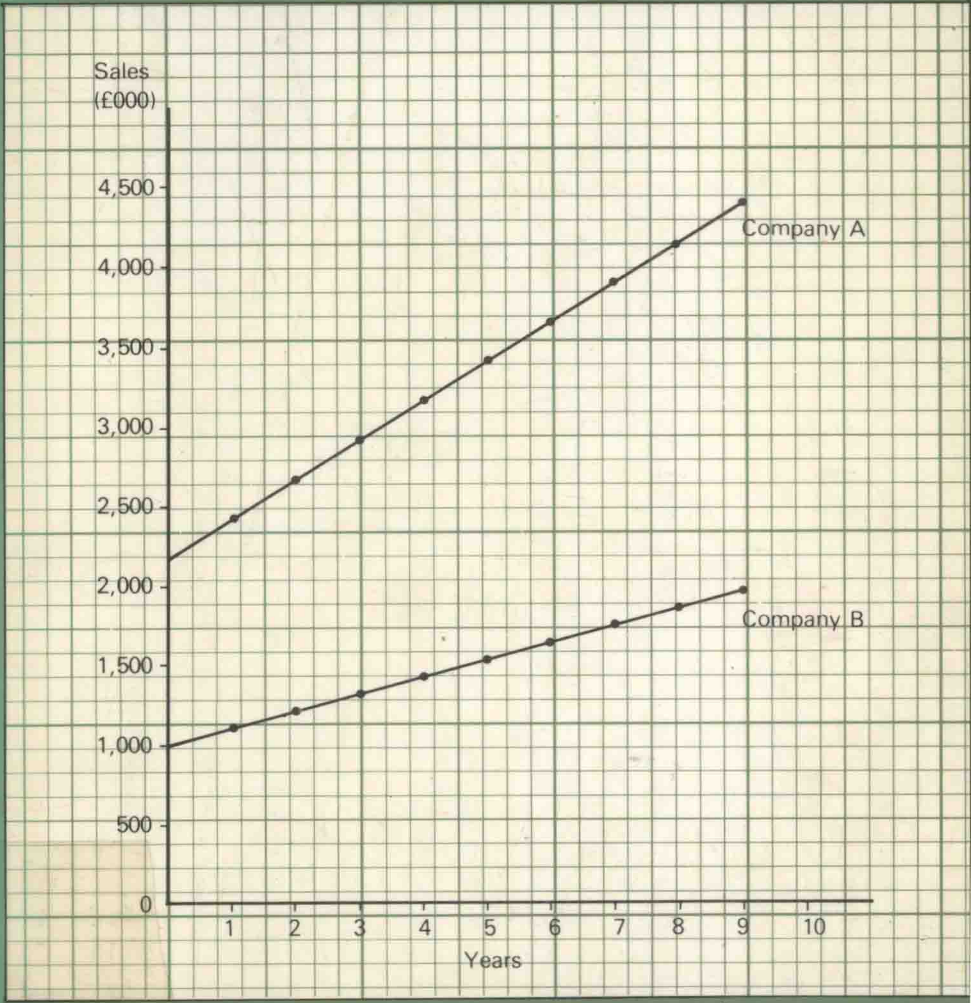
Statistics

Second Edition

Frank Owen and Ron Jones



Statistics

Second Edition

Frank Owen
Ron Jones

Pitman

PITMAN PUBLISHING LIMITED
128 Long Acre, London WC2E 9AN

A Longman Group Company

Second edition published in Great Britain 1982
Reprinted 1986

ISBN 0-273-02639-9

Produced by Longman Group (FE) Ltd
Printed in Hong Kong

Preface to First Edition

Recent years have seen many changes in both the content and style of examinations of the standard which used to be called 'Introduction to Statistics'. These changes reflect the growing belief that it is not enough to be able to perform statistical calculations. Important as it s to have some degree of calculative ability, this alone is no longer sufficient to ensure a pass in the examination room. More and more it is becoming necessary for the student to understand, not only what he is doing, but also the meaning of the results he obtains.

In surveying the literature, it seems to us that although there are many excellent books on statistical method there is a marked deficienty of textbooks designed to cover the present first-year syllabus of the major professional bodies. It is this gap that this book is designed to fill. We hope that it will prove to give full coverage of the Ordinary National Certificate Statistics examinations in both Business Studies and in Public Administration, the examination of the Institute of Certified Accountants, and the Institute of Secretaries and Administrators. Additionally it covers the content of the Statistics section of the Institute of Cost and Management Accountants paper in Mathematics and Statistics, and should prove useful for large parts of the Quantitative Methods paper of the Institute of Public Finance and Accountancy.

Every author knows that his book is never entirely his own work. We owe a great deal to the comments and criticisms of our colleagues at Liverpool Polytechnic; we believe that all students using this book will have so willingly allowed us to use their past examination questions. We are deeply appreciative of the encouragement and practical help of our publishers without which this work might never have seen the light of day.

The debt we owe to previous writers is beyond measure. If we have not acknowledged every single one it is because their ideas are so much a part of our own that it is impossible to identify with certainty what is theirs. We hope that each one of them will accept our acknowledgement of their contribution to the existing state of knowledge.

Such merits as this book may have are direct results of the help we have received from these and from many others. But the final manuscript is ours and the responsibility for undetected errors that remain must be ours alone.

Above all our gratitude is expressed to our wives, who have endured many hours of loneliness during the writing of this book. Their forbearance and encouragement have contributed in no small part to the completion of this manuscript. To them it is dedicated.

Frank Owen
Ronald H. Jones

Liverpool Polytechnic

Preface to Second Edition

Opportunity has been taken to revise completely the content of this book and to incorporate a new chapter on Statistical Decisions which have become increasingly important since the book was first published.

We would like to thank the many people who have made suggestions for improvement, most of which we have incorporated in the revised text. We would like to thank too the many students using this book whose encouragement we have found highly rewarding.

Frank Owen
Ronald H. Jones

Acknowledgements

We would like to express our thanks to the following examinations bodies who have allowed us to make such a liberal use of their past examination questions:

The Association of Certified Accountants
The Chartered Institute of Secretaries and Administrators
The Institute of Cost and Management Accountants
The Association of International Accountants
The Chartered Institute of Public Finance and Accountancy

Contents

Chapter One

The Organisation of Data

The modern business world has a great hunger for facts and data. Well organised data improves our understanding of problems, and helps us to take decisions wisely. Badly organised data is little better than worthless. Unfortunately, you will all-too-often come across data that is not organised – most firms have filing cabinets full of data that someone intends to organise 'one day'. In this chapter we will suggest methods of how data can be organised for meaningful analysis – we will take our first steps in the rewarding (though often confusing) world of statistics.

Most people are vaguely aware that Statistics is concerned with figures in one way or another. Equally, we think, most people are rather distrustful of the statistics that they see quoted in the press or on television. We must admit that we ourselves have some sympathy for the housewife who is told on the news one evening that the cost of living has gone up by only 2% this month, and then finds in the shops next morning that everything she buys has, in fact, risen in price by between 5% and 10%. When this sort of thing happens it is no wonder that people get the impression that statistics can be made to prove anything. And yet — if our figures are accurate and the information is presented properly — how can this be so? We would like you to believe right from the start that no genuine statistician will ever deliberately misrepresent information or use it to mislead people. It can be done of course. In life many people are unscrupulous, and later in this course we will tell you how they misrepresent information, with the strict warning that *you* must never do it.

The great weakness of Statistics, is that to the man in the street who has never studied it, the methods used by statisticians are a closed book. We hope that as you work through this course your own personal book will be opened and that you will understand the dilemma in which our housewife finds herself.

But before we begin to think of the techniques you will use and the calculations you will perform, let us stop for a minute to consider the raw material you will be dealing with.

Suppose that the student union in your school or college wishes to obtain information about its members — their age, sex, home area, whether they live in a flat, or at home, or in lodgings and so on. How would the union secretary go about collecting this information? The most obvious way is for each student to be issued with a questionnaire, posing the relevant questions, and asking for it to be returned to the secretary's office. No doubt some of the forms will be incorrectly completed: some students may genuinely misunderstand the questions: some may refuse to answer certain questions which they regard as personal: doubtless some, in the fashion of the great petitions of the nineteenth century, will be signed by Queen Victoria or Karl Marx. Yet, with all its faults, this mass of completed questionnaires is the basic raw material for the statistical report that the union secretary wishes to produce.

Raw material such as this, collected at first hand, in response to specific questions is known as *primary data*; its characteristics are that it is obtained directly for the purpose of the survey which is being undertaken, and is, as yet, unanalysed.

Now, if your union secretary is lucky, he may also be able to obtain a great deal of information from the College administration, who, using enrolment forms as their primary data, may already have produced for their own purposes a fair amount of statistical information about students. Such information will, of course, have been produced for college purposes and may not be exactly what the union wants: but it is often useful additional information. Such data, which has already been collected for another, and different purpose, we know as *secondary data*. Usually it is of less use than primary data since it has already been processed and the original questionnaire is unlikely to have asked all the questions you would like to have asked. But whether it is primary or secondary, there can be very little statistical information which was not at one time to be found only in a pile of completed forms or questionnaires. The main task of any writer on statistics is to explain what the statistician does with his raw data between collecting it and presenting his report. So let us go back to your union secretary.

It is obvious that no-one would sit down and write a report in the form of

'Mary Smith is 17, lives in Durham, and is in lodgings here; Susan Yeung comes from Singapore and is in lodgings here ...'

We might as well hand over the completed forms to anyone who is interested since all that this type of report does is detail the information which is already given in detail on the questionnaire.

We can get a clue about the next stage of the analysis if we ask ourselves what it is that the union really wants to know. Surely the sort of information that is really wanted is how many students are 16, how many are 17 and so on; what percentage of students live at home; what proportion of students come from overseas. It is not the individual we are interested in so much as total numbers in given categories. The categories in this investigation may be age, sex, type of residence, number of hours a week spent on study and so on. Within each category students will vary. Some are 16; others are 17; some live at home, others in lodgings. We call each of these categories a *variable* because within each category students will vary. So we may now say that we are interested in a number of variables such as age, and more specifically in the value we can assign to each student within the range of values over which the variable extends. We may find, for example, that when we consider the variable 'age', 267 students are aged 17, 164 are aged 18 and so on up to the eldest student. The numbers of students whom we can place at each value of the variable we will call the *frequency,* because it tells us how often we will come across a student with this particular characteristic (that is, aged 18, or doing 27 hours a week private study, or travelling more than 15 miles to college). Thus the first step we must take is to decide what aspects of student life we are interested in and count up how many students are found within each of these categories. In so doing we are simplifying our data — reducing it to a more manageable form. In the process some detail is lost. We no longer know how old Brenda Jones is; but if we are interested we still have her completed questionnaire. On the other hand we do know that 267 students are aged 17 as well as much other general information.

Once we have reached this stage we are in a position to summarise our results in the form of a table and our work begins to look more like that of a statistician. Probably as a first tentative step we would produce a simple table dealing with only one variable. It might appear like this:

Age of Students attending ABC College

Age (the Variable)	Number of Students (the Frequency)
17	267
18	164
19	96
20	74
21 and over	23
	624

There is nothing wrong with our producing 15 or 20 tables like this, each concerned with a single variable, but surely it is better for presentation purposes if we could produce a small number of compound tables each showing several variables at once. Thus we could construct a double table showing the two variables, age and sex of students at the same time.

We have constructed this table by listing one of our variables vertically (age) and the other horizontally (sex). There is no golden rule, but it generally looks better if we tabulate the variable with the greater number of values vertically and that with the smaller number of values horizontally. Notice too that we have totalled both the vertical and the horizontal columns and that this adds to our information. We not only have the age distribution of male students and of female students but also the age distribution of the entire student population, and the total number of male and female students.

Age	Number of Students		
	Male	Female	Total
17	151	116	267
18	98	66	164
19	70	26	96
20	52	22	74
21 +	18	5	23
Total	389	235	624

You may of course still argue that the table is still concerned with only one variable, age, and that all we have is two age distributions. Let us then extend our table to consider three variables, age, sex and type of accommodation. Obviously now we must further subdivide either the horizontal or the vertical columns. Again it is a good general guide to say that we believe it better to subdivide the horizontal rather than the vertical columns. But in doing this the variable in the vertical column tends to become the more important. So we must consider which is the most important variable, and this often depends on what we are trying to show. Let us suppose that in this case we are aiming to show that the type of accommodation a student occupies depends on his or her age. In this case we will list the ages vertically and subdivide the horizontal columns. Our table may now appear like this:

	Number of Students						
Age	At Home		In Lodgings		In Flat		Total
	M	F	M	F	M	F	
17	112	92	16	20	23	4	267
18	64	42	24	16	10	8	164
19	31	12	28	7	11	7	96
20	8	4	16	10	28	8	74
21 +	2	3	3	1	13	1	23
Total	217	153	87	54	85	28	—
	370		141		113		624

You will readily appreciate what a vast amount of information a table such as this can give us: the number of students who live at home, subdivided into male and female and classified according to age, as well as the same information for those who are living in lodgings or living in a flat. You can understand too how much more information could be incorporated if we subdivided further the horizontal axis as well as some subdivision of the vertical axis such as the area of origin followed in each case by the age range.

There is one problem — the more we subdivide, the more complicated our table becomes, and there comes a time when it is so difficult to read it and understand it that we find that clarity has been lost rather than gained. It is true that one treble table, such as the one above, is better than three single tables. It is equally true that if we are considering eight or nine variables, three treble tables are better than one very complex one. And if you are wondering why clarity is so important think again what we have been doing. We have collected primary data, simplified it and classified it, and are now trying to present it to our union executive in a readily digestible form. How much notice do you think the executive will take of us if they cannot understand what our tables are all about?

Just in case you are ever in the position of having to construct tables to present the raw material you have collected, there are several points you should bear in mind. Let us call them the 'Principles of Good Tabulation'.

(a) Every table should have a short explanatory title at the head. At the end you should put a note of the source of the information you have used, whether it is based on your own survey or secondary data.

(b) The unit of measurement should be clearly stated, and if necessary defined in a footnote. Not many people, for example, would know offhand what a 'Long Ton' is. In addition the heading to every column should be clearly shown.

(c) Use different rulings to break up a larger table – double lines or thicker lines add a great deal to the ease with which a table is understood.

(d) Whenever you feel it useful insert both column and row totals.

(e) If the volume of data is large, two or three simple tables are better than one cumbersome one.

(f) Before you start to draft a table be quite sure what you want it to show. Remember that although most people read from left to right, most people find it easier to absorb figures which are in columns rather than rows.

As with most things practice is the best way of learning, and these principles will soon become second nature after you have drafted a few tables for yourself.

You might well ask at this stage whether this is all there is in the subject of Statistics. It if were you would all end up with distinctions. But the most important part of the work is still to come. No statistician (or student) worth his salt is content with a mere list of figures. He now begins to ask questions, the most important of which is 'What do the figures tell me?' We now begin, to analyse the figures, and statistical techniques are largely methods of extracting the utmost possible information from the data we have available. We could, for example, calculate the average age of students living at home, and compare it with the average age of students living in flats to try to determine whether we are right in assuming that the younger student will tend to live at home and the older students tend to be flatdweller. We can do the same thing for both male and female students to see if they behave differently. Let us say that there are many questions that the statistician can ask even from the simple data we have used so far.

We said earlier that the most obvious way for the union secretary to collect his data was to issue a questionnaire to each and every student. The results of his enquiry would cover every single student in the college — it will refer to what statisticians call the *population* of students. Beware of this term population. In statistics it does not mean the number of people living in a particular area. What it implies is that we have examined or obtained information about every single member of a particular group we are investigating. Thus we can talk of a population of telegraph poles, a population of shaggy-haired dogs, a population of ball-bearings and so on.

But is there any need for us to examine the population of students attending the college? If we wish to save time and money can we not do as so many public opinion polls do and take a sample of students? We could issue the questionnaire to, say 60 or 70 students only, or perhaps to every tenth student, and so reduce our raw data considerably. The *sample results* we obtain can then be applied to the population of students: if 12% of the sample live at home, we will argue that about 12% of all students in the college live at home.

Now, you may well argue that this can lead to wildly inaccurate results; and if you consider some of the results of public opinion polls in recent years it is apparent that things can, and do, go wrong. The sample chosen may be too small; it may not be representative of the population; the error arising as a result may mislead us. At this stage we will merely point out that in taking a sample we are in good company: an extremely high percentage of government statistics such as the statistics of Household Expenditure are based on samples which, on the face of it, appear to be ludicrously small.

If you think back now to the questions we suggested that you might ask about what our tables can tell us, you will realise that most of them involve a more detailed study of one variable only — the age of females living at home; the age of males living in flats. When we do begin to analyse you will appreciate that this is usual. The table presents several variables at once, but we extract just one of them at a time for further examination. In a few cases we will use two variables at once, when we are asking if there is a relationship between them such that one affects the other or that both move in sympathy. But in this foundation course we will never ask you to get involved in the analysis of three or more variables at once — which is indeed a complex matter.

The Grouped Frequency Distribution

Let us examine again the table showing the age of students attending the ABC college.

Age (x)	Number of Students (f)
17	267
18	164
19	96
20	74
21 and over	23
	624

Such a tabel is called a *frequency distribution*. It shows how many students (f) have the stated age (x). For example it tells us that 96 students are 19 years old. We stated that this table was ideal for summarising data dealing with only one variable – in this case age. However, it is not always convenient to arrange single variable data into a table like this. Suppose, for example that a scrap metal dealer buys a job lot of metal pipes. To get some idea of the lengths of pipes he selects 100 and carefully measures them. Now it is highly likely that no two pipes will have *exactly* the same length. It is no use then arranging the data into a distribution like the one we had for the ages of students. It might well be a table consisting of one hundred rows. It would be far more sensible to *group* the lengths together into classes like this:

Lengths of 100 copper pipes

Length (cm)		Frequency
10 but under	20	3
20	30	7
30	40	10
40	50	16
50	60	34
60	70	13
70	80	7
80	90	6
90	100	4

A table such as this is known as a *grouped frequency distribution*.

Certain points about this distribution can be noted.

1. There is no ambiguity about the way the classes have been stated. We are left in no doubt about the class to which a pipe belongs. For example, a pipe with a length of exactly 20 centimetres would go into the second class, i.e. 20 and under 30. However, if we were to state the classes like this

   10 – 20
   20 – 30

   into which class would we then put the pipe?

2. Although it is certainly more convenient to put the lengths of pipes into this distribution (much more convenient than a list of 100 lengths), we have lost something. For example, without going back to the original data, we would have no idea of the *actual* lengths of (say) the three pipes in the group 10 and under 20. We have sacrificed detail for the sake of presenting a picture which can be absorbed fairly easily. It might seem, of course, that the use of class intervals will prevent our using the frequency distribution as a basis for further work. Naturally, it does create a problem, and to overcome it we must make an assumption. We will assume that all three pipes in the class 10 and under 20 have lengths precisely at the centre of this class. Now, the smallest length that could be in the class is exactly 10 cms., and the largest length is 19.9999 cms. The centre then (or the *mid-point* as it is called) is

$$\frac{10 + 19.999\dot{9}}{2} = 15$$

3. We stated that the largest pipe in the first group could have a length of 19.9999 cms. and this throws light on the nature of the data we are dealing with. This data is what statisticians call *continuous* data – it can take any value within a particular group. There is, however, no reason why the data should have *particular* values within a group. It may well be that the scrap metal dealer, when taking measurements, has

recorded only certain values. He might, for examples, have recorded only to the nearest half centimetre, 11.0, 11.5, 13.5 and so on. But this way of measuring has been decided by the dealer and not by the *nature* of the data. The pipes may still have any length within the group, no matter how that length has been recorded. Generally speaking, any data that is obtained by measuring rather than by counting is continuous data, and, as we have said, should be listed in classes arranged in the form

x and under y

4. Let us now turn to another example. Suppose we count the number of telephone calls made by 100 firms on a particular day. Clearly this is not continuous data as the number of telephone calls made cannot take any value within a group. Thus, in the group 10 and under 20, we cannot make 10.2 or 14.3 telephone calls. The number of calls made can increase only in uniform steps of 1. Such data is called *discrete* data,[1] and it is usual to arrange it in a frequency distribution like this:

No of Calls	No of Firms
10 – 19	9
20 – 29	14

There is no ambiguity in stating the classes in this way as we cannot have 19.1, 19.2, 19.3 etc. calls. Again we have no precise idea of the exact number of calls made by firms in any particular class, and we assume that all firms have made a number of calls equal to the mid point of the class, i.e.

$$\frac{10+19}{2} = 14.5 \text{ calls.}$$

Now you may think that this is nonsense. After all, we have just stated that the number of calls must be a whole number, so how can we assume that 9 firms each made 14.5 calls? Well, don't let this worry you. This is something we assume merely for convenience of calculation.

At this stage it would be useful to give a few words of warning about using continuous and discrete data. We have implied that it is better to use classes of the type

10 and under 20
20 and under 30 for continuous data (Type A)

and

10 – 19
20 – 29 for discrete data (Type B).

Now although it would be wrong to use Type B for continuous data, there is nothing wrong with using Type A for discrete data, though it does create problems. Suppose we had arranged our data on telephone calls like this

10 but under 20	9
20 but under 30	14

[1] The uniform stepped increase for discrete data need not be 1 – shoe sizes increase by ½, hat sizes by ⅛, and money (in pounds sterling) by £0.005.

The unwary student might be tempted to say that the mid point of the first class is $\frac{10+20}{2} = 15$; but we know it is really 14.5. To avoid problems like like this, never assume that data is continuous merely because it is in a Type A distribution. Read the heading at the top of the table and decide what is the smallest and what is the largest possible value in each class. The mid point is then

$$\frac{\text{smallest} + \text{largest}}{2}$$

Another warning we must give you concerns rounded data (i.e. data given to the nearest whole number, the nearest 10 etc.) Suppose, for example, that the scrap metal dealer had rounded his measurements to the nearest centimetre. Now we know this data is continuous, so we may be tempted to use the Type A distribution.

10 but under 20
20 but under 30

However, suppose we have a pipe with an actual length of 19.6 centimetres. this would be rounded up to 20 and put into the second class *even though, really, it should be in the first class.* Problems such as this can be avoided if classes are carefully stated. It would not occur if, in this case, we had stated the classes as

9.5 but under 19.5
19.5 but under 29.5

The pipe with a length of 19.6 would be rounded up and (correctly) placed in the second class. Now, admittedly, many people use a type B distribution for rounded data, but the implication of doing this is that the class limits are 9.5 to 19.5 – another reason for reading carefully the heading at the top of the table and deciding for yourself whether the data is continuous or discrete.

A final warning concerns the class interval used in a frequency distribution. If it is a Type A distribution say

10 but under 20

then this of course covers a (continuous) range of ten numbers and the class interval is 10.

If it is a Class B distribution, say

10 – 19
20 – 29

you must remember that the groups cover 10 discrete values (10 to 19 inclusive) and the class interval is also 10 in this case.

Finally, you will find, if you look at any published statistical tables, that in many cases no limits are given for the first and last classes. An income distribution showing annual income might begin merely with 'Under £660' and end up with the group 'Over £50,000'. Such open-ended classes create problems and we will give you a few hints on how to handle them later.

One of the most difficult problems you will have in building up a frequency distribution from raw data is to decide on what class intervals to use. Obviously, a great deal will depend on the data you have available, but a few general guidelines may help. Firstly, try not to choose class intervals which will reduce the number of groups below five or six. If you do the data will be so compressed that no pattern emerges. Naturally the rule is not infallible — E.E.C. have published statistics of farm sizes giving only three classes. These three, however, correspond to a generally accepted international definition of small, medium and large farms. Our advice is that you should try not to emulate E.E.C. Equally, at the other extreme, do not have too many classes. About fifteen or sixteen is the maximum. The problem here is not only the difficulty of absorbing lengthy tables, but also the fact that each group will have a very low, or in some cases, even a zero frequency. And this leads to another point. At the upper end of the table, if you stick slavishly to a single class interval you may well find that several consecutive groups have no members while a higher group has a frequency of two or three. In these circumstances you should sacrifice the idea of equal class widths and combine the several classes into a single wider class.

A good general guide is to take the difference between the minimum and maximum value of the variable (which we call the *range*), and divide by ten. This will give you the right class width (or thereabouts) for the majority of classes, provided that you realise that class width of five or ten, or fifty is better than one of four, or seven or sixty-two, and provided that you take care with the extreme values of the variable.

Before we leave this brief description of the frequency distribution it would be an advantage if we show you how to tackle examination questions which ask you to construct a frequency distribution from a mass of figures. For this purpose we will look at a typical examination question.

Example

The following is a record of the percentage marks gained by candidates in an examination:

65	57	57	55	20	54	52	49	58	52
86	39	50	48	83	71	66	54	51	27
30	44	34	78	36	63	67	55	40	56
63	75	55	15	96	51	54	52	53	42
50	25	85	27	75	40	37	46	42	86
16	45	12	79	50	46	46	59	57	50
56	74	50	68	52	61	40	38	57	31
35	93	54	26	67	62	51	52	54	61
93	84	28	66	62	57	45	43	47	33
45	25	77	80	91	67	53	55	51	36

Tabulate the marks in the form of a frequency distribution, grouping by suitable intervals.

Looking at the figures we find that there are 100 marks given ranging from 12 to 96. We have laid down a principle of aiming at somewhere in the region of 10 classes in our frequency distributions and it certainly seems that