YOUR GUIDE TO BOOSTING EXAM RESULTS
upgrade

The challenge of numbers 2 — measures of location

Many students find the numerate elements of business studies difficult. In the second of a series, ***David Dyer*** *shows students how to summarise data effectively using the measures of location*

In most investigative studies a lot of data are produced and it is essential to reduce these to manageable proportions whilst at the same time maintaining their usefulness in solving the problem. I want to look at some of the choices available in making this decision, namely the mean, median, mode, bar chart, histogram, frequency table and ogive. Collectively they are known as *measures of location*.

Using data to make a point

People often quote or misquote figures to make a point because figures always seem to speak louder than words and they are credible because they are figures. Take the following examples:

- one in four marriages ends in divorce
- the average weekly wage is £62 (union leader)
- the average weekly wage is £110 (employers' organisation)
- 67% of the workers did not like their jobs.

The first statement was made on television and obscures the fact that three quarters of all marriages, on this evidence, do not end in divorce. The second and third statements were quoted in a dispute — does this mean one of them was lying? No, it means they were selective in the evidence they chose to emphasise their individual cases. The last statement was taken from a Research Assignment. Although it would seem to be conclusive it is representative of a common error. The assignment was undertaken in a firm which employed 350 people at the level being considered but the student had only interviewed 12 of them. The sample was therefore far too small to support such a statement.

The lesson here is that before we can rely on the interpretation given to any statistical statement we have to be aware of:

- the source of a statement
- the size of the sample
- the way they were chosen
- the question asked
- the conditions of the survey.

Any statistical statement should be treated with great care, using common sense as well as statistical awareness to interpret it. It is in this spirit you should use and interpret the techniques which follow.

Selecting a representative measure

Consider Table 1.

Table 1 *Monthly sales of the* Evening Herald, *summer 1997*

	August		July		June	
M					2	1340
T			1	600	3	620
W			2	720	4	690
Th			3	1200	5	1100
F	1	1125	4	1238	6	1026
M	4	1420	7	1250	9	1300
T	5	620	8	610	10	660
W	6	840	9	828	11	900
Th	7	900	10	921	12	980
F	8	730	11	756	13	769
M	11	1346	14	1502	16	1489
T	12	641	15	607	17	695
W	13	802	16	794	18	891
Th	14	911	17	976	19	1006
F	15	723	18	779	20	819
M	18	1621	21	1389	23	1665
T	19	534	22	664	24	723
W	20	778	23	904	25	991
Th	21	777	24	992	26	1033
F	22	643	25	659	27	921
M	25	1239	28	1234	30	1746
T	26	867	29	765		
W	27	765	30	741		
Th	28	665	31	722		
F	29	609				

Data like these are useful for reference purposes, as a background to the ideas you want to talk about, but they have to be interpreted and summarised according to your needs. You may, for example, want to work out:

- various totals — overall and weekly
- figures for each day e.g. Mondays
- best and worst day for each week
- the trend across 3 months
- a frequency distribution for which you must decide the intervals
- average sales in a number of different patterns e.g. per week, for Tuesdays, for each month
- a sales graph.

Having summarised the data, two obvious questions arise:

(1) Why are Monday sales much larger than the rest?

(2) Why are average sales in June better than July and August?

The answer to the first question might be that the shop supplying the data is near a main line station and many of the purchasers are weekly commuters. Alternatively, there might be something in the paper on Mondays which is attractive to readers. The answer to the second question might be that fewer people are on holiday in June than in July and August. On the other hand, it might be that there were five Mondays in June and four in July and August.

Taken as a whole, what the data have done is to pose questions that must be asked rather than to offer any explanations. This is normally the case. Data usually tell you where to start and rarely give any answers. The questions you might ask are:

- Are there any patterns to the data which need explanation?
- Is there a distinctive trend?
- Are there any peculiarities which need explanation?
- Are there any variations which might be important?
- Are there any lessons about sales which the figures might contain?

Measures of location

With this kind of measure we are seeking for a value which will be typical of the distribution and therefore represent it as well as any one measure can. Please refer to 'The challenge of numbers' in Vol. 3. No. 4 for consideration of the arithmetic mean (average).

The mode

This is the most frequent number in a distribution. This may be clear or a distribution may have more than one mode e.g.

1 222 3 444 7 888 9

Here there are three modes so the concept cannot really be of much use although the overall pattern might be once it is displayed. But the mode can be applied to the newspaper sales in Table 1 once they have been grouped into classes as part of a frequency distribution. Once you have done that the individual numbers disappear and you only have the number of observations in each class. To use the mode the assumption is made that the class with the largest number of entries is the *modal class*. This may not be very useful but it is likely to be of greater help because a range of numbers is involved and not just one figure. Even then, it is easy to misuse.

Suppose you took the absentee or lateness figures for a firm for a year and grouped them. There may be one or more modal groups but they might not be much use to you. If you divided the distribution into male employees, mothers with young children and other female employees, these distributions might well help you to pose questions which will aid solution. But by and large the mode is a simple figure which will rarely be useful.

The median

The median is the middle number in the range once it has been ordered, e.g.

Distribution 1 (unordered) 1 9 8 7 6 3 7 9 4 3 9

Distribution 2 (ordered) 1 3 3 4 6 **7** 7 8 9 9 9

The median is the middle number in distribution 2 i.e. the sixth number from the left which is 7. The median, in many ways represents what we tend to do in real life — most of us go for a middle value rather than an extreme. Adding a further observation and therefore having 12 of them, means there is no middle value and we overcome this by taking the middle two and dividing the total by 2 e.g.

1 3 3 4 4 **6 7** 7 8 9 9 9

$$\frac{6+7}{2} = 6.5$$

The weakness of the median is that the data can have any pattern and it will not influence the nature and position of the median. For example:

1 5 5 5 5 **7** 7 9 12 14 17

is an entirely different array in both pattern and range and yet it has the same median as our first example.

The median is easy to find when the number of observations is small, but it takes far too long when they are large and we need a formula to save time. The middle number of distribution 2 is the sixth number i.e. the 11 numbers plus one more divided by two and this is the general rule which we can represent as

$$\frac{(n+1)}{2}$$

where n is the number of observations and the distribution is ordered.

The median can be useful in those situations when the mean is affected by very high or low figures at the extremes of the range. Ranges A and D in Table 2 are like this. If there are range-widening single values in an otherwise well-grouped set of data, use the median, otherwise use the mean. The median is really the only available datum when we have only a part of the distribution. It is often a good measure when considering conclusions from a sample of a large population.

Table 2 *The deceptive average*

							Average
Range A	66	20	17	16	13	12	24
Range B	26	26	24	24	23	21	24
Range C	46	34	18	18	18	10	24
Range D	44	24	24	24	24	4	24

The median graphically

A cumulative frequency table for results in A-level Business Studies results is shown in Table 3. Note that the figures are an example and are not actual.

Table 3 tells us at a glance several important things. We can plot these figures as a graph of candidates (y axis) against

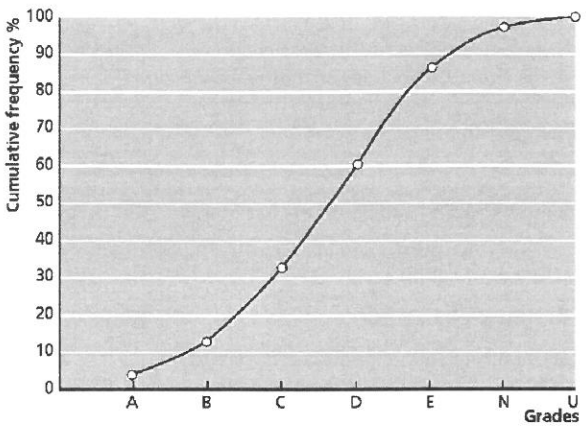

Figure 1 *The ogive*

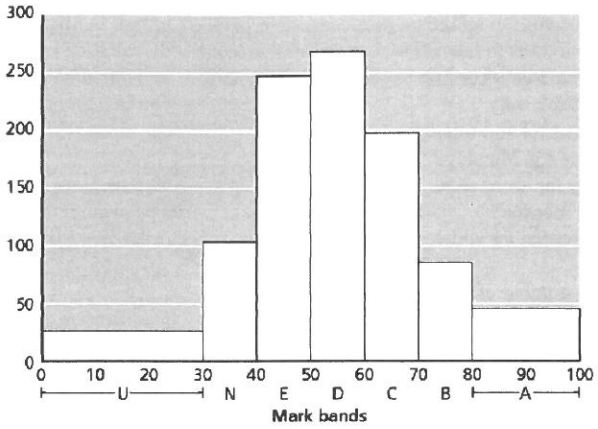

Figure 2 *A histogram showing Business Studies grades*

cumulative results (x axis) and this is called an *ogive*. It will show the number of candidates getting more than a particular mark as it rises from zero (see Figure 1). Since there are 940 candidates the median lies between candidates 470 and 471 and we can see from Table 3 that such a candidate would get a grade D. We can read it off the graph a little more precisely on the assumption that candidates' marks are spread evenly throughout each grade.

Table 3 *Cumulative frequency A-level Business Studies results*

Grade	A	B	C	D	E	N	U	Total
Actual	40	80	190	260	240	102	28	940
Cum.	40	120	310	570	810	912	940	
% of total	4.3	12.8	33.0	60.6	86.2	97.0	100	

The histogram

Distributions vary, e.g.:

- The distribution of marks gained by ten people in a very bright set might be expected to skew towards the top of the range.
- The distribution of marks for an exam taken by a large number of people of all abilities is likely to be normal.
- The distribution of the number of days absence through illness of a group of workers is likely to skew towards the bottom with many on 0 or 1 day because many people will have the odd day off or not be absent at all whilst a few will be absent persistently.

The beauty of the histogram is that it visualises these differences well. It also shows small differences, as does the bar chart, because the line across the top is broken in even the closest of cases. The histogram differs from the simple bar chart in that there is a numerical scale on the horizontal axis and the width of each arm of the histogram is determined by this. If one of the classes represented by the histogram were twice as wide as the others its base would be twice as wide (see Figure 2). Histograms also relate to continuous information and so there are no gaps between the arms. There are numbers but no other information on the vertical axis of an histogram (horizontal if it is drawn the other way round) because this axis is not measuring anything — it is the *area* that counts. Each block represents two factors which must be multiplied together to represent the full picture. In the set of examination results above we have at Grade D a range of marks (less than 60 but more than 49) and the number of candidates achieving that grade (260). If we drew a histogram of the number of candidates getting grades in the exam results (Table 3) its structure would be:

Class	A	80 marks or more × Observations 40
Class	B	Less than 80 but more than 69 × 80
Class	C	Less than 70 but more than 59 × 190
Class	D	Less than 60 but more than 49 × 260
Class	E	Less than 50 but more than 39 × 240
Class	N	Less than 40 but more than 29 × 102
Class	U	Less than 30 × 28

Classes B to N inclusive will take up the same space on the horizontal axis because they have the same fixed range whilst classes A and U will take up more space because they are broader classes. The vertical line will have marks graduated from 0–100.

It follows that when data are ungrouped and not continuous you would use a bar chart and otherwise use a histogram. Where classes are 'open' i.e. have one of the limits missing it will be at one end or the other of a distribution and common sense must be used in supplying the limit. For example, the obvious limit for the top class is 100 and for the bottom 0. In some cases you have to supply a reasonable limit e.g. the upper limit to an age class of 75 and over might reasonably be 100.

Histograms give a general picture of a distribution and its characteristics but many (and I am among them) consider that they are more trouble than they are worth. Most things can be achieved with the ogive or the bar chart.

In the next issue I consider *measures of spread* including: the range and sections of spread, the standard deviation (particularly useful for those taking the Cambridge Linear exam or who might want to use some simplified ideas about the normal curve in an assignment), scatter diagrams and sampling methods.

David Dyer *is Chief Examiner for Cambridge Modular and Oxford and Cambridge Modular and Linear courses. He is also Chairman of the* Business Review *team.*