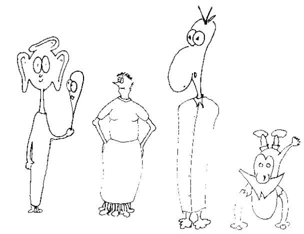
What is normal?



UNIT 3

The unit deals with a number of statistical topics - mean, mode and median as averages; range, mean deviation and standard deviation as measures of spread; and an introduction to the idea of the normal distribution. This leads into a discussion of what it means to be "normal".

Mathematical content

AT 4

- **Averages and measures of spread**

Spiritual and moral development

The unit aims to encourage pupils to think for themselves and form their own attitudes and opinions, rather than feel they have to conform to the norms of their peers and of society.

Using the unit

The unit is designed for students working for Higher Level GCSE, and it starts by revising work on averages. Some attention is given to:

- Quick ways of calculating means, using working means and scaling;
- The appropriateness of different averages in different circumstances.

We next introduce the concept of the spread of a distribution, and find a way to evaluate it by calculating the standard deviation.

The Normal distribution is then introduced, with some work on its calibration in terms of standard deviations.

The unit concludes with consideration of the danger of using the Normal distribution to prescribe what is right.

Work on the unit should take about three hours.

Students will need a calculator. Many calculators have a facility for calculating means and standard deviations, and you may wish to discuss the use of these buttons with your students.

Notes on the activities

Averages

Task 1 answers:

a) $^{52}/_{9} = 5.778$ (to 3 d.p.)

b) Use 4 as working mean, get
$4 + (^{0.4}/_{5}) = 4.08$

c) $^{31.3}/_{6} = 5.217$ (to 3 d.p.)

d) Use 34 (say) as working mean and get
$34 + (^{10}/_{7}) = 35.429$ (to 3 d.p.)

e) Take out factor of 4 and get
$4 (^{27}/_{9}) = 12$

f) Take out factor of 250 and get
$250 (^{53}/_{8}) = 1656.25$

Task 2 answers:

2 i) Modes: 7, none, none, none, 12, 1250
2 ii) Medians: 6.5, 4.1, 5.05, 35, 12, 1625

Task 3 answers:

a) $^{100}/_{8} = 12.5$ so mean salary is £12,500

b) $^{(8 + 11)}/_{2} = 9.5$ so median salary is £9,500

c) Modal salary is £8,000
d) The mean salary
e) The modal salary or, possibly, the median salary
f) Averages differ more if there is one number significantly different from the rest.

Measures of spread

Task 4 answers:
Means: Slogger 60, Steady 60

Task 5 answers:
Ranges: Slogger 146, Steady 24

Task 6 answer:
Mean of untreated deviations is always zero.

Task 7 answer:
Mean deviation for Steady: $^{38}/_{5} = 7.6$

Task 8 answers:

1. Standard deviation for Steady: 8.56 (to 2 d.p.)
   Much lower than Slogger's

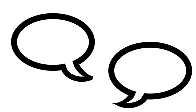
2. Mean	Mean deviation	Standard deviation
a) 20	$^{32}/_{7} = 4.57$	5.24
b) 5.2	2	2.32
c) 14.83	1.5	1.86
d) 4.5	2	2.29
e) 11.02	0.98	1.03
f) 62.6	2.68	3.2

(Rounded answers are given correct to 2 d.p.).

Task 9 answers:

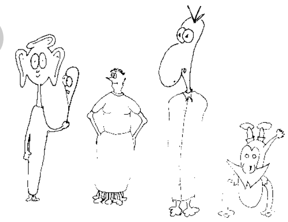
1. i) 16% ii) 2.25% iii) 18.25%

The Normal distribution



It is suggested that students are asked to read the final section, "What is normal" and reflect on it, prior to a lesson. They could be asked to think of areas where it is good to conform, and others where it isn't, in advance, to enable a useful discussion to take place.

What is normal?



UNIT 3

Statistics is about collecting, displaying and analysing data.
Often, there is a lot of data, such as the 1991 Census, collecting information on all 56 million inhabitants of the UK, or school 'league tables' presenting statistics about 6000 secondary schools in England and Wales. Clearly, no-one can hope to discuss such vast amounts of data easily, and so they are often simplified. The most common way in which this is done is to work out an average.

Averages

We are all familiar with the idea of the Mean - the average obtained by adding up a set of numbers and dividing by the number of items.

For example, the mean of 2, 5, 7, 4, 8, 3, 4, 11 is

$$\frac{(2 + 5 + 7 + 4 + 8 + 3 + 4 + 11)}{8}$$

which equals 5.5.
Sometimes it is possible to use some 'tricks' to speed up the arithmetic.

1

Work out the means of the following sets of numbers, using quick methods where possible.

a) 3, 8, 7, 6, 5, 9, 4, 3, 7, 7

b) 4.1, 3.9, 4.2, 4.4, 3.8

c) 2.5, 4.8, 5.3, 9.1, 6.0, 3.6

d) 34, 36, 37, 32, 33, 35, 41

e) 4, 8, 12, 12, 4, 20, 16, 20, 12

f) 1500, 1250, 1250, 1750, 2000, 750, 2500, 2250

2

Other 'averages' are possible; two common ones are: the mode - the number that occurs most often, and the median - the middle number when all the data is arranged in order.

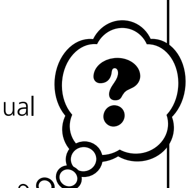
Find the i) mode and ii) median of the sets of data in 1a) to f) above.

3

Using different averages can give quite different results.
Suppose the various workers in a company received the following annual salaries (in £1000s): 6, 8, 8, 8, 11, 12, 16, 31.

1. Calculate the a) mean, b) median and c) modal salaries.

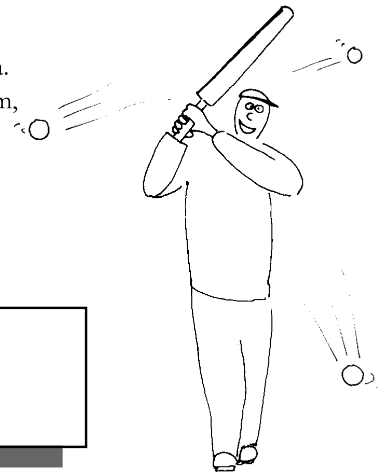
d) If you were the chairman of the company, trying to keep the wage costs down, which figure would you use when suggesting to the workers, in the annual pay negotiations, that they were already highly paid?



e) If you were the union representative, arguing for a higher pay rise, which figure would you use to suggest that the workers are not very well paid at the moment?

f) Why do you think the three averages differ so much? Which is 'best'?

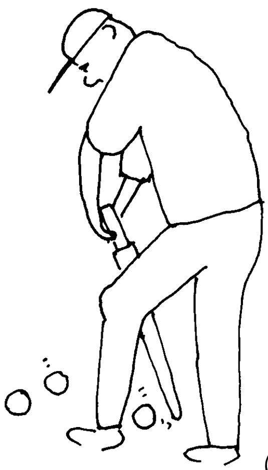
Measures of spread



Sometimes, a single average does not give us sufficient information about data. Suppose, for example, that you had to choose between two batsmen for a team, Slogger and Steady, whose scores in the last five innings were;

Slogger 106, 31, 148, 2, 13
Steady 58, 66, 49, 54, 73

4 Find the mean score for each batsman.



In fact, the mean scores are the same. However, the distributions of the numbers are quite different. For example, if a team needed 95 to win, it seems very unlikely that Steady would get it, whereas Slogger might (although he might equally get a very low score). Steady is reliable, but Slogger could be a match-winner.

To present these different types of data, statisticians try to measure how 'spread out' the data is. Is it all fairly consistent, like Steady's, or is it much more erratic, like Slogger's?

As with averages, there are various ways of doing this, and the simplest measure of spread is called the Range. It is just the highest figure minus the lowest.

5 Calculate the range for Steady's score, and for Slogger's.

This clearly shows the difference in the two sets of numbers.
However, the range is a rather simplistic answer, which can easily be distorted by one single unusual result. For example, 2, 5, 7, 3, 5, 1, 4, 3, 19, 2 has a range of 19 - 1 = 18, but almost all of the data lies within a range of 1 to 7. Here the single number, 19, has produced a very misleading result.

A better method is to use all the numbers, and find out how far they are, on average, from the mean.
We will now see how to do that.

Deviations

The distance of each number from the mean of its set is called its deviation.

6 For Slogger, the deviations are
46, -29, 88, -58, -47

Calculate the average of these numbers.

Adding these numbers up and dividing by five gives a strange result. This is because the negative deviations cancel out the positive ones. This method would therefore give an average deviation of zero every time and this is not at all helpful as a measure of how spread out the data is!
To overcome this difficulty, statisticians have devised two alternative methods.

One solution is called the Mean Deviation. This just neglects the minus signs and then takes the mean. For Slogger it is calculated as follows:

Continued from page 2

Slogger's scores	106	31	148	2	13
Deviation from mean (60)	46	-29	88	-58	-47
Deviation without minuses	46	29	88	58	47

So the mean deviation is $\frac{(46 + 29 + 88 + 58 + 47)}{5} = 53.6$

7 Calculate the mean deviation for Steady's scores.

This again shows clearly that Steady's scores are more consistent than Slogger's.

The mean deviation is one measure of how spread out the data are. The process of 'dropping the minus sign' can cause problems.

The alternative measure of spread of the data is called the Standard Deviation and this is the one that is used most of the time by statisticians. (It is the one for which calculators are usually programmed.)

To work out the standard deviation, the minus signs of the deviations are eliminated by squaring the deviations. (Remember that any number squared is positive, regardless of whether it was positive or negative originally.)

We then calculate the mean of the squared deviations, and finally we obtain the square root of this mean.

Let us see how it works for the batsmen:

Slogger's scores	106	31	148	2	13
Deviation from mean (60)	46	-29	88	-58	-47
Squared deviations	2116	841	7744	3364	2209

Their mean is $\frac{(2116 + 841 + 7744 + 3364 + 2209)}{5} = \frac{16274}{5} = 3254.8$

The standard deviation is $\sqrt{3254.8} = 57.05$ (to 2 d.p.)

8

1. Calculate the standard deviation of Steady's scores, and compare it with Slogger's. What does it show?
2. Calculate the mean deviation and standard deviation of:
   a) 23, 24, 15, 24, 19, 25, 10
   b) 2, 6, 5, 4, 8, 9, 3, 6, 7, 2
   c) 12, 14, 15, 16, 18, 14
   d) 1, 2, 3, 4, 5, 6, 7, 8
   e) 10.3, 12.4, 12.1, 9.8, 10.5
   f) 63, 65, 59, 57, 66, 63, 68, 60, 61, 64

With the mean and standard deviation, it is now possible to compare any two sets of statistical data.

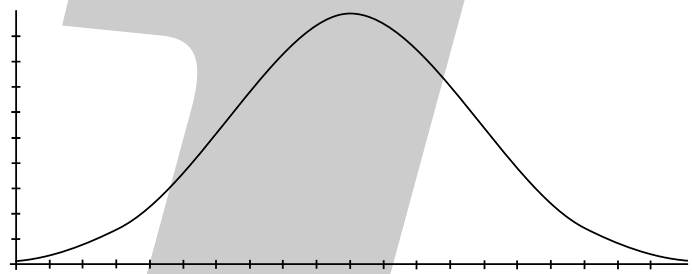
The Normal distribution

The way in which the numbers are distributed in a set of data is very important. In some cases, most numbers are near one end, perhaps very small, with just a few large numbers. For example, winnings on the National Lottery are mostly £10, with very few over £1,000,000.

There is, however, one pattern that is very common, and which occurs in a wide variety of situations. This is where most numbers are in the middle, with fewer occurring the further away from the mean (above or below) one gets.

One example of this might be the length of grass blades measured by a botanist. The mean might be 6cm, with some longer and some shorter. Very long blades and very short ones would be uncommon.

Another example, which is a slight approximation, is the number of heads scored when a coin is tossed 200 times. On average, there will be 100 heads. In practice, you would be almost as likely to get 97, or 103, and so on, but less likely to get 130 or 70. A graph of the frequency of the various scores would look like this:



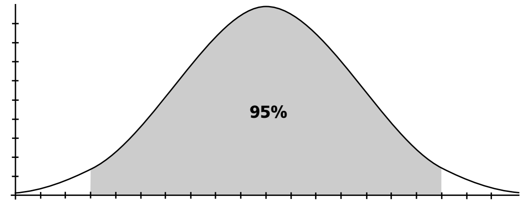
This is a typical Normal distribution, sometimes called the 'bell shaped curve'. All Normal distributions, whatever their mean and standard deviation, have the same relative proportions, e.g., they all have the top 10% occurring above 1.28 standard deviations above the mean. Also it should be noticed that the mean, median, and mode are all the same.

In this way, the mean and standard deviation (SD) provide a scale for the distribution (this is another reason why the standard deviation is so important).

Mathematicians have published detailed tables showing what percentage of the population lie within certain distances from the mean, and these Normal tables are used by all statisticians.

Some of the numbers are used so often that they are known by heart:

No. of SD's away from the mean	% of population included
0	0
1	68
1.65	90
1.96	95
2	95.5
2.33	98
2.58	99



In theory, 100% is never quite reached, as the curve never quite touches the axis; in reality, almost all of the population will be included between 3 standard deviations above and below the mean.

1. A population is described by a normal distribution, with mean 28 and standard 4. Find the percentage of the population who are
   i) above 32, ii) below 20, iii) between 24 and 36

2. Make a list of the heights of the students in your class and calculate the mean and the standard deviation of these heights.

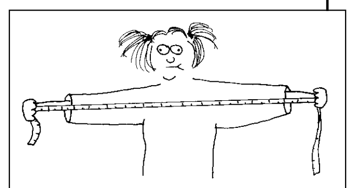
3. Find what height is 1.65 standard deviations above the mean and what height is 1.65 standard deviations below the mean.

4. What percentage of the students' heights lie between these two numbers?

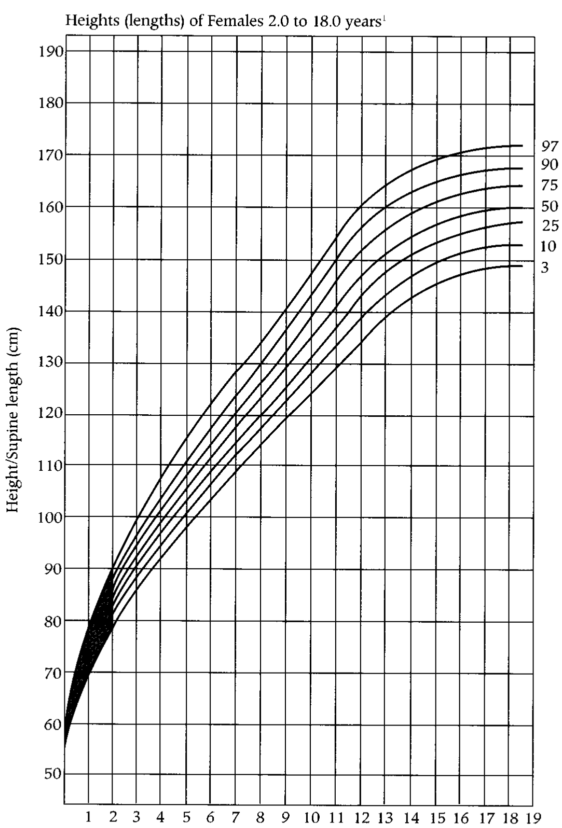
The answer to the last question should be approximately 90%. However, your class is relatively small as compared with the whole population of students of your year-group in the country (or the world!) so the percentage may be a bit above or below 90%.

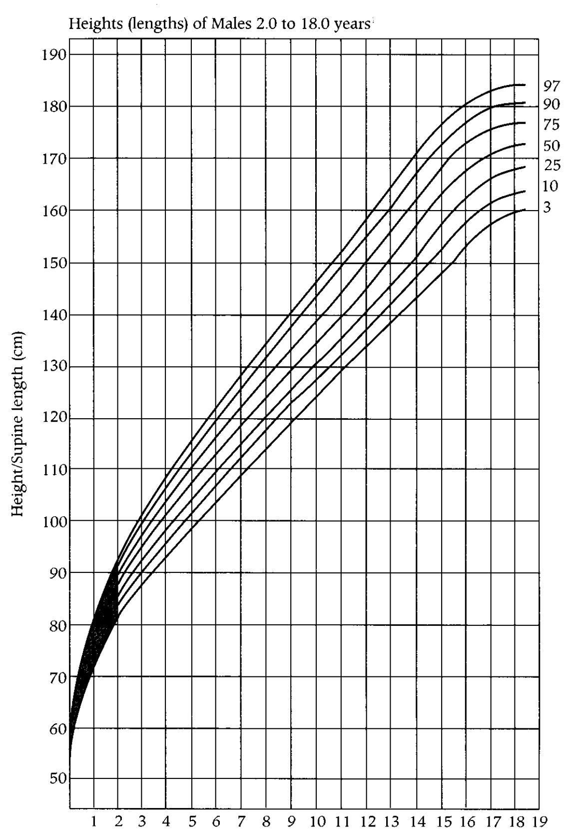
5. Repeat the previous three questions for handspans of students in the class.

6. Repeat them for armspans of students in the class.



Doctors have compiled charts showing how tall/heavy children should be at various ages (see below for height). They become concerned if a child is in the top or bottom 1% of the population.





To find the possible extremes of measurements, have a look in The Guinness Book of Records.

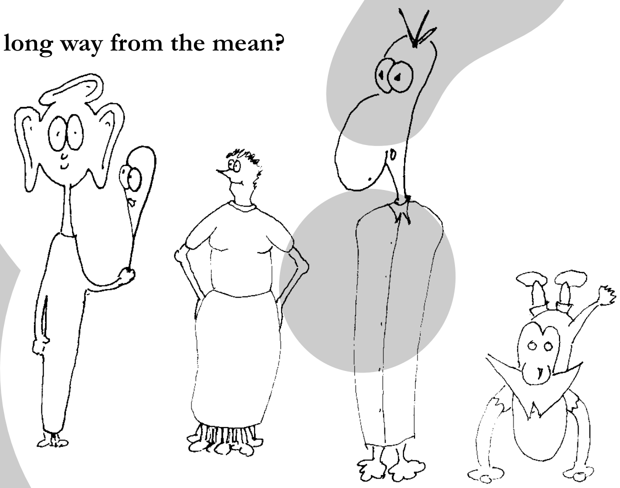
What is normal?

One effect of working with the Normal distribution is that one becomes used to considering the items in the middle. Indeed, items which are a long way from the mean become regarded as very unusual and strange.

For example in the case of the 200 coins, 99% of the time the result will be between 81 and 119 heads. What happens if you actually threw 76 heads? This is very unlikely - almost, but not quite, 'impossible'. Should we therefore disregard it - perhaps assuming the coin is biased? Statisticians generally reckon that any event with a probability of less than 1% is so unlikely as to be impossible.

But in a wider context, should we ignore events which are unlikely, or unusual, or odd? In schools, for example, should we just cater for the 99% of pupils who are in the middle range of intelligence, and ignore those at the ends of the Normal distribution - the geniuses or those with severe learning difficulties?

Is it wrong to be a long way from the mean?



Many people feel more comfortable being similar to everyone else - no-one likes being regarded as different, or odd. Yet in many areas, it is the people who are exceptional who lead the rest of us - in scientific discovery, in athletic achievement, in artistic and cultural activities. Einstein, Linford Christie and Beethoven would all have been, or indeed are, at the far end of the spectrum of ability in their respective fields, but they were willing to use their talents and not to hide them.

In society, we also tend to respect categories of people by how many there are of them. A group that consists of 1% or less of the population, such as those in wheelchairs, might have difficulty making its collective voice heard, whilst a group of 50% of the population, e.g. women, are more capable of promoting reform.

Is it right to discriminate against any particular group because of its small size, e.g., those who want to maintain a quiet Sunday, with shops closed, are a small percentage, but should their wishes be ignored? Equally, because such people are a small proportion of the population, should you feel uncomfortable with being identified as one of them?

It sometimes needs courage to identify yourself as a member of a minority group. Ethnic groups, suffragettes, trade unionists, Christians and members of other faith groups have all suffered discrimination. Statistics can tend to encourage conformity, but we need not feel that there is something wrong with us if we are not "within 2.58 standard deviations of the mean".