# UNIT 9 *Data Analysis* Activities

## Activities

# ACTIVITY 9.1 *Averages*

Sometimes data sets have to be summarised by a *single v*alue, usually called an *average*.

There are three types of average measures commonly used:

*Mean* *Median* *Mode*

## A *Calculating each average measure*

27 members of a class were set a 'logic' question and the times (in minutes) each pupil took to solve it were noted.

*Times (in minutes) taken to solve 'logic' question*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 19 | 14 | 15 | 9 | 18 | 16 | 10 | 11 | 16 |
| 4 | 20 | 10 | 14 | 11 | 9 | 13 | 15 | 13 |
| 12 | 2 | 17 | 15 | 14 | 10 | 11 | 10 | 12 |

1. The **MEAN** value of a set of data is $\frac{\text{sum of values}}{\text{number of values}}$.

   What is the *mean* (to 2 d.p.) of the times given in the table?

2. The **MEDIAN** is the *middle* value of an ordered set of data.

   (a) Write down the times in the table above in *ascending* order, i.e. smallest first.

   (b) How many values are there? (c) What is the *median* ?

3. The **MODE** is the value which occurs most often, i.e. the most popular.
   What is the *mode* of the times in the table above?

4. Which of the three measures do you think is most representative of the average time? Give your reasons.

## B *Choosing which measure to use*

5. In a clothes shop, the sizes of a particular dress sold during one week were noted and are shown in the table opposite.

   (a) Find the *mean*, *mode* and *median* for this data.

   (b) Which measure is of most use to the sales staff?

*Dress sizes sold in one week*

| | | | | |
|---|---|---|---|---|
| 10 | 14 | 12 | 16 | 18 |
| 16 | 12 | 10 | 14 | 16 |
| 16 | 14 | 18 | 8 | 14 |
| 12 | 16 | 10 | 10 | 16 |
| 16 | 18 | 14 | 16 | 8 |

6. The wages of factory employees are shown in the table.

   (a) Find the *mean*, *mod***e** and *median* of the weekly wages.

   (b) Which of these measures is the most useful?

*Factory wages paid per week*

| | | |
|---|---|---|
| 10 | are paid | £120 |
| 35 | are paid | £140 |
| 25 | are paid | £160 |
| 30 | are paid | £180 |

# ACTIVITY 9.2 *Inter-quartile Range*

The % of first class letters delivered the next day from a number of Sorting Offices in the UK is shown in the tables below for 1989–90.

**Table 1** ***Southern England***

| *Letters posted in* | *Within district* | *Neighbouring district* | *Distant district* |
|---|---|---|---|
| BRIGHTON | 92 | 78 | 73 |
| LONDON | 86 | 80 | 70 |
| GUILDFORD | 89 | 75 | 74 |
| OXFORD | 81 | 68 | 65 |
| PORTSMOUTH | 89 | 85 | 64 |
| READING | 85 | 67 | 65 |
| WATFORD | 91 | 73 | 75 |
| EXETER | 95 | 89 | 67 |
| BOURNEMOUTH | 92 | 86 | 63 |
| MILTON KEYNES | 84 | 70 | 62 |
| CHELMSFORD | 86 | 79 | 70 |

**Table 2** ***Northern England and Scotland***

| *Letters posted in* | *Within district* | *Neighbouring district* | *Distant district* |
|---|---|---|---|
| LEEDS | 93 | 88 | 70 |
| NEWCASTLE | 91 | 88 | 68 |
| CARLISLE | 92 | 87 | 68 |
| MANCHESTER | 90 | 82 | 63 |
| ABERDEEN | 91 | 91 | 65 |
| EDINBURGH | 89 | 87 | 63 |
| GLASGOW | 93 | 83 | 67 |
| INVERNESS | 83 | 79 | 40 |
| PERTH | 91 | 90 | 63 |

So much data is given here that it is difficult to see what it all means! There are many ways of analysing and presenting data to make it more understandable. One method is shown below for the *Within District* column in **Table 1**.

STEP 1 The data are written out in ascending order and the number of values ($n$) noted as 11.

81, 84, 85, 86, 86, 89, 89, 91, 92, 92, 95

STEP 2 The data are divided into 4 equal parts (or *quartiles)* as shown below:

81, 84, $\boxed{85,}$ 86, 86, $\boxed{89,}$ 89, 91, $\boxed{92,}$ 92, 95

LQ Median UQ

The *Lower Quartile* (LQ) is the $\frac{1}{4}(n+1)$ th number = 3rd number = 85.

The *Median* is the $\frac{1}{2}(n+1)$ th number = 6th number = 89.

The *Upper Quartile* (UQ) is the $\frac{3}{4}(n+1)$ th number = 9th number = 92.

If the data set had an *even* number of values and a quartile had worked out as, e.g. the $3\frac{1}{2}$th number, then the average of the 3rd and 4th numbers would be taken.

STEP 3 The *Inter-quartile Range* (IQR) is the difference between the Lower Quartile and the Upper Quartile and shows how spread out the data are. In this case,

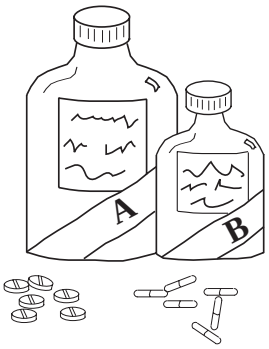
$$\text{UQ} - \text{LQ} = 92 - 85 = 7.$$

This means that the middle half of the data points lie within a range of 7 units.

1. Find the median and IQR for the other two sets of data in *Table 1*.
2. Find the median and IQR for the three sets of data in *Table 2*.
3. Compare and contrast all these data sets, identifying their main characteristics.

# ACTIVITY 9.3 *Box and Whisker Plots*

A hospital has been trying out two new pain relieving drugs. Patients taking the drugs were asked to record how many hours it was before any pain recurred.




The results are summarised below.

| *Drug* | *Recurrence Times* (hours) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *A* | 6 | 7 | 10 | 7 | 11 | 5 | 8 | 10 | 6 | 7 | 11 |
| *B* | 3 | 12 | 5 | 10 | 2 | 13 | 9 | 15 | 3 | 6 | 14 |

The hospital's problem is to decide which of the drugs is preferable.

1. (a) Find the mean recurrence time for each drug.

   (b) Does this give you sufficient information to recommend which drug to use?

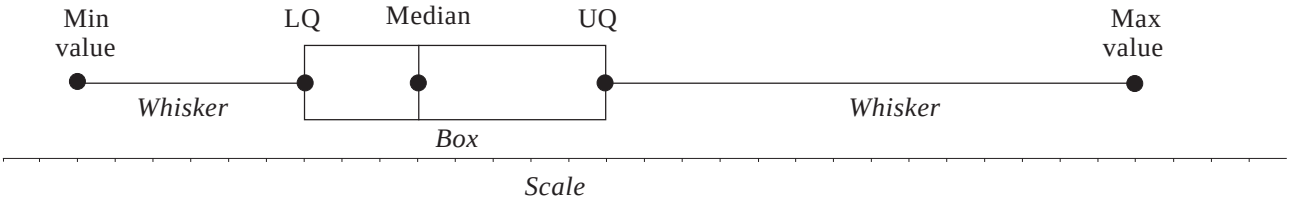
A more useful method of analysis, which summarises more of the data, is called a *Box and Whisker Plot*. You must first find the *median* and *upper* and *lower quartiles*.

2. For the data from *Drug A*, what is

   (a) the minimum time (b) the lower quartile (c) the median

   (d) the upper quartile (e) the maximum time?

You can now construct a *Box and Whisker Plot* for the data, of the type illustrated below.




3. (a) Using a *Box and Whisker Plot*, with the same scale for each, illustrate the data for

   (i) *Drug A* (ii) *Drug B*.

   (b) Which drug do you think should be used now and why?

***Extension***

A worker travelling daily from Reading to London has a choice of three types of train to catch:

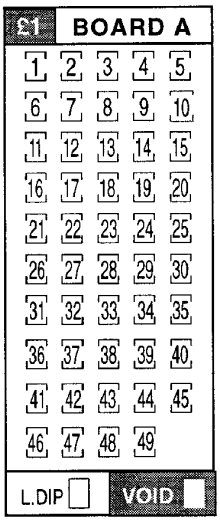
• *stopping* • *semi-fast* • *inter-city*.

She is concerned about the punctuality of the service, and has kept records for all three trains over a period of three weeks. These are shown in the table below.

| *Type of Train* | *Minutes Late* ( * This means the train was early! ) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Stopping* | 0 | 3 | 2 | 10 | 5 | 0 | 1 | 7 | 12 | 10 | 4 | 0 | 0 | 5 | 7 |
| *Semi-fast* | 1 | 2 | 5 | 4 | 3 | 2 | 2 | 5 | 0 | – 1* | 2 | 7 | 2 | 0 | 3 |
| *Inter-city* | 3 | 0 | – 2* | 0 | 10 | 3 | 1 | 35 | –3* | 2 | 5 | 30 | 0 | 0 | 2 |

Illustrate the data using *box and whisker plots*. Which type of train should she catch and why?

# ACTIVITY 9.4 Sheet 1 *National Lottery*




To play the National Lottery, you pay £1 to enter your choice of *six* distinct numbers from

1, 2, 3, . . . , 49

Each week, six numbers (plus a *bonus* number) are drawn randomly and you win a prize if you have three, four, five or six numbers the same.

The complete list of 6 winning numbers (plus *bonus* number) in the first 100 draws of the National Lottery is given on a separate sheet.

Over a long period of time, if the numbers drawn are truly random, you would expect that each number will occur roughly the same number of times, i.e. will have the same frequency.

However, some commentators insist that certain numbers are luckier than others! This activity will investigate their claim.

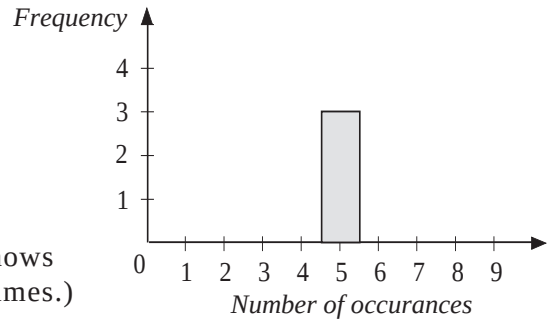
1. Over 49 consecutive weeks, how many times, on average, should each number be drawn?

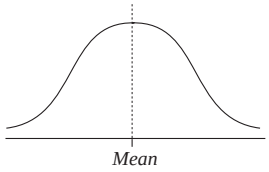
Of course, in practice some numbers will be drawn more, and others less, than this average.

2. (a) From the *Data Sheet, Activity 9.4, Sheet 2,* choose any 49 consecutive draws and find the frequency of occurrance of each number.

   (b) Plot a frequency diagram of the 49 frequencies found.

   (For example, the bar chart opposite shows that 3 particular numbers occurred 5 times.)




Over a long period, this frequency diagram should approximate to a *normal distribution* (similar to the diagram opposite).

To see if your distribution is a reasonable one, on the assumption that the numbers drawn are random, you can use random number tables (*OS8.6*) to generate your own draws and find the corrresponding frequency diagram.

To use the table of random numbers, you should start anywhere and take pairs of digits. *Ignore 00 and 50*. The digits 01–49 will give exactly corresponding numbers, but the digits 51–99 can also be used for the numbers 1–49 by subtracting 50. *Ignore any repeated numbers.*

For example, the random numbers 93319, 51747, 56137, . . . will generate draw numbers

43, 31, 45, 17, 47, 6.

3. Use random number tables to simulate 49 draws and compare your frequency diagram with that found in Question 2. What do you conclude?

# ACTIVITY 9.4 Sheet 2 *National Lottery Data*

| *Draw* | *Six winning numbers* | | | | | | *Bonus* |
|---|---|---|---|---|---|---|---|
| 1. | 30 | 3 | 5 | 44 | 14 | 22 | *10* |
| 2 | 16 | 6 | 44 | 31 | 12 | 15 | *37* |
| 3. | 21 | 11 | 17 | 30 | 29 | 40 | *31* |
| 4. | 26 | 47 | 49 | 43 | 35 | 38 | *28* |
| 5. | 13 | 3 | 38 | 5 | 14 | 9 | *30* |
| 6. | 27 | 29 | 39 | 3 | 44 | 2 | *6* |
| 7. | 17 | 44 | 36 | 32 | 9 | 42 | *16* |
| 8. | 21 | 32 | 2 | 5 | 25 | 22 | *46* |
| 9. | 23 | 38 | 17 | 7 | 32 | 42 | *48* |
| 10. | 47 | 6 | 16 | 31 | 30 | 20 | *4* |
| 11. | 31 | 16 | 25 | 43 | 4 | 26 | *21* |
| 12. | 46 | 42 | 1 | 38 | 7 | 37 | *20* |
| 13. | 48 | 38 | 15 | 29 | 18 | 35 | *5* |
| 14. | 45 | 16 | 36 | 19 | 21 | 29 | *43* |
| 15. | 18 | 33 | 8 | 31 | 5 | 10 | *28* |
| 16. | 17 | 36 | 11 | 12 | 42 | 26 | *13* |
| 17. | 2 | 22 | 13 | 46 | 29 | 27 | *36* |
| 18. | 41 | 19 | 31 | 18 | 9 | 24 | *21* |
| 19. | 4 | 49 | 41 | 44 | 42 | 17 | *24* |
| 20. | 43 | 41 | 22 | 25 | 30 | 32 | *29* |
| 21. | 42 | 17 | 22 | 24 | 47 | 14 | *34* |
| 22. | 1 | 23 | 26 | 4 | 6 | 49 | *8* |
| 23. | 33 | 36 | 8 | 20 | 38 | 18 | *46* |
| 24. | 31 | 9 | 15 | 34 | 48 | 22 | *23* |
| 25. | 35 | 14 | 48 | 17 | 43 | 5 | *22* |
| 26. | 41 | 16 | 28 | 25 | 7 | 26 | *19* |
| 27. | 46 | 15 | 17 | 28 | 6 | 32 | *22* |
| 28. | 45 | 12 | 25 | 37 | 44 | 13 | *9* |
| 29. | 31 | 1 | 29 | 40 | 21 | 32 | *27* |
| 30. | 44 | 15 | 26 | 46 | 12 | 49 | *14* |
| 31. | 48 | 30 | 40 | 27 | 38 | 33 | *2* |
| 32. | 5 | 43 | 45 | 21 | 15 | 42 | *20* |
| 33. | 25 | 7 | 8 | 5 | 48 | 44 | *3* |
| 34. | 3 | 14 | 11 | 20 | 1 | 40 | *45* |
| 35. | 1 | 4 | 43 | 20 | 31 | 41 | *38* |
| 36. | 3 | 21 | 22 | 2 | 23 | 40 | *24* |
| 37. | 41 | 34 | 49 | 28 | 46 | 45 | *11* |
| 38. | 35 | 1 | 25 | 30 | 45 | 8 | *15* |
| 39. | 25 | 33 | 28 | 47 | 11 | 34 | *48* |
| 40. | 24 | 23 | 48 | 5 | 8 | 28 | *19* |
| 41. | 21 | 41 | 18 | 38 | 16 | 27 | *26* |
| 42. | 40 | 49 | 28 | 15 | 1 | 22 | *44* |
| 43. | 12 | 22 | 41 | 2 | 20 | 45 | *47* |
| 44. | 37 | 14 | 25 | 41 | 10 | 2 | *5* |
| 45. | 10 | 34 | 24 | 19 | 5 | 46 | *28* |
| 46. | 11 | 33 | 40 | 10 | 32 | 29 | *16* |
| 47. | 28 | 37 | 10 | 30 | 36 | 22 | *45* |
| 48. | 25 | 30 | 9 | 5 | 4 | 47 | *17* |
| 49. | 17 | 19 | 2 | 21 | 6 | 47 | *5* |
| 50. | 16 | 33 | 44 | 27 | 35 | 7 | *5* |

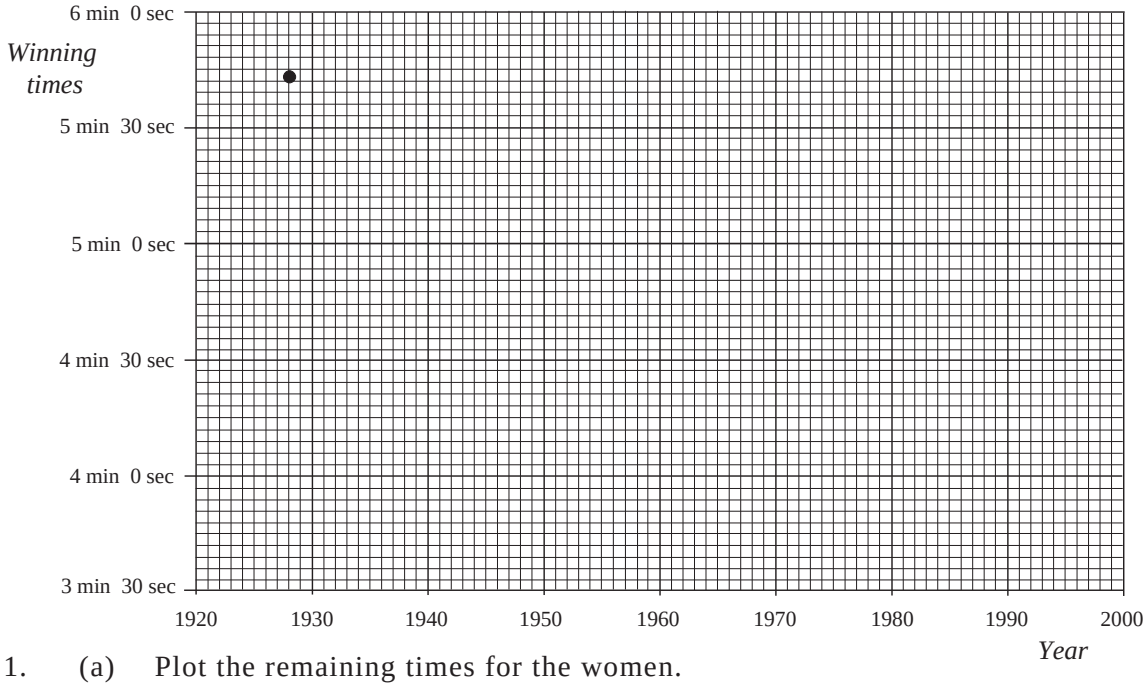
| *Draw* | *Six winning numbers* | | | | | | *Bonus* |
|---|---|---|---|---|---|---|---|
| 51. | 6 | 14 | 18 | 48 | 27 | 44 | *1* |
| 52. | 23 | 28 | 48 | 10 | 7 | 30 | *3* |
| 53. | 33 | 7 | 4 | 48 | 18 | 45 | *1* |
| 54. | 46 | 42 | 28 | 16 | 30 | 23 | *45* |
| 55. | 26 | 16 | 19 | 46 | 15 | 35 | *7* |
| 56. | 5 | 26 | 29 | 12 | 11 | 33 | *20* |
| 57. | 23 | 49 | 7 | 28 | 35 | 8 | *10* |
| 58. | 40 | 47 | 6 | 49 | 34 | 11 | *16* |
| 59. | 6 | 43 | 42 | 39 | 45 | 32 | *36* |
| 60. | 4 | 13 | 2 | 3 | 42 | 44 | *24* |
| 61. | 31 | 32 | 48 | 21 | 29 | 34 | *25* |
| 62. | 23 | 37 | 33 | 30 | 25 | 5 | *3* |
| 63. | 16 | 41 | 38 | 17 | 43 | 42 | *28* |
| 64. | 2 | 32 | 44 | 22 | 9 | 26 | *40* |
| 65. | 14 | 28 | 11 | 4 | 15 | 42 | *6* |
| 66. | 18 | 14 | 16 | 22 | 4 | 15 | *33* |
| 67. | 5 | 24 | 44 | 2 | 7 | 35 | *30* |
| 68. | 41 | 11 | 9 | 24 | 45 | 12 | *6* |
| 69. | 45 | 30 | 37 | 16 | 29 | 14 | *7* |
| 70. | 19 | 38 | 48 | 12 | 28 | 2 | *45* |
| 71. | 30 | 18 | 5 | 14 | 43 | 7 | *28* |
| 72. | 37 | 12 | 49 | 27 | 26 | 28 | *43* |
| 73. | 38 | 4 | 14 | 1 | 17 | 6 | *9* |
| 74. | 38 | 47 | 23 | 44 | 49 | 40 | *12* |
| 75. | 40 | 31 | 9 | 29 | 28 | 48 | *23* |
| 76. | 18 | 11 | 31 | 48 | 6 | 4 | *41* |
| 77. | 47 | 6 | 33 | 25 | 26 | 34 | *49* |
| 78. | 7 | 48 | 12 | 10 | 22 | 34 | *11* |
| 79. | 33 | 46 | 4 | 40 | 13 | 12 | *41* |
| 80. | 8 | 26 | 42 | 20 | 34 | 43 | *25* |
| 81. | 35 | 45 | 24 | 37 | 36 | 39 | *20* |
| 82. | 32 | 15 | 17 | 11 | 25 | 46 | *29* |
| 83. | 47 | 25 | 18 | 44 | 13 | 46 | *34* |
| 84. | 4 | 7 | 11 | 17 | 3 | 40 | *20* |
| 85. | 34 | 35 | 17 | 27 | 46 | 4 | *7* |
| 86. | 44 | 47 | 45 | 43 | 26 | 13 | *36* |
| 87. | 11 | 5 | 42 | 41 | 10 | 12 | *2* |
| 88. | 14 | 44 | 6 | 25 | 34 | 20 | *45* |
| 89. | 13 | 21 | 45 | 2 | 19 | 32 | *9* |
| 90. | 26 | 28 | 36 | 31 | 13 | 17 | *44* |
| 91. | 41 | 23 | 36 | 45 | 3 | 38 | *44* |
| 92. | 28 | 42 | 33 | 39 | 44 | 2 | *46* |
| 93. | 8 | 11 | 14 | 33 | 44 | 18 | *34* |
| 94. | 27 | 3 | 5 | 47 | 14 | 44 | *43* |
| 95. | 5 | 13 | 15 | 44 | 18 | 32 | *41* |
| 96. | 10 | 9 | 38 | 48 | 11 | 2 | *1* |
| 97. | 41 | 35 | 8 | 7 | 30 | 12 | *47* |
| 98. | 19 | 26 | 23 | 39 | 36 | 31 | *3* |
| 99. | 47 | 45 | 9 | 48 | 6 | 25 | *14* |
| 100. | 25 | 15 | 45 | 16 | 39 | 30 | *14* |

# ACTIVITY 9.5 *Olympic Swimming Records*

One of the events in the Olympic Games is the 400 metres Freestyle Swimming. The tables below give the names of the winners for the years 1928 to 1992 and the time they took (to the nearest second).

| *WOMEN* | | | *MEN* | | |
|---|---|---|---|---|---|
| *Year* | *Name* | *Time* | *Year* | *Name* | *Time* |
| 1928 | Martha Norelius | 5 m 43 s | 1928 | Alberto Zorilla | 5 m 1 s |
| 1932 | Helene Madison | 5 m 29 s | 1932 | Buster Crabbe | 5 m 48 s |
| 1936 | Henrika Mastenbroek | 5 m 28 s | 1936 | Jack Medica | 4 m 45 s |
| 1948 | Ann Curtis | 5 m 18 s | 1948 | William Smith | 4 m 41 s |
| 1952 | Valerie Gyenge | 5 m 12 s | 1952 | Jean Boiteux | 4 m 31 s |
| 1956 | Lorraine Crapp | 4 m 55 s | 1956 | Murray Rose | 4 m 27 s |
| 1960 | Chris Von Saltza | 4 m 51 s | 1960 | Murray Rose | 4 m 18 s |
| 1964 | Virginia Deunkel | 4 m 19 s | 1964 | Don Schollander | 4 m 12 s |
| 1968 | Debbie Meyer | 4 m 32 s | 1968 | Mike Burton | 4 m 9 s |
| 1972 | Shane Gould | 4 m 19 s | 1972 | Brad Cooper | 4 m 0 s |
| 1976 | Petra Thuemer | 4 m 10 s | 1976 | Brian Goodell | 3 m 52 s |
| 1980 | Ines Diers | 4 m 9 s | 1980 | Vladimir Salnikov | 3 m 51 s |
| 1984 | Tiffany Cohen | 4 m 7 s | 1984 | George De Carlo | 3 m 51 s |
| 1988 | Jane Evans | 4 m 4 s | 1988 | Uwe Dassier | 3 m 47 s |
| 1992 | D. Hase | 4 m 7 s | 1992 | E. Sadevyi | 3 m 45 s |
| 1996 | Michelle Smith | 4 m 7 s | 1996 | D. Loader | 3 m 48 s |

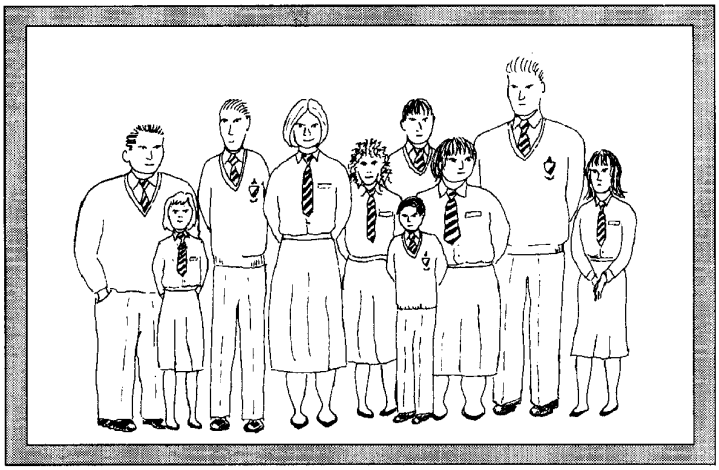
The first womens's time (5 minutes 43 seconds in 1928) is shown on the graph below.




1. (a) Plot the remaining times for the women.
   (b) In a different colour, plot all the times for the men.

2. For each set of data, draw a *line of best fit* and use it to estimate the 400 metres Olympic Freestyle Swimming times for men and women in the year 2000.

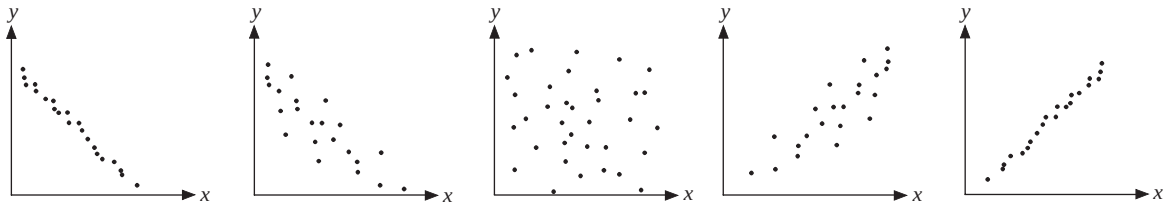
# ACTIVITY 9.6 *Correlation*



The bodies of most people are in proportion. If you are particularly tall, then you will probably also have long arms and legs and large hands, etc. The purpose of this activity is to attempt to see how clear-cut these relationships are.

We say that two variables have *positive correlation* if they increase in proportion.

Different types of correlation are sketched below in scatter diagrams.




*Strong negative correlation* | *Weak negative correlation* | *No correlation* | *Weak postive correlation* | *Strong positive correlation*

1. For your class members, or a group of 20 to 30 people, find out their

   (i) height (ii) feet size (iii) arm length

   (iv) hand size (iv) waist size (v) circumference of head.

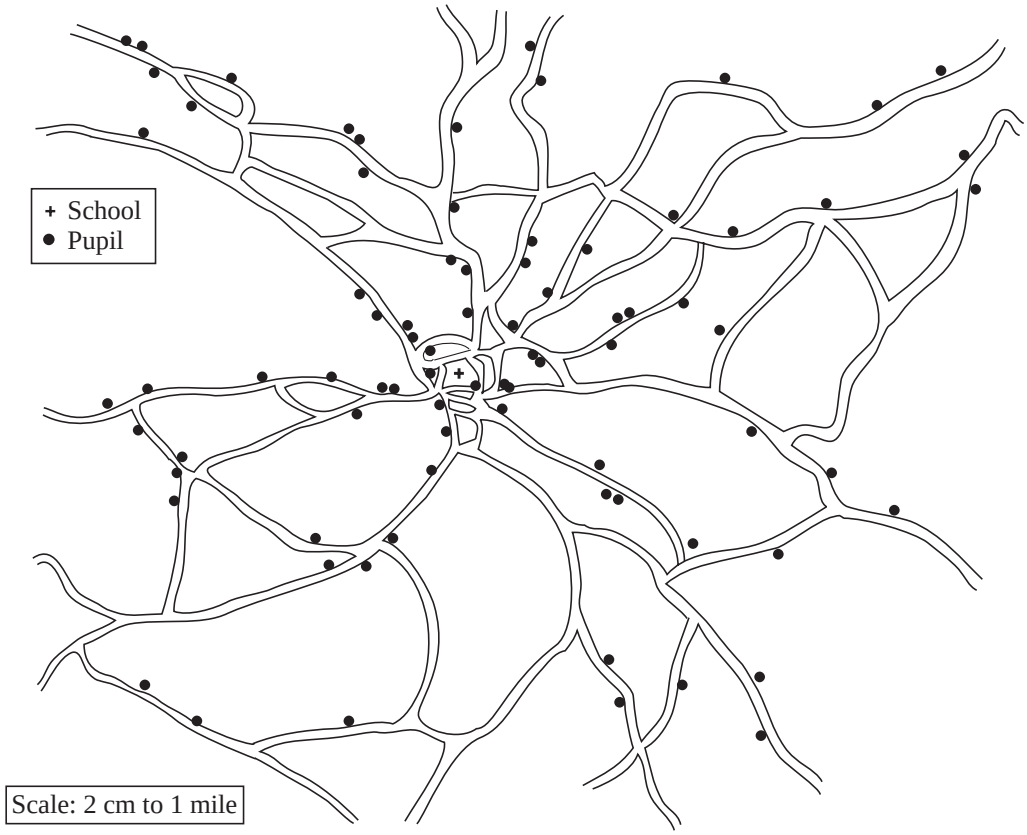
2. (a) Taking the $y$ variable as height and $x$ as one of the other variables, draw scatter diagrams for each $x$ variable.

   (b) Determine the type of correlation in each case.

# ACTIVITY 9.7 *Estimation of Mean Distances*

A small primary school wants to develop a new policy for funding contributions to pupils' travel costs to school. At the moment, only pupils who live more than 3 miles from the school 'as the crow flies' are given free transport, paid for by the local council.

To estimate the mean distances travelled by all its pupils, the Head Teacher has pinpointed on a scale map the homes of each of its pupils. This is shown below.



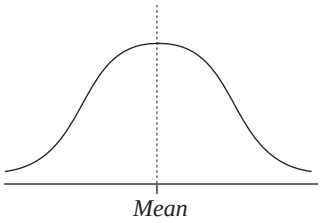

1. By first drawing concentric circles of radius 2 cm, 4c and 6cm, with the school as centre, use the data from the map to complete a copy of the table opposite.

| *Distance from school* | *No. of pupils* |
|---|---|
| Less than 1 mile | |
| Between 1 and 2 miles | |
| Between 2 and 3 miles | |
| More than 3 miles | |

2. Estimate the average distance from school of all pupils, except those who live more than 3 miles away.

3. Estimate the weekly cost of a transport subsidy which pays:
   - *nothing* to pupils who live less than 1 mile from school;
   - *£1 per mile* per week to pupils who live 1 mile or more, but not more than 3 miles, away from the school.

# ACTIVITY 9.8 *Normal Distribution 1*




The *normal distribution* (opposite) plays a key role in advanced statistical analyses, including

- significance testing
- confidence intervals
- quality control.

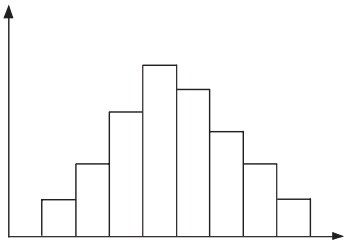
The graph is characterised by being *bell-shaped* and *symmetrical* about the mean.

Its central role is due to the fact that so much data found in the real world fit this type of distribution.

1. Find a reasonable sample (about 50 items) from one or more of the following categories:
   (a) actual weights of crisps in crisp packets,
   (b) lengths of leaves from one particular tree,
   (c) sizes (weight or volume) of pebbles on a beach,
   (d) heights of children of a similar age,
   (e) heights (or weight) of adults of one sex,
   (f) times taken by class members to run 100 metres.

2. Draw a histogram of your results, grouping the data in suitable widths.

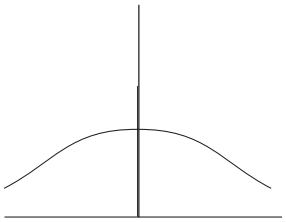
For most (if not all) of these activities you should end up with a histogram similar to the one opposite.



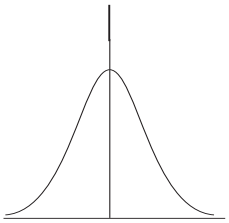
If you imagine having a much larger sample of items (say in excess of 500) and using smaller widths for each class boundary, then you will end up with a *normal distribution*.

3. (a) Combine your sample with all the other similar samples gathered by your class-mates and draw a new graph to represent the full set of data.
   (b) What do you notice about the distribution of the combined samples compared with that of your own initial sample?

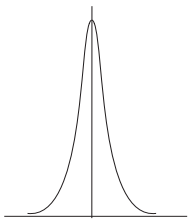
Some examples of normal distributions with different standard deviations are shown below.



*Larger standard deviation*



*Normal distribution*

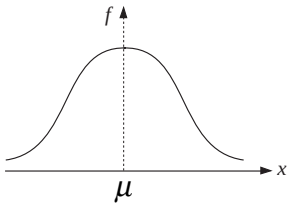


*Smaller standard deviation*

# ACTIVITY 9.9 Sheet 1 *Normal Distribution 2*

The shape of the normal distribution with mean $\mu$ and standard deviation $\sigma$ is given by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)}{2\sigma^2}} \quad \Rightarrow$$




Fortunately, you do need never use this formula! Its derivation comes from generalising the number of successes of an event in which the probability of success is $\frac{1}{2}$; this makes it symmetric, as we will see below.

The best (and simplest) example to consider is that of tossing a fair coin and noting the number of heads.

1. Toss a fair coin *twice* and let $X$ be the number of HEADS obtained.

   Copy and complete the table below.

| *No. of heads (X)* | 0 | 1 | 2 |
|---|---|---|---|
| *Probability (p)* | ? | ? | $\frac{1}{4}$ |

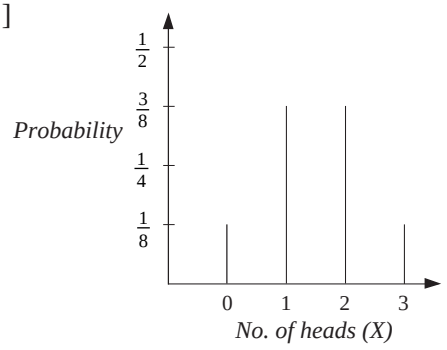
2. Carry out the experiment again but this time toss the coin *three* times.

   Copy and complete the table below.

| *No. of heads (X)* | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| *Probability (p)* | ? | ? | ? | ? |

*[Hint:* You might find it helpful to use a tree diagram.]

You can begin to see an approximation to the shape of a *normal distribution* by plotting the probability against the values of $X$.




3. (a) Repeat Question 2 for

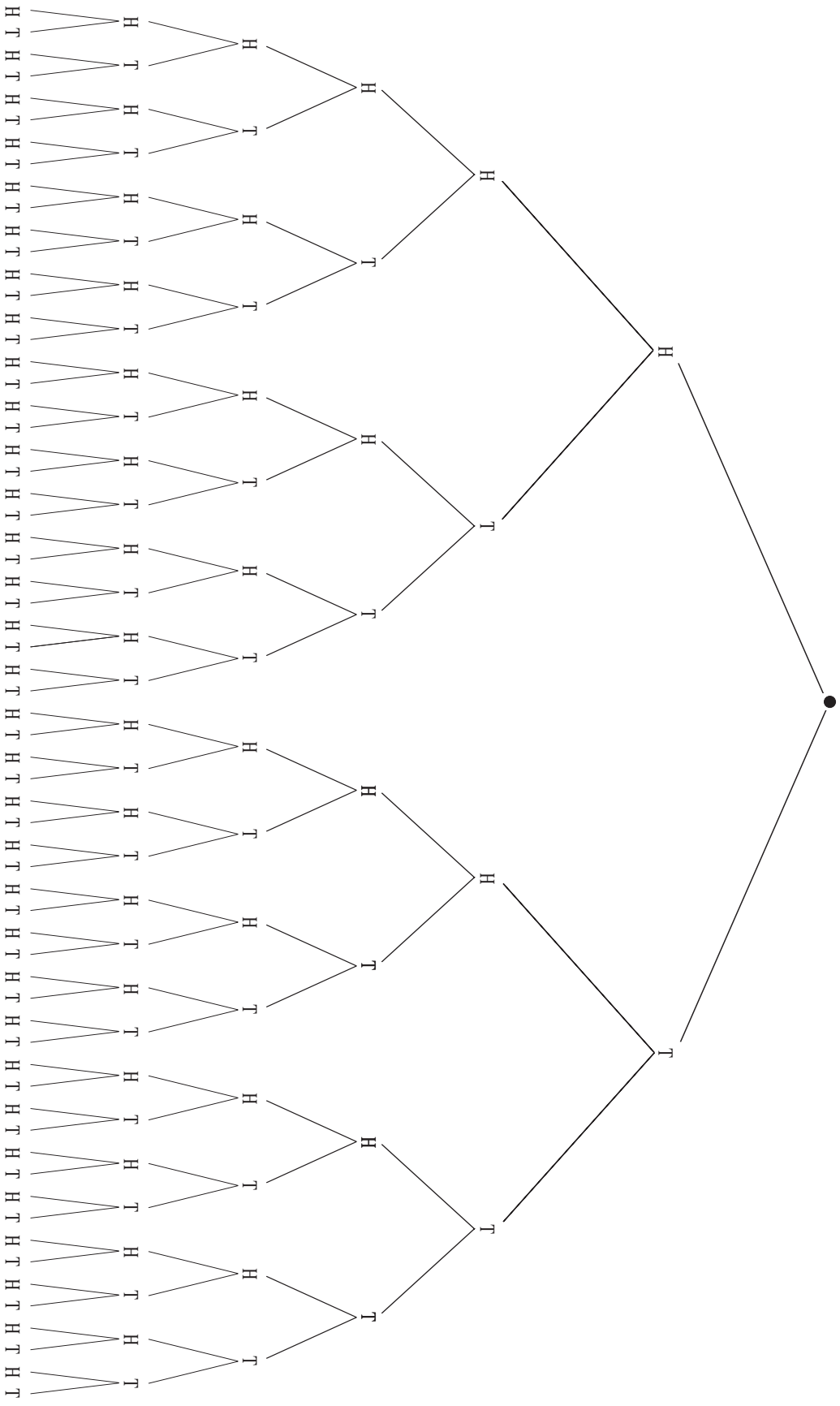
   (i) four tosses (ii) five tosses (iii) six tosses

   of the coin.

   (b) In each case, plot the probability against the number of HEADS obtained.
   [*Hint:* You might find it helpful to use the tree diagram given on a separate sheet.]

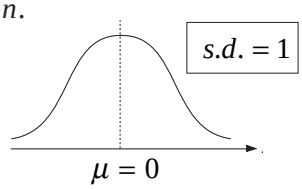
As the number of tosses increases, the closer the frequency diagram gets to the shape of a *normal distribution*. This is in fact how its formulae was first developed by the prolific German mathematician, *Carl Gauss* (1777–1855).

# ACTIVITY 9.9 Sheet 2 *Tree Diagram*



# ACTIVITY 9.10 *Standardised Normal Distribution*

In this activity, we will show how to fit a normal distribution to a set of data and then how to transform it into what is called the *standardised normal distribution*.




This is a normal distribution which has a mean of zero and a standard deviation of 1.

You will need a set of data (for example, the data from the first problem in *Activity 9.8*), for which the histogram looks roughly 'normal' in shape.

1. For your set of data, find the mean, $\mu$, and standard deviation, $\sigma$.
   [*Hint:* This is best done using the 'STATS' mode on a scientific calculator or using a spreadsheet on a computer.]

2. Transform each value, say $x_i$, $(i = 1, 2, 3, \ldots, n)$ into a new value, $z_i$, using the formula

$$z_i = \frac{x_i - \mu}{\sigma}$$

3. Find the mean and standard deviation of your transformed set of data values, $z_i$.

4. Repeat Questions 1–3 several times using other sets of data. What do you notice?

This method of transforming is a key concept in using a normal distribution with mean $\mu$, and standard deviation, $\sigma$. The formula given above will transform any normal distribution into a *standardised normal distribution* with mean 0 and standard deviation 1.

In fact, using the formula (see *Activity 9.9, Sheet 1*) for the shape of a normal distribution with mean 0 and standard deviation, $\sigma$,

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}},$$

you can check the relationship between the shapes of the three distributions on *OS 9.7 – 9.9*.

[*Note*: The number, *e*, is an irrational number. You can find its value by using the inverse of the ln button on your scientific calculator.]

5. What is the function which represents the shape of a normal distribution with mean 0 and standard deviation:

   (a) $\sigma = 1$ (b) $\sigma = 2$ (c) $\sigma = \frac{1}{2}$ ?

6. Calling these functions $y = f_1(x)$, $y = f_2(x)$ and $y = f_{\frac{1}{2}}(x)$ respectively, show that

$$2\, f_2(2) = \frac{1}{2}\, f_{\frac{1}{2}}\left(\frac{1}{2}\right) = f_1(1).$$

   Verify this result using *OS9.7–9.9*.