Statistics - Activity 1

Entering, importing and organising raw data.

Statistical measurements and observations are called raw data. Usually raw data is difficult to understand without some form of organization. Maths Helper Plus makes it possible to organise raw data in many different ways.

1) Start Maths Helper Plus now, or create a new document.

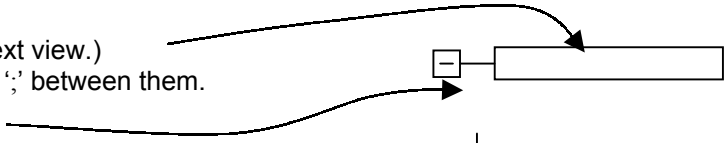
Manual data entry.
For modest amounts of data, you can enter the data manually from the computer keyboard.

The following data are measurements of the density of the earth, obtained by Henry Cavendish in 1798 using a torsion balance. Density is presented as a multiple of the density of water.

```
5.50, 5.57, 5.42, 5.61, 5.53, 5.47, 4.88, 5.62, 5.63, 4.07, 5.29, 5.34, 5.26
5.44, 5.46, 5.55, 5.34, 5.30, 5.36, 5.79, 5.75, 5.29, 5.10, 5.86, 5.58, 5.27
5.85, 5.65, 5.39
```

2) Enter these statistical data values as follows:

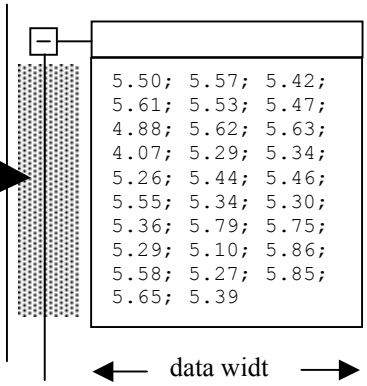
- Click on the input box. (On the text view.)
- Type the data values with a ',' or ';' between them.
- Click outside of the input box.



If no error messages are displayed, then the data has been accepted.

3) Display the options dialog for this data:

- Double click beside the data on the text view. This will display the options dialog.
- Change the 'Data Width' setting to 6 (inches) or 15 (centimetres). This will widen the data display on the text view, giving more room for tables and calculations.



4) Organise the data into a table.

Each different data value is called a score and represented by the symbol 'x'.

- Select the 'Data Tables' tab in the options dialog.
- Click to select the following columns for your table:

Table Column	Description
' x '	Displays the scores in order from smallest to largest.
' f '	Displays the number of times a given score occurs in the data set, also called the frequency of each score. The sum of the 'f' values is displayed at the bottom of the table, and equals the total number of data values.
' fx '	Displays the product of each score and its frequency. The sum of the 'fx' values equals the sum of all data values entered.
cumulative frequencies	The number of data values less or equal to this score.
cumulative percentages	The percentage of data values less or equal to this score.

Click OK to close the options dialog. To view the data table, hold down a 'Ctrl' key and press 'T' to view the text view only. Scroll to view different parts of the table.

HINT: To change the table options you have set, simply double click beside the table on the text view.

5) Use the data table created above to answer these questions:

a) Without actually counting them, how many data values were entered ? _____

b) What is the middle score (also called the median) ? _____

c) Calculate the 'range' = largest score – smallest score _____

Data entry from a text file.

It is not practical to enter large amounts of data manually. In these cases, you can import the data into Maths Helper Plus from a text file.

The data file: '1984 Olympic Records.txt' contains the national record times for men before the 1984 Olympics. The record times for eight races are listed by country.

You will import data for two races: the 100 m and the marathon.

6) Create a new Maths Helper Plus document.

7) Select the 'Import data...' command from the 'File' menu.
- This displays a 'file open' dialog.
- Click to select the file: '1984 Olympic Records.txt', then click the 'Open' button.
- The 'Import data' dialog will be displayed.

The data is displayed in the editing window of the dialog box. Use the scroll bars to study the data briefly. The first row and the first column are data labels describing the type of race and country.
- Click to select the 'data labels in first column' option.
  (Otherwise it will think the country names are data values.)

Counting data columns only, the 100 m times are in column 1, and the marathon times in column 8.
- To import data for the 100 m and the marathon only, click on the 'columns to import' edit box and type: 1,8

- Click the 'Create' button to import the data into Maths Helper Plus, then click the 'Cancel' button to close the dialog box.

8) Repeat steps (3) and (4) above to change the data width and create data tables for the marathon data and the 100 m data.

9) Use the data tables for the marathon and 100 m data to answer these questions:

a) What was the fastest time for the 100 m before the 1984 Olympics ? _____

b) What is the range in times for the 100 m (slowest - fastest) ? _____

c) What is the range in times for the marathon ? _____

d) How many countries are represented by this data ? _____

e) What is the median of the 100 m data ? _____

f) What is the median of the marathon data ? _____

10) Answer the following questions from the data tables as well as by looking at the original data file. (Look at the file by using the 'Import data...' command as in step 7 above.)

a) Which countries recorded the best and the slowest times in the marathon and the 100 m prior to 1984 ?

Marathon: fastest _____ slowest _____ 100m: fastest _____ slowest _____

b) Suggest a reason for your findings in (a) above.

Statistics - Activity 2

Putting data into classes.

If the number of scores in a statistical data set becomes too large, even sorting the data into tables is not very helpful. Using well chosen class intervals effectively reduces the number of scores and makes the data manageable.

In this activity, you will use classes to make a concise table from a statistical data set containing thousands of scores.

A class is defined by two numbers, called the class limits. Say the class limits are: 0 and 5, then 0 is the lower class limit and 5 is the upper class limit. Any data values within the class limits are part of that class.

Consider these data values: 2,4,6,10,2,0,7,1,5,6,9,8,3,4,11,3

The three boxes below represent three classes for organising this data.

For classes arranged like these, so that the upper class limit of one class is the same as the lower class limit of the next class, a score equal to the upper class limit always goes in the next higher class.

1) Write each data value from the list above in the correct box for that class:

Class: 0 to 5	Class: 5 to 10	Class: 10 to 15

All data values in a given class are considered to have the same value, called the class midpoint.

$$\text{class midpoint} = \frac{\text{lower class limit} + \text{upper class limit}}{2}$$

2) Complete this table for the classes defined above:

Class	Class Midpoint, x	frequency, f
0 - 5	2.5	
5 - 10		
10 - 15		

Now you will use classes to create a meaningful table from a huge sample of statistical data.

---

3) Start Maths Helper Plus, then load the document 'Statistics - Sunspots.mhp'.

This document contains one very large set of statistical data that consists of monthly sunspot numbers observed from Zurich between 1749 and 1983.

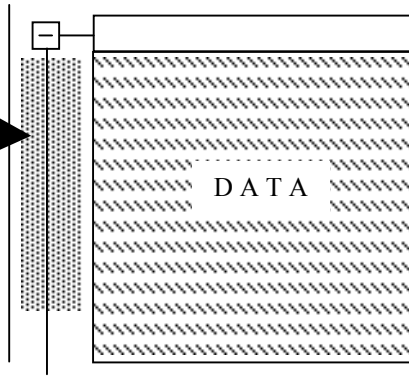
4) Attempt to organise this data into a table as follows:

a) Double click beside the data on the text view:

b) Select the 'Data Tables' tab in the options dialog.

c) Click to select the following columns for your table:
'x', 'f' and 'cumulative frequencies'

d) Click OK to close the options dialog.



To view the data table, hold down a 'Ctrl' key and press 'T' to view the text view only. Scroll to view different parts of the table.

5) The text view can only be up to 3.5 metres high. Can you view the bottom of the data table ?

As the table stands, every score occurring in the data creates one row in the table. This way, you can end up with a table having thousands of rows that is too big to display or print!

Here are the first 10 rows of the table you created:

x	f	cum.f
0	67	67
0.1	2	69
0.2	7	76
0.3	7	83
0.4	4	87
0.5	9	96
0.6	5	101
0.7	4	105
0.8	5	110
0.9	4	114

Grouping this data into classes will reduce the size of the table so that it can be displayed and printed.

Say we choose class limits, like this: 0 to 1, 1 to 2, 2 to 3 and so on.

The first row of our table now includes all scores from 0 up to (but not including) 1:

class	x (mid)	f	cum.f
0 to 1	0.5	114	114

All 114 scores that fall into this class are considered to have a score of x = 0.5 which is the class midpoint.

From the data table displayed on the text view, determine how many data values will fall into the class: 1 to 2 ? _____

6) Set up these classes to group this large data set.

a) Double click beside the data on the text view to display the options dialog box. (As in (4) above)
b) Select the 'Classes' tab.
c) For 'class number' n, set 'lower class limit' to 0, and 'upper class limit' to 1
d) For 'class number n+1', set 'lower class limit' to 1.
e) Select the option: 'apply the above class grouping to the data'.
f) Click OK to close the options box.

Examine the new table on the text view. Correct your answer to question 5 above.

7) Answer these questions from the grouped table:

a) How many scores are present in the original data ? _____

b) Comment generally on the distribution of the observed numbers of sunspots. How likely is it that a very large number of sunspots will be observed in a given month ?

c) Calculate the median number of sun spots observed in a month from this grouped table.
(Your answer will be a class midpoint value, not one of the original scores.)

8) Even this grouped data table is still very big. Try these other class intervals to group the data into fewer classes:

Firstly, try these classes: 0 to 10, 10 to 20, etc, then examine the data table on the text view.
Also try these classes: 0 to 50, 50 to 100, etc. Once again, examine the data table.

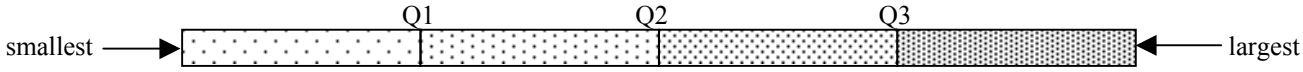
a) What class grouping to you think is best for this data ? (Explain why).

b) What are the disadvantages in making the class intervals too large ?

Statistics - Activity 3

Finding Quartiles.

What are quartiles ? When a set of data are arranged in order from smallest to largest and divided into four equal groups, then the three data values at the boundaries are the quartiles.



From smallest to largest, we have the first quartile (Q1), the second quartile (Q2) and the third quartile (Q3).

Quartiles provide a useful summary of the properties of a set of data, and are used to compare data sets with one another.

This activity explains one way of calculating quartiles of a statistical data set.

The second quartile, Q2, is found first.

For an odd number of data values, Q2 = the middle data value. (Q2 is also the median.)

For these 5 sorted data values: 1, 2, [4], 6, 9 the middle data value = 4.
So the second quartile, Q2 = 4.

To obtain the position of the middle data value for an odd number of scores, add 1 to the number of values and divide by 2. (So in this example, the middle value is at position (5+1)/2 = 3.)

What is the position of the middle value for a data set with 257 values ? _____

For an even number of data values, Q2 = the average of the two middle data values.

For these 6 sorted data values: 1, 2, [4 , 5,] 6, 9 the middle data values are 4 and 5.
So the second quartile, Q2 = (4+5)/2 = 4.5

What are the positions of the two middle values for a data set with 488 values ?

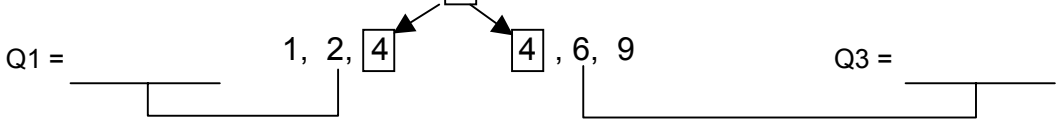
_____ and _____

The first and third quartiles, Q1 and Q3, are found next. You divide the sorted data into two equal halves, then find the middle value of each half.

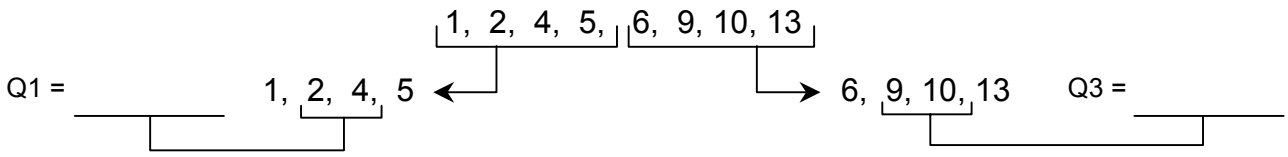
The halves may also have an odd or even number of values, so finding Q1 and Q3 is just like finding Q2.

For an odd number of data values, split the data into two halves, each half including the middle value:

For these 5 sorted data values: 1, 2, [4], 6, 9 the middle data value = 4.



For an even number of data values, split the data into two equal halves:



1) Start Maths Helper Plus, then open the document: ‘Statistics - Births.mhp’.
This dataset lists the number of live births per 10,000 23-year-old women in the United States between 1917 and 1975.

Hold down a ‘Ctrl’ key and press the ‘T’ key to display only the text view.

You will now find the quartiles: Q1, Q2 and Q3 of this data set.

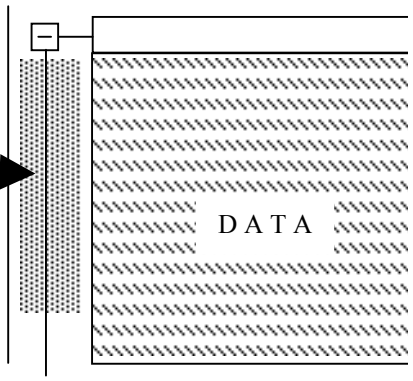
2) Create a sorted table of the data values.

a) Double click beside the data on the text view:

b) Select the ‘Data Tables’ tab in the options dialog.

c) Click to select the following columns for your table:
‘x’, ‘f’ and ‘cumulative frequencies’

d) Click OK to close the options dialog.



3) Use the information in the table and the totals at the bottom of the table to answer these questions:

a) How many data values are in this set: _____

b) What is the position of the middle value(s) in the table: _____

c) Using the cumulative frequency column in the table to help, find Q2.

Q2 = _____

d) Dividing the data into two halves, how many data values will be in each half ? _____

e) Calculate Q1

f) Calculate Q3

4) Maths Helper Plus can correct your work.

a) Display the options dialog for the data set. (As in (2) above).

b) Click to select the ‘Calculations’ tab. The value of Q1, Q2 and Q3 are shown in the middle of the dialog.

c) Click to select the option: ‘five number summary’. The complete working for finding the quartiles will be displayed on the text view. Use this to correct your work.

Statistics - Activity 4

Graphing histograms.

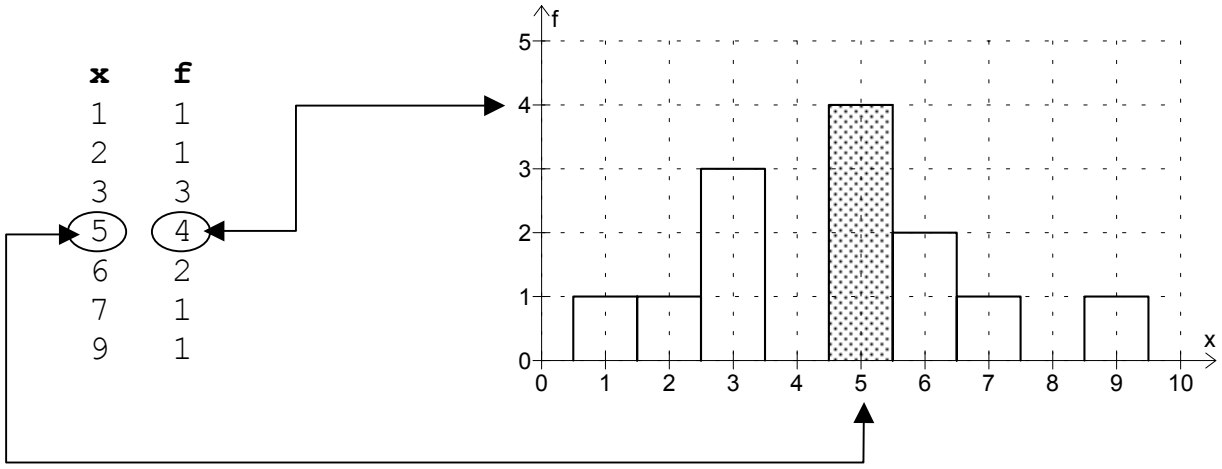
A '**histogram**' is a graphical display of the frequency distribution of statistical data. To construct a histogram, draw a rectangle for each score in the data set so that each rectangle ...

- sits on the 'x' axis so that the midpoint of the base equals the score, 'x'.
- has a height equal to the frequency of that score.

All rectangles (or bars) are of the same width, and neighbouring rectangles touch each other with no gaps.

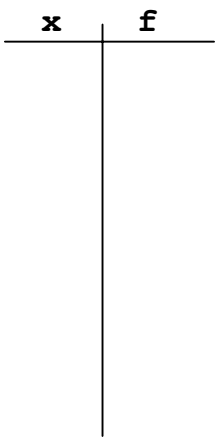
For example, consider this data: 2, 3, 3, 5, 6, 5, 7, 5, 9, 3, 6, 5, 1

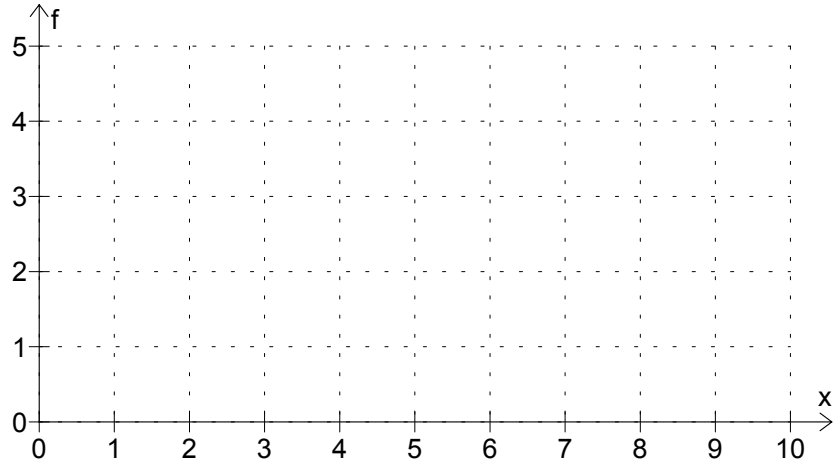
The frequency table and histogram are shown below side by side. There are seven different scores in the table, and there is a histogram bar for each one. The bar for a score of '5' is shaded. Note how it is centred on x = 5, and has a height of 4, which is the frequency of this score.



1) In the space below, construct a frequency table and then a histogram for this data:

4,7,7,2,2,1,3,2,6,3,8,7,9,7,2,4,6,2,3





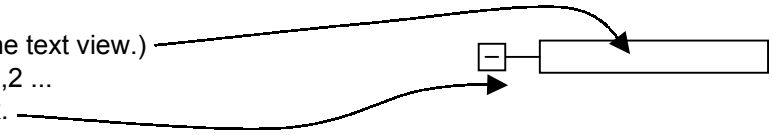
2) The following steps describe how to draw a histogram with Maths Helper Plus.
Use the data from question (1) to perform each step as it is explained.

a) Start Maths Helper Plus or create a new document. (Ctrl+N creates a new document.)

b) Enter your statistical data values.

To enter data you:

- Click on the input box. (On the text view.)
- Type the data, like this: 4,7,7,2 ...
- Click outside of the input box.



Enter the data from the last exercise: 4,7,7,2,2,1,3,2,6,3,8,7,9,7,2,4,6,2,3

c) Study the data to find out how to set up the graph scales.

For example, in this data: 4,7,7,2,2,1,3,2,6,3,8,7,9,7,2,4,6,2,3

- The scores range from 1 to 9, so to leave a gap on each side of the histogram, make the 'x' scale of the graph go from at least 0 to 10.
- The maximum frequency is for the score x = 2, which has frequency f = 5, so make the y axis scale go from zero to at least 5.

d) Set up the histogram scales.

The easy way to do this is to load the basic histogram setup template, then modify it to suit your histogram.

- Choose the 'Use Template...' command from the 'File' menu.
- Select: 'Graph setup - basic histogram.tpl' then click 'Open'.

The basic histogram sets the 'x' and 'y' scales from 0 to 10, which is just what we want for the example data.

For other sets of data, you may need to change the scale values.
To do this, press F8, then select the 'Graph Scale' tab. Type new scale limits in the edit boxes and click OK.

Here are some hints for selecting scale numbers to use with the basic histogram setup:

- The difference between the 'x' scale numbers should equal 5 or a multiple of 10
- The bottom 'y' scale value should always be zero.
- The top 'y' scale value should equal 5 or a multiple of 10

e) Customise the look of your histogram.

To choose colours and other options for your histogram, you need to display the options dialog box for your data.
Double click beside the data on the text view to display the options dialog.
Now click to select the 'Histogram, Frequency polygon' tab.

- Display the options box for your data, then experiment with the histogram display options.

---

3) Use Maths Helper Plus to create histograms for these sets of data values:

a) 2,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,5,5,5,5,6,6,6,8

b) 51,45,47,53,55,50,51,52,52,49,50,50,48,51,53,47

c)

score:	1	2	3	4	5	6	7
frequency:	2	3	0	11	15	12	4

d) 13,14,16,18,17,16,13,15,15,12,17,14,17,16,16,13,15,16,15,16,16

[HINT 1: To enter a frequency with a data value, type the frequency, then a colon, then the value, E.g.: '3**:**2' means 'three twos', in other words, x = 2 and f = 3]

[HINT 2: Ignore scores with f = 0.]

Statistics - Activity 5

Scatter plots.

Scatter plots are used to investigate relationships between sets of statistical data.

To create a scatter plot, a set of (x,y) points is plotted, where the 'x' and 'y' coordinates are the data values from two different statistical data sets, taken in a particular order.

If the plotted points in a scatter plot show no pattern and look completely random, then it is likely that there is no relationship between the variables. If the points do show a pattern, (commonly a straight line or curve), then this suggests that there may be a relationship between them.

For example, this table shows Olympic record distances (in inches) for three events over a number of years:

high jump	discus	long jump	year
71.25	1147.5	249.75	1896
74.8	1418.9	282.875	1900
71	1546.5	289	1904
75	1610	294.5	1908
76	1780	299.25	1912
76.25	1759.25	281.5	1920
78	1817.125	293.125	1924
76.375	1863	304.75	1928
77.625	1948.875	300.75	1932
79.9375	1987.375	317.3125	1936
78	2078	308	1948
80.32	2166.85	298	1952
83.25	2218.5	308.25	1956
85	2330	319.75	1960
85.75	2401.5	317.75	1964
88.25	2550.5	350.5	1968
87.75	2535	324.5	1972
88.5	2657.4	328.5	1976
92.75	2624	336.25	1980
92.5	2622	336.25	1984

It is quite obvious from the table that all three events improved as time went on.

Scatter plots can tell us a lot about the trends in these three events.

Here are some ideas:

a) Plot the high jump heights as 'y', against year as 'x'. The first two points in this scatter plot will then be: (1896,71.25) (1900,74.8)

b) Plot the discus throw length as 'x', against the long jump length as 'y'.

What will be the first two points in this case ?

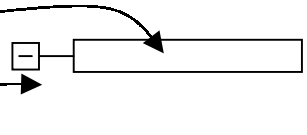
_____

Follow this procedure to use Maths Helper Plus to plot high jump heights against year.

1) Start Maths Helper Plus and open the document: 'Statistics - Olympic gold.mhp'.

2) Plot the (x,y) points for your scatter plot.
For very small data sets, you may type your points directly into the input box, like this:

- Click on the input box. (On the text view.)
- Type the data, like this: (1896,71.25) (1900,74.8) ...
- Click outside of the input box.



Most real data sets are too big to type by hand. You must then import the data into Maths Helper Plus from a text file. This has already been done in this case, but if you want to try doing this yourself, the text file is available and is called: 'Olympic Gold.txt'.

This document contains the statistical data sets from the table above. The data sets appear on the text view and are labelled: 'year', 'long jump', 'discus' and 'high jump'.

To plot a set of points as: (year,high jump), you type this into the input box: ({year}, {high jump})

A name in curly brackets is shorthand for the entire set of data that has that name, so: {year} means all of the year data, and {high jump} means all of the high jump data. Thus ({year},{high jump}) represents a set of points where the 'x' coordinates are the 'year' values, and the 'y' coordinates are the 'high jump' values. This saves a lot of typing !

Try it now! Type: ({year},{high jump}) into the input box.

You need to set the graph scale before the plotted points will appear in the graphing area...

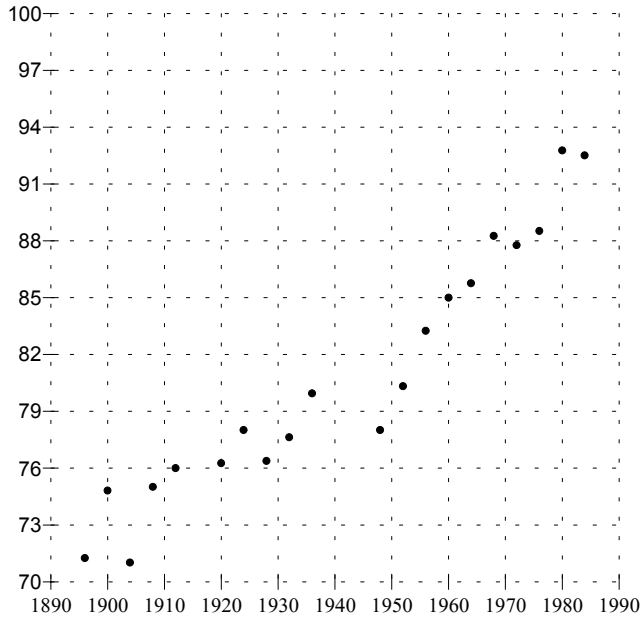
3) Set the scales so that your plotted points will all fit within the graphing area.

- The horizontal 'x' scale will display the year. The years range from 1896 to 1984, so and 'x' scale setting of 1890 to 1990 includes all values.

- The vertical 'y' scale goes from 71.25 to 92.5, so a 'y' scale setting 70 to 100 would include all values.

Press the F8 key and click to select the 'Graph Scale' tab. Change the numbers to set the scale values you require, then click OK to close the dialog box.

Your scatter plot should now look like this: →



5) Plot the discus and long jump distances on the same graph with the high jump data.
This will enable us to compare the trends in the three sports over the years.

- This plot command will plot year against discus throw lengths: ({year}, {discus})
- Use the input box to enter the plot command as in step (2) on the front of this sheet.

- What is the plot command for the long jump distances: _____
- Enter the command for plotting the long jump distances.
- Change the 'y' scale so that all data is visible. (0 to 3000 works well.)
  Can you compare the events now ?

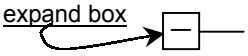
There are two problems that make this scatter plot hard to interpret:
(a) All of the plotted points look the same (i.e. red dots), and
(b) The discus distances are huge in comparison to the high jump numbers.

6) Change the look of the plotted points for each event.

- Double click with the mouse as close as possible to the centre of a plotted point. This will display the options dialog for the points. In the 'Markers' section, click the 'Colour' button and choose a different colour. Close the dialog boxes. Every point in this set will have the new look!

- Change the look of one of the other sets of plotted points as well.

[HINT: To find out which plotted points go with which event, carefully click the mouse on one point. The expand box of that data set on the text view will turn yellow.]



7) Modify the data so it is more easily compared.
One way of doing this is to divide all of the distances in each event by the first distance recorded for that event. The resulting decimal fractions show relative improvement.
We can do this by modifying the plot commands. The first discus distance recorded in the year 1896 was 1147.5 inches. So we change the plot command to this: ({year}, {discus} / 1147.5) This will divide ALL of the discus values by 1147.5 before displaying them.

- Change the plot commands for all three events. To do this, click on the plot command on the text view, make your change, then click outside of the edit box.
- Change the 'y' scale of your graph to suit the new plotted points. (From 1 to 3 is a good choice.)

8) Suggest possible reasons for the much greater improvement in discus distances since 1896 compared to high jump and long jump distances.