

The Concepts and Practice of Statistics

Part 3 - Graphics

Alan L. McLean B.Sc.(Hon), M.Sc., M.Admin., Dip.Ed. *
Melbourne, Australia

Abstract: The process of analysis of a set data aims to detect patterns, especially relationships between variables, in the data. Statistical techniques can be described broadly as techniques to detect particular types of patterns, and to test whether such patterns 'really exist'. The use of graphs to aid in pattern detection is long established. A picture can be used to make a point quickly and easily. In the context of data analysis, it is making the point more powerfully than does a table of data. A well chosen and well drawn graph makes its point - reveals a pattern - very powerfully. Equally, a badly chosen or dishonestly drawn graph makes a point incorrectly - but equally powerfully.

Reference: Aust.Orthod.J. 10(1): 15 - 20 (1987)

MeSH words: Statistics

Key words: Statistics, graphs, time series, distribution

5. A PICTURE MAY BE WORTH A THOUSAND NUMBERS

A graph is a two-edged sword. It should be approached by both writer and reader with considerable enthusiasm, but with caution.

5.1 Graphing a time series

The type of graph most commonly encountered by most people is the graph of a time series: a sequence of values of a scalar, usually continuous, variable. This is, like most graphs, a 'cartesian' graph: it uses two axes to represent the values of the two variables involved. In this case, the independent variable is time (occasionally, some surrogate for time) and is customarily the horizontal axis. In drawing graphs there are a number of conventions which sometimes seem to be laws, but are not. For example, the angle between the axes need not be a right angle; the independent variable need not use the horizontal axis. However, the conventions should be followed unless there is some good reason for not doing so.

In its simplest form, a time series graph consists of a sequence of points. Since it is a graph of data, not that of an abstract mathematical relationship, this set of points constitutes the entire informational content of the graph. But left like this, the graph will not pack much of a punch, so some enhancement is used. The usual approach is either to draw vertical bars or blocks below the points, or to join successive points by straight line segments. The results are called, not surprisingly, *bar graphs* and *line graphs*. Further ornamentation may be added, of which the most useful is to shade or colour the area below the graph.

It should be remembered that drawing a graph is only basically a mathematical technique. If a line graph is drawn, the line segments are usually *not* to be interpreted in the mathematical sense of showing values of the variable. They are *drawn only to highlight the positions of the points*, and so to show the behaviour of the series. In some circumstances, however, the lines do have mathematical meaning: a continuously recording device such as a seismograph generates a graph in the mathematical sense. Further, if values of a continuous variable are recorded at very brief intervals, the line segments can be interpreted as approximations to the 'true' continuous graph.

Whether a bar graph or a line graph is drawn is

largely a matter of individual preference, although line graphs are more common for time series. Bar graphs are often used when each data point refers to a whole period rather than to an instantaneous measurement; for example, yearly expenditure on research. If two or more series are being drawn together, line graphs should be used. The decisions to be made in drawing a graph, apart from the graphic questions of proportions, style, etc, are based on what, precisely, the graph is to show. As a general rule, a graph should be used to present *one* (simple) idea; this applies whether the graph is expository or exploratory.

For a time series, the feature that is of most interest is the *variation* in the values: how much and in what way.

It is sometimes useful to compare the magnitudes of two or more series by drawing them on the same axes. This of course requires the series to be measured in the same units. It is also necessary that the variable's axis have a full scale: that is, start at zero. If the *amount* of variation is the point of interest, it is again necessary to use a full scale; otherwise the wrong impression, usually an inflated one, is received. This is the most common crime committed in the name of graphing - drawing a graph to show the magnitude of variation, with the scale starting other than at zero.

There are times when it is appropriate to use an incomplete scale: when the interest lies in the *pattern* of variation. For example, to illustrate a seasonal effect, or to show that two series tend to move in step, it would be useful to expand the graph by using a narrow range in the vertical scale. But many people tend to use this when it is not the pattern of variation that is of interest, but its magnitude. Of course, it has the effect of making any change look greater! If an incomplete scale is used, this should be indicated clearly by showing a break in the axis in some way.

A graph may be used to show absolute variations, or relative variations. The latter may be shown by dividing the series through by the first value, so that each value is then expressed as a percentage of the first value. This is a simple form of indexing. It is particularly useful when it is desired to graph on the same axes two time series which are measured in different units, or are totally different in magnitude.

The effect of an action such as indexing is only to change the scale of each graph. This may make a superficial change in its appearance, but does not affect its basic shape, so the behaviour of the series is not affected. An alternative approach to this is to graph the changes, absolute or relative, rather than the original measurements.

* Senior Lecturer in Statistics, Chisholm Institute of Technology

Examples

Figures 1-7 show the behaviour of two artificial series:

Time	1	2	3	4	5	6	7	8	9	10
X	15	19	23	30	25	21	31	26	24	22
Y	220	190	170	150	160	200	180	210	200	230

Figure 1 shows the basic time series graph of X. Figure 2 shows the effect of drawing it as a bar graph; in practice these bars are likely to be drawn much thicker. Figure 3 shows the effect of drawing it as a line graph.

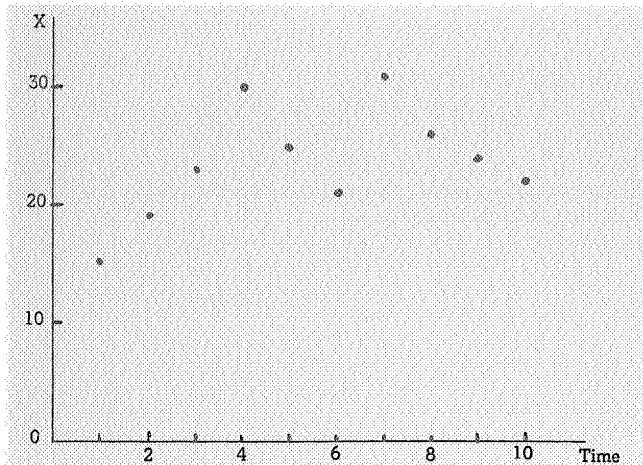


Figure 1. The basic time series graph for X, not enhanced

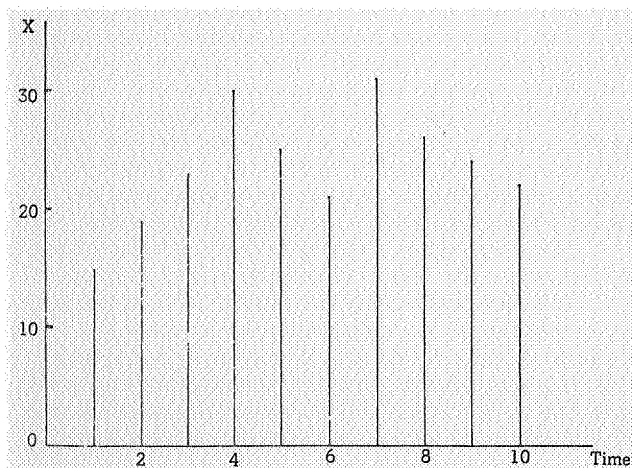


Figure 2. The time series graph for X, drawn as a bar graph

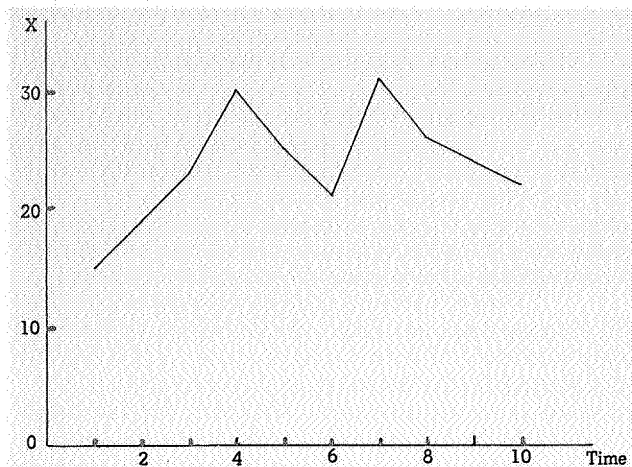


Figure 3. The time series graph for X, drawn as a line graph

In Figure 4, the series is graphed with an incomplete scale: the scale starts at 14. The use of an incomplete scale is not indicated. The amount of variation looks considerably greater. This effect is heightened by 'slipping in', unnecessarily, an expansion of the scale.

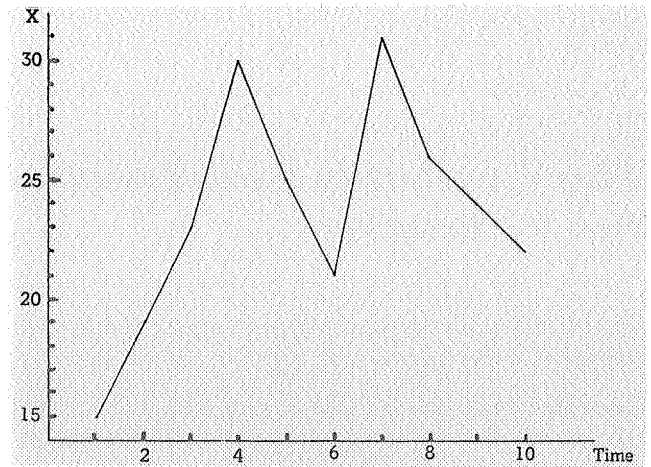


Figure 4. The line graph for X drawn with an incomplete scale; note the dramatic fluctuations in X!

In Figure 5, the two series are graphed together. Note that two scales (one incomplete) are used. This is permissible because the aim of this graph is to show that the series tend to behave in unison: when one increases, the other tends to decrease, and vice versa.

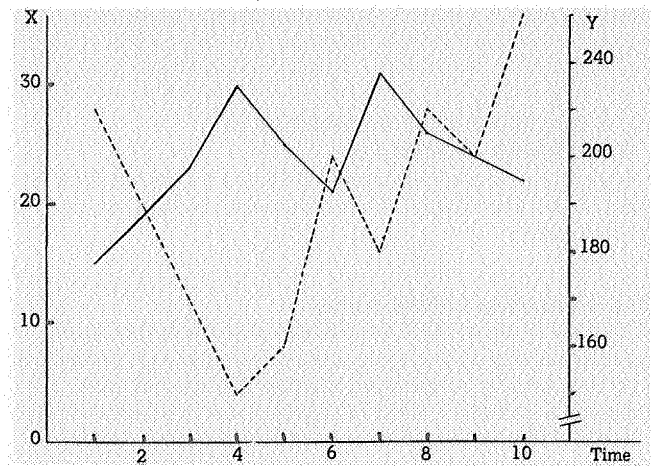


Figure 5. The line graphs for X and Y drawn together, with separate scales, to show that the variables are related

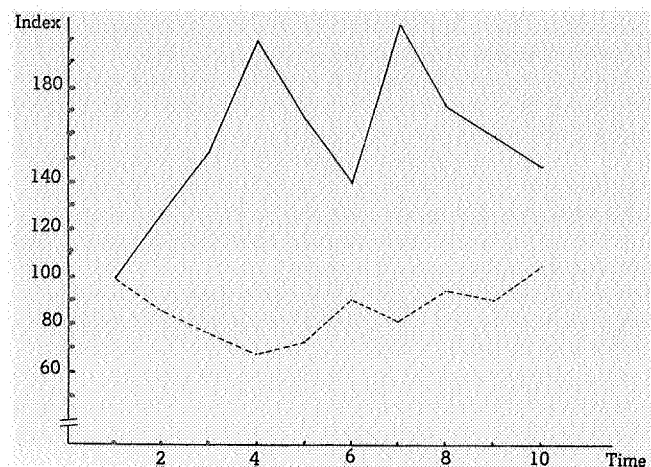


Figure 6. The line graphs for X and Y drawn together, after indexing both variables, to show that they are related

For Figure 6, the two series have been indexed. The first series has been divided through by 15 (and multiplied by 100), the second by 220, to give:

Time	1	2	3	4	5	6	7	8	9	10
X	100	127	153	200	167	140	207	173	160	147
Y	100	86	77	68	73	91	82	95	91	105

Generally, the scale used with indexing should be complete. In this case, the main reason for choosing to break the scale

was simply to fit the scale to the graph paper used – it would have been quite difficult to draw the graph with a full scale.

For Figure 7, the percentage changes have been calculated:

Time	1	2	3	4	5	6	7	8	9	10
X		27	21	30	17	16	48	16	8	8
Y		14	11	11	7	25	10	17	5	15

Note the loss of a data point.

Working with changes tends to be done in economics, where many variables change exponentially with time, so that percentage changes tend to be linear.

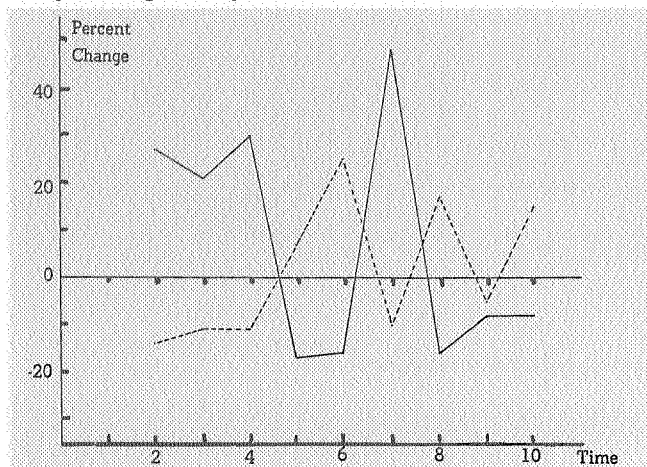


Figure 7. Line graphs showing percentage changes in X and Y, drawn together

Graphing a distribution

Given a set of values of a variable, the usual first step in 'getting a hold on' the data is to collate it into a distribution; the usual second step is to draw a graph of that distribution. The type of graph to be drawn depends on the type of variable.

For a scalar variable, the distribution consists of data pairs, with each value paired with the percentage of times it occurred, so it is straightforward enough to graph them as a cartesian graph; the percentage is the dependent variable, so it usually takes the vertical axis. As with a time series, the resultant graph consists simply of a set of points, so bars or line segments are used to highlight it, again giving a *bar graph* or a *line graph*. The result is a picture showing how the values of the variable are spread out along the scale.

What is the objective in graphing a distribution? A minor purpose may be to show the general magnitude of the values – the 'location' or 'average' of the distribution. This is particularly true when a comparison is being made between two distributions. But by far the most important reasons are the same as with a time series: to investigate the amount and pattern of the *variation* in the values.

A useful way of picturing a distribution is as a time series compressed into an instant. Example 8 shows the two ways of regarding the same set of data: taking time into account, and not. In this example, discarding the time effect is not wise, since there is a pronounced trend. In this sense, it is not surprising that graphing time series and graphing distributions have the same general objectives, since they are basically the same thing.

For a variable which takes only a few different values, the bar graph or line graph is quite straightforward. For a variable which takes many different values, the graph is likely to be useful only for a very large set of values. For a small set, each value will occur only once or twice, if at all.

Such a variable is, for practical purposes, continuous, and is often called so even if it is actually discrete. As discussed in Part II, the usual approach to treating a continuous distribution is to work with the average percentage; or better, to use the continuous model variable. In the latter case, the average percentage would be called the *percentage density*.

Since for a continuous variable the percentage density has values for all values of the variable, the resultant graph is continuous (technically, it is piecewise continuous). It is also, for each class, horizontal, so the graph consists of a sequence of horizontal line segments, giving the appearance of a set of rectangular blocks. Such a graph is called a *histogram*.

The word 'histogram' is often used for any graph which looks like a set of rectangles; including a bar graph with wide bars. This is very loose usage, and is not recommended. Note that the correct way of drawing a histogram follows naturally from this development. In particular, the vertical axis is not 'percentage', as with a bar graph, but is 'percentage density'.

For a bar graph, the height of each bar represents the percentage of observations; so the total of bar heights is 100 percent. For the bar graph obtained when grouping, the total height of the bars within each class represents the percentage for that class. For the histogram, the *area* of each block represents the percentage for that class, so the total of the block areas is 100 percent.

It follows from the mathematical logic that the class percentage is represented by area. But this is also correct by the logic of perception. Quite simply, we perceive areas.

Examples

In Figure 8 the following data set is graphed:

Time	1	2	3	4	5	6	7	8	9	10	11	12	13
X	4	4	5	5	4	5	4	5	6	4	6	7	5
Time	14	15	16	17	18	19	20	21	22	23	24	25	
X	7	9	6	7	8	7	6	9	8	9	9	8	

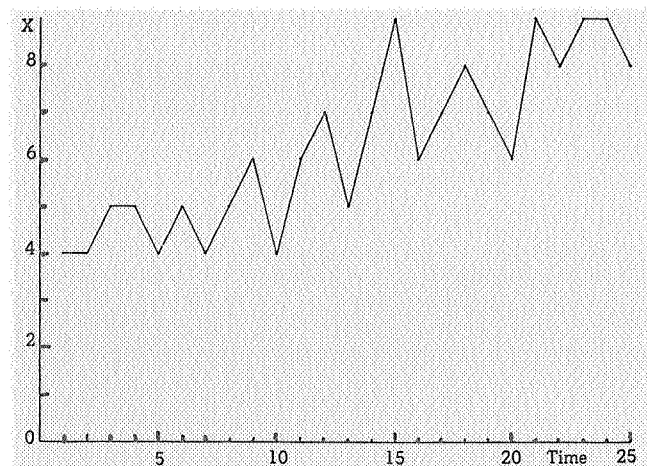


Figure 8(a). Data for the (second) variable X, graphed as a time series

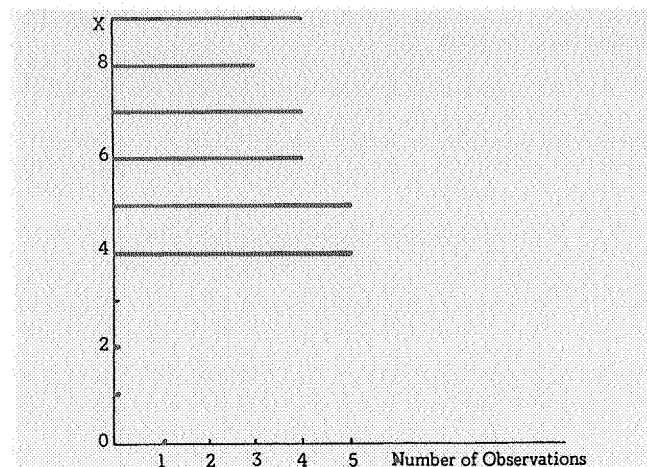


Figure 8(b). Data for X, graphed as a frequency distribution

In Figure 8A the time series is drawn; in Figure 8B the series is collapsed into the frequency distribution:

X	Number of observations
4	5
5	5
6	4
7	4
8	3
9	4
	25

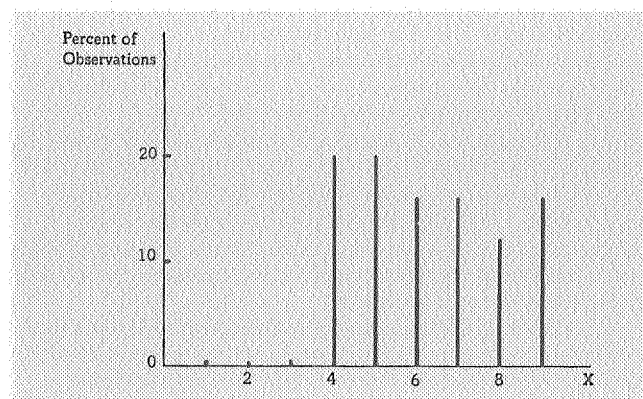


Figure 8(c). The percentage distribution for X with usual orientation

In *Figure 8C* the frequencies have been converted into percentages (of 25) and the graph drawn in the usual way. Graphs 8B and 8C have the same shape - their only difference is in orientation.

This data set has no clearly defined central value; in this sense it shows quite a lot of variation. One of the sources of this variation is the *trend* that is clearly seen in the time series. If the purpose of analysing the data is to forecast future values of X, then replacing the time series by the distribution may have been a mistake.

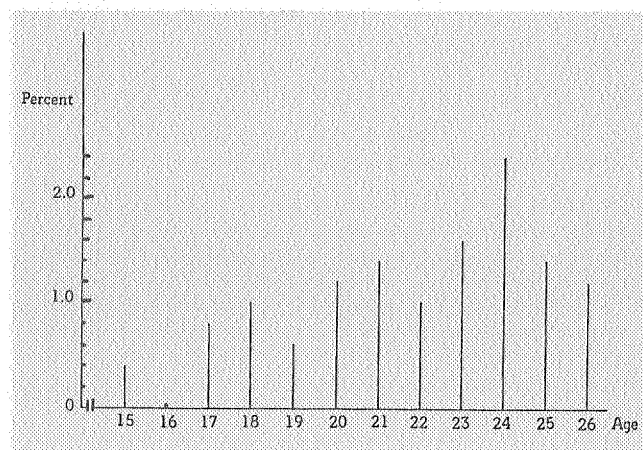


Figure 9(a). The percentage distribution for Age, using raw values, graphed as a bar graph

In *Figure 9* we look again at the variable 'age'. Again the original data consists of a sequence of values which can be graphed as a time series, or the order of the sequence ignored and the data graphed as a bar graph (*Figure 9A*). Suppose the first few values are:

Age	Percent of observations
15	0.4
16	0
17	0.8
18	1.0
19	0.6
20	1.2
21	1.4
22	1.0
23	1.6
24	2.4
25	1.4
26	1.2
	etc.

This could be grouped as:

Age	Percent of observations	Average percent
15-19	2.8	0.56
20-24	7.6	1.52
25-29	11.4	2.28
	etc.	

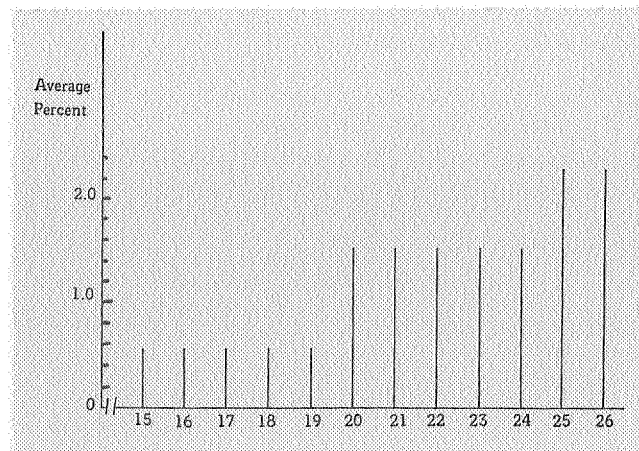


Figure 9(b). The percentage distribution for Age, using average class values, graphed as a bar graph

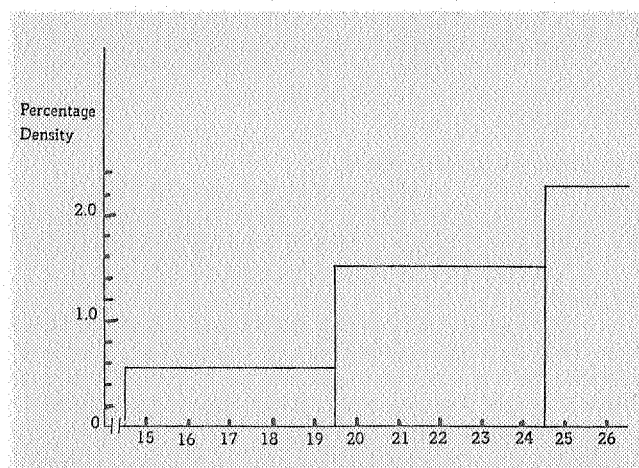


Figure 9(c). The percentage distribution for Age, graphed as a histogram

In *Figure 9B* the bar graph corresponding to the grouped data is shown; in *Figure 9C* the corresponding histogram (the graph of the continuous model) is drawn.

Suppose that the complete age distribution is:

Age	Percent of observations	Average percent
15-19	2.8	0.56
20-24	7.6	1.52
25-29	11.4	2.28
30-39	21.6	2.16
40-59	45.2	2.26
60-64	5.4	1.08
65 and over*	6.0	0.30

* upper limit taken to be 84

Following the procedure described above, the average percentage in each class is found. Note that to do this for the final class ('65 and over') an upper limit must be assumed. If the raw data is available, the correct midpoint of this class is the mean of the values in the class. If this data is not available, a reasonable estimate has to be made.

Commonly, a distribution tails off at the upper end, so if the grouping is well done the upper class will contain only a small percentage of observations. Hence the error in choosing the upper limit should be small.

The correct graph is shown in *Figure 10A*. *Figure 10B* shows a common error in drawing the histogram: using 'class percentage' as a vertical axis. As can be seen, the visual error is

considerable. *Figure 10C* shows another, even worse error: treating the classes as discrete values, and drawing a bar graph.

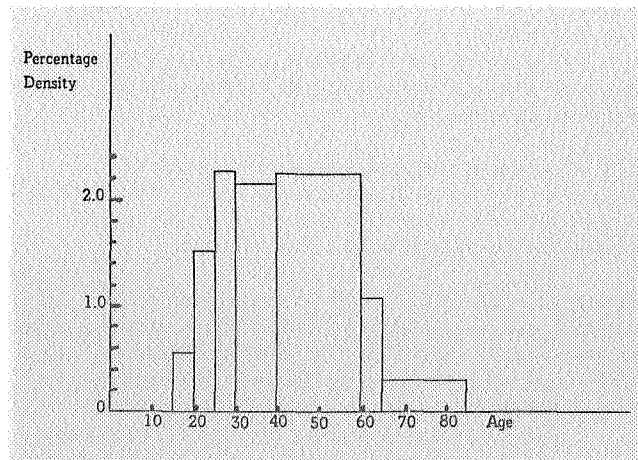


Figure 10(a). The complete distribution for Age, graphed as a histogram

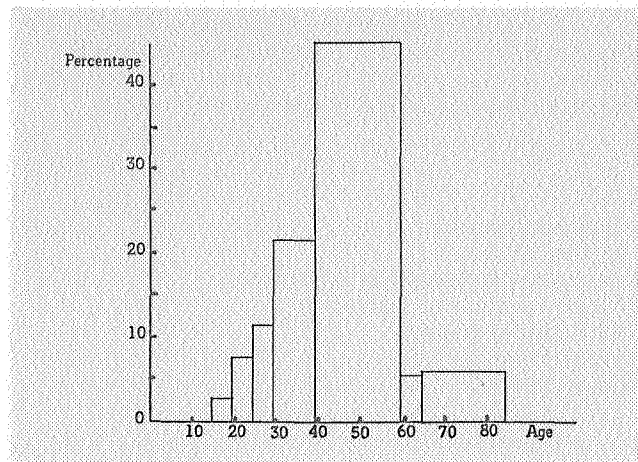


Figure 10(b). The complete distribution for Age, graphed incorrectly with percentage as vertical axis. Note the visual error

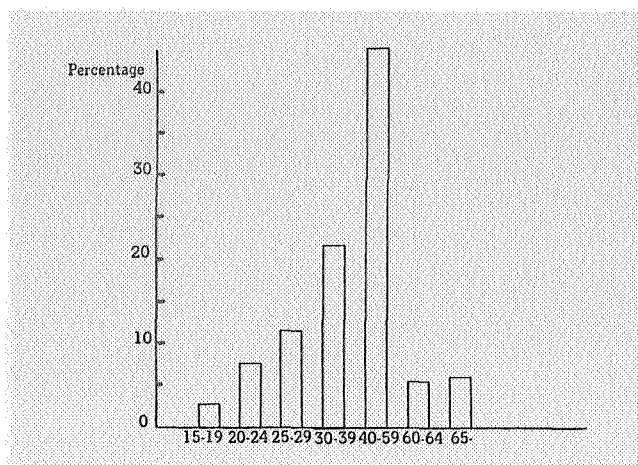


Figure 10(c). The complete distribution for Age, graphed incorrectly with percentage as vertical axis, and with incorrect horizontal scale

To graph the distribution of a *nominal* variable one can proceed as for a discrete scalar variable and generate a bar graph. By definition, however, there is no scale for the variable, so the location - and even the order - of the points representing the variable values is quite arbitrary. The resultant picture can be changed simply by changing the scale or order. It follows that such a graph is of very little use.

It is more common to use a more primitive form of graph for a nominal variable: a pie graph, in which the percentage breakdown is shown by the sectors of a circle. Note that all a pie graph does is to convey the relative sizes of the individual percentages.

5.3 Graphing a cumulative distribution

The graph of a cumulative distribution is called an *ogive*.

For a discrete variable, the ogive is a step graph, with jumps at the variable values. For a simple distribution, the bar graph consists of spikes; at any point between the spikes, the percentage is zero. For a cumulative distribution, this is not so. For the distribution used in *Figures 8 and 11*, there are no observations *equal to* 4.5, for example, but there are five observations *less than or equal to* 4.5. Hence the graph consists of steps.

For grouped data the approach is analogous to that for a simple distribution. Calculate the *average (percentage) increase per unit* of the variable, to give the approximate cumulative distribution; then *model* it by using a straight line approximation. As with a histogram, points are drawn for the 'true' class end-points, and joined by line segments. For a histogram, the segments differ in *level*; for an ogive they differ in *slope*.

Examples

For the example of *Figure 8*, the cumulative percentages are:

X	Percentage of observations less than or equal to X
4	20
5	40
6	56
7	72
8	84
9	100

This is graphed in *Figure 11*.

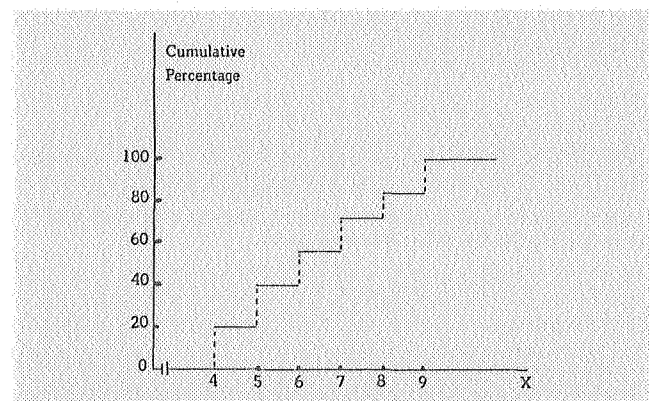


Figure 11. The graph of the cumulative distribution of X

For the example of *Figure 9*, the lower end of the cumulative distribution is:

Age	Percent of observations less than or equal to age	Approximate percent of observations less than or equal to age
15	0.4	0.56
16	0.4	1.12
17	1.2	1.68
18	2.2	2.24
19	2.8	2.80
20	4.0	4.32
21	5.4	5.84
22	6.4	7.36
23	8.0	8.88
24	10.4	10.40
25	11.8	12.68
26	13.0	14.96

This is graphed in *Figure 12A*. Using the average percentages gives the last column, which is graphed in *Figure 12B*. The model graph is shown in *Figure 12C*. *Figure 13* shows the complete ogive for the age distribution.

5.4 Statistical packages

The major statistical packages such as SPSS^X and BMDP include capacity for drawing graphs. They seem at present to be less than satisfactory. This is partly due to limitations in printers,

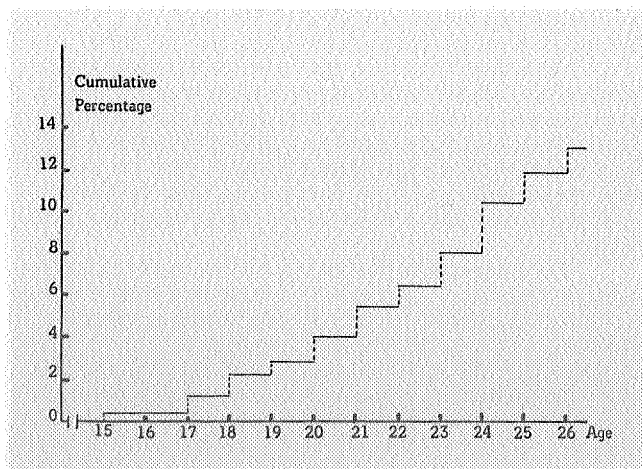


Figure 12(a). The graph of the cumulative distribution for Age, using raw values, drawn as a step graph

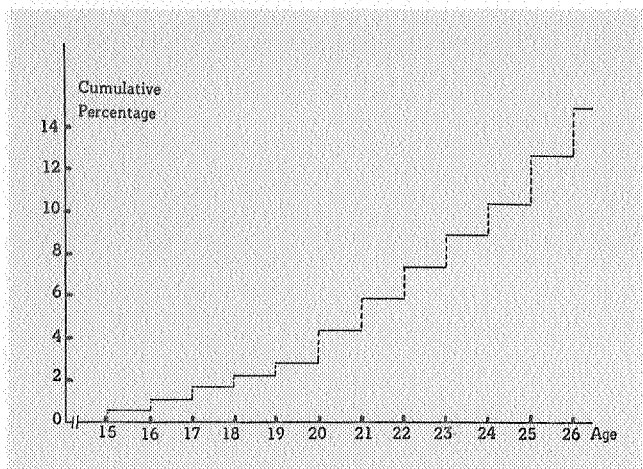


Figure 12(b). The graph of the cumulative distribution for Age, using average class values, drawn as a step graph

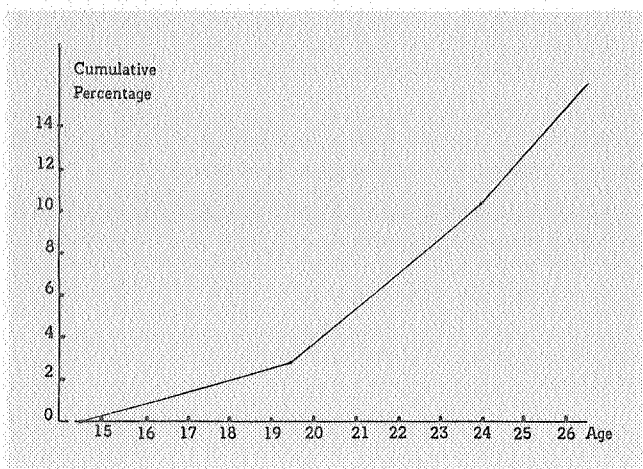


Figure 12(c). The graph of the cumulative distribution for Age, drawn as a line graph

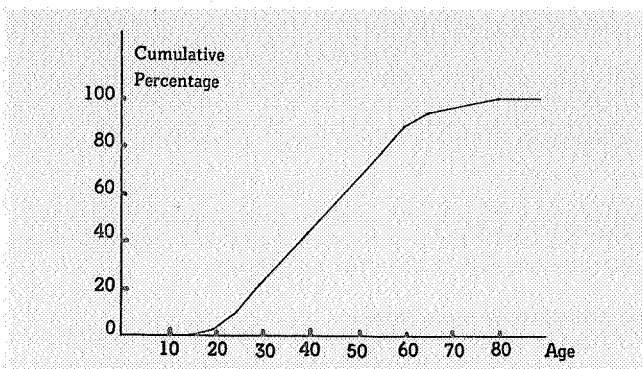


Figure 13. The graph of the complete cumulative distribution for Age

which are not specifically designed for graphics, and partly because good graphics - and layout of results generally - appear to have been low on the list of priorities of the software writers.

Microcomputers, with their greater emphasis on graphics, should offer more scope, but the software still seems to be inadequate.

5.5 Some general comments on graphs

The types of graphs discussed above are the standard ones underlying univariate statistics. In a later article graphs underlying bivariate statistics will be discussed.

There are a number of variations on these graphs, all aiming to show the distribution pattern. 'Exploratory data analysis' includes such graphical techniques as stem and leaf diagrams (see Section 4.4, page 284 AOJ vol 9 no 3, 1986) and box plots. One can use normal probability paper to graph data, to see if its distribution is 'normal'.

With any set of data there may also be particular graphs which can be drawn to show some features peculiar to that set of data. In any analysis it is worthwhile looking for such special graphs.

It is worth finishing this section with a reminder that graphics are very powerful. This power should be used correctly: to expose rather than to conceal; to clarify rather than to confuse; to convey truth rather than untruth.

Correspondence to:
Mr Alan McLean
Department of Mathematical Sciences
Chisholm Institute of Technology
900 Dandenong Road
Caulfield East
Victoria 3145