Pitfalls of Statistical Presentation

S.A. Grenville and I.J. Macfarlane
Reserve Bank of Australia

Paper presented to the 17th SEANZA Central Banking Course,
Kirribilli

11 October 1988

1. Introduction

This paper is not intended to be a technical piece on economic statistics, but to be of some help to people who are not professional statisticians or who do not spend most of their time dealing with statistics. It sets out a few simple rules that can help make statistics illuminating, rather than confusing or, even worse, deliberately misleading. If there is one major aim of the paper, it is to encourage readers to develop a healthy scepticism towards statistics that are presented to them, and to make readers aware of how to ask relevant questions if a statistic which is presented to them does not seem right.

There are two main parts to the paper. The first deals with how raw data should be viewed or processed in order to get them to yield useful information. The second deals with the presentation of statistics through tables, graphs and econometric procedures. The conclusion summarises some simple rules which can help in the use and presentation of statistics.

2. Collecting and Processing Statistics

(a) Choosing the data

When analysing a subject, a user of statistics is often faced with a mass of data. From these, he needs to select and pre-process, so as to separate important statistics from those less important and reduce the task of interpreting the data to manageable levels.

The first rule is to "know the data". Users should understand how the statistics are collected and what simplifications and short-cuts have been used in their compilation.[1] There is the well-known example of the economic researcher who produced an econometric study showing that he could predict how travel receipts in the balance of payments would move. The equation was able to predict with remarkable accuracy the movements in these receipts by relating them to short-term arrivals of passengers. What the economic researcher did not know was that the official balance of payments statistics were also compiled in exactly the same manner; the country's statistical office merely collected the figures for short-term passenger arrivals and multiplied them up by an average dollar amount to obtain an estimate of travel receipts.

Most of the data economists use are from samples rather than full enumeration; this raises a series of possible problems relating to sample bias. While most official data are free from the most blatant forms of sample bias (such as sample self-selection as typified by the telephone "ring-in" poll), they are not always free from subjectivity. An example is the response to income and expenditure surveys, where it is difficult to reconcile individuals' expenditures with their income. Sampling errors are also important. If the increase in the consumer price index this quarter is 0.2 higher than last quarter, this may not have much significance if the statistician reports that the standard error for his CPI estimate is 0.5 per cent.

Another important lesson when using data is always to be suspicious about any very large movements. A good rule of thumb is that if a number *looks* wrong, it probably is wrong.

At the very least, any large movement merits further investigation; it may be an incorrect figure, a large movement in response to an identifiable cause, or a legitimate movement of unknown cause. Whatever it is, the user and presenter of the data should be able to account for it.

(b) Pre-processing

While it might be nice in theory to present all the data and let the reader draw his or her own conclusion, even the most objective person must do some pre-processing to bring some generalisations out of the mass of data. There are a number of reasons for this. One is to help the reader draw the most plausible conclusion. The other is to save time for the reader. A particular statistic may be collected annually, quarterly and monthly, be available in seasonally-adjusted and seasonally-unadjusted form, and in nominal and real terms. The reader should expect to have the field narrowed down to easily digestible proportions. The main forms of pre-processing are as follows:

(i) Averages

Averages are a simple concept and allow large volumes of data to be reduced to manageable proportions. They also reduce the risk of using an observation that may be unrepresentative. But, there are traps. Averages can hide important variations. The story of the statistician who drowned while crossing a river that had an average depth of one metre provides a good example.

Some important points to remember when using averages are:

- do not mix unlike periods - the average should be representative of the period;
- there is a role for medians rather than means (especially with income statistics); and
- it may be useful to give some notion of distribution; perhaps a range will be more appropriate than a single figure.

(ii) Growth rates

It is often very useful to summarise the behaviour of a statistical series by calculating its growth rate, rather than showing a number of observations of its level. In doing so, however, it is most important to use good practice; it is easy to mislead the reader by using the wrong method of calculating the growth rate.

Unfortunately, there is no unique way of calculating a growth rate and, of course, the period over which the growth rate is calculated is to some extent a matter of discretion. Many arguments and misunderstandings about the state of the economy, for example, arise simply because

1. For a good discussion of these problems, see Morgenstern (1963), Chapter 2.

different people have used different ways to calculate growth rates of economic time series.

The three main ways of calculating a growth rate are:

- a "year-on-year" growth rate;
- a "through-the-year" growth rate; and
- a growth rate for the last observation over the previous one.

For people who are regular users of economic time series, the difference between the first two growth rates becomes second nature to them, but for those who have not had a lot of exposure to statistics, it is necessary to become familiar with these two different concepts.[2] Although the differences between these two growth rates is commonly encountered in practical statistical applications, it is not covered in any of the standard statistical textbooks. The difference between these two arises because most statistics are available in different periodicities, e.g. annual, quarterly and monthly. The third concept is relatively straightforward so little needs to be said about it.

An appendix sets out an example of where a "year-on-year" growth rate would be misleading and another example of where a "through-the-year" growth rate would be misleading. It is useful to go through the arithmetic to get a feel for the mechanics involved. The principles, however, can be stated simply:

- if you are worried about being misled by large atypical movements, you should use a growth rate that contains a lot of smoothing but suppresses the influence of recent movements. This would point you towards a "year-on-year" growth rate;
- a "through-the-year" growth rate also contains some smoothing but less so than the "year-on-year" growth rate. It should be used if the within-year movement in the most recent year suggests a different trend to that shown by the "year-on-year" growth rate (see Appendix);
- if there is a high priority on being up to date, and you are prepared to tolerate aberrations (because you can explain them, or perhaps you are wishing to highlight them for some reason), you would use the growth rate of the most recent observation. Some people, for convenience, will often scale up the growth rate to an annual rate, but this can be misleading as it tends to give the impression that the growth rate has been sustained for a year when it may only be a monthly or a quarterly aberration.

A lot of judgment is required in choosing the most appropriate growth rate, but it is possible to provide some guidance. For example, in a time series where there is a lot of semi-random variation from quarter to quarter, such as the national accounts, there is a tendency to rely on "year-on-year" growth rates. Where there is no sampling error or unexplainable variation, such as some population statistics, use of a growth rate that involved smoothing would be unnecessary and only introduce an unhelpful lag into the interpretation.

(iii) Seasonality

Economic data often show strong seasonal patterns. In order to be able to interpret movements, it is important to be able to distinguish between a movement which is simply the typical seasonal change and one which represents a more fundamental shift.

Seasonal patterns can arise for various reasons. Agricultural production or exports, for example, usually have pronounced seasonal patterns because planting, harvesting, etc are dictated by weather conditions. Other seasonal patterns are due to conventions such as school and public holidays; some of these, such as Easter, are moveable feasts. An example of a series that has a large seasonal pattern due to public holidays is holdings of notes and coin; they tend to rise sharply ahead of holidays.

There are a number of standard methods of seasonal adjustment which attempt to remove the "purely" seasonal movements in an economic time series so as to make the underlying movements more apparent. In essence, they estimate the *average* seasonal factors (for each month or quarter) that have operated over a run of years. These average factors are applied to the raw data to produce data which are seasonally adjusted.

Problems can arise when the seasonal pattern is changing. In this case, the seasonal adjustment procedures, which rely on measuring the average seasonal patterns over a number of years, will fail to remove some purely seasonal movements and will adjust for some seasonal patterns that are no longer present. In principle, there are statistical tests to tell whether seasonality is stable or unstable. In practice, they are relatively weak, and there is no substitute for a full understanding of the data and the extent to which they might be affected by an institutional change that has caused seasonal characteristics to vary.

There is a common misconception that the difficulties of seasonal adjustment can be avoided by using unadjusted data and calculating growth rates on a "twelve-months-ended" basis - i.e. looking at growth this January over last January, this February over last February, etc. This is, in fact, just a crude form of seasonal adjustment. It has the advantage of simplicity, but it does not avoid any of the other pitfalls of seasonal adjustment - e.g. changing seasonality. It also reduces timeliness in that it reflects growth over the previous eleven months, not just the latest month.

In particular, if a "twelve-months-ended" growth rate falls when the calculation is moved ahead one month, it does not necessarily mean that growth in the most recent month was weak; it may only mean that the growth rate twelve months ago (which was dropped out of the calculation) was strong. (See the Appendix for an example of this.)

2. "Year-on-year" and "through-the-year" are the most common terms used to describe these growth rates, though they are not the only ones. The "year-on-year" growth rates are often also referred to as "year average on year average" or "average of period over average of previous period". The "through-the-year" growth rate is often referred to as "growth over the course of the year", "end point on end point", "twelve-months-ended" or "four-quarter-ended" growth rate.

In general, seasonally-adjusted statistics should be used where available, as failure to do so can produce misleading results. However, as in all uses of statistics, it is necessary to know the data and beware of distortions caused by changing seasonal patterns.

(iv) "Normalising" or "Standardising"

It is often necessary to "normalise" or "standardise" data before a meaningful comparison can be made. For example, in comparing air safety over time, it would be misleading to compare the number of fatalities in 1920 with the number in 1988. Allowance needs to be made for greater use of air travel in 1988. The data should, therefore, be "normalised" by dividing by, say, the number of passenger-miles flown. The essence of normalisation is to ensure that like is compared with like, to focus on the elements which are of interest, and to eliminate any extraneous factors.

Normalisation can take many forms. The most common form in economics is to allow for changes in prices. For example, in comparing government expenditure on defence today with income twenty years ago, one would want to take account of how prices have moved over that period - i.e. changes in real expenditure are probably what are of most interest. Many economic time series, particularly national accounting statistics, are therefore measured in real terms as well as nominal terms.

Another form of normalisation is to calculate ratios - e.g. income per head, consumption as a proportion of national income, external debt as a proportion of national income, etc. This form of normalisation allows for the fact that most variables grow in real terms as well as due to inflation.

There are many other forms of normalisation; all are variations on a simple theme - the need to compare like with like and in a way which is relevant to the particular problem at hand.

3. Presentation of Data

Having drawn together and analysed a series of numbers on a particular subject, the user will normally want to present the findings. Communication of results can be greatly assisted by the use of tables and graphs.

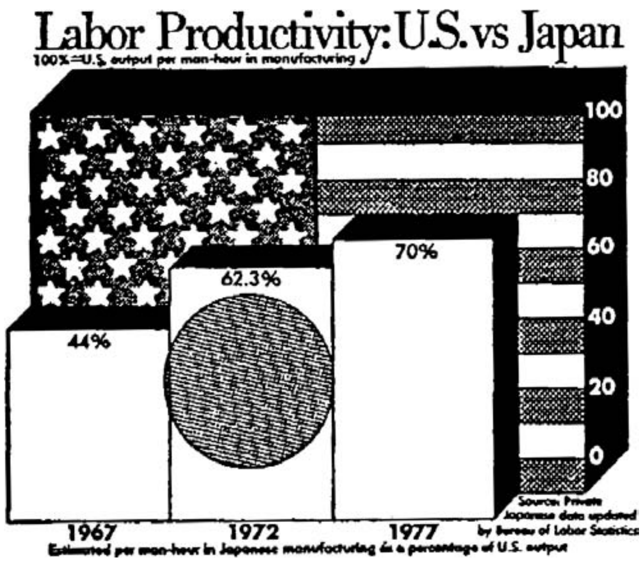
However, tables and graphs are not always necessary or desirable. The key point in getting a message across to the reader is "simplicity". Tables or graphs should only be used if they are necessary to illustrate a point which the user wants to make. If they are not necessary, they will only make the reader's job harder. Data should not be presented simply as a matter of reference, other than in specialised volumes.

In some cases, the data can be condensed to a couple of very simple facts and it is more efficient simply to present it as part of the text. The example below illustrates an unnecessary graph.

Despite the mass of lines, shapes and shading,[3] this graph has only three observations. The information could just as easily have been conveyed by writing a sentence saying that "labour productivity in Japan was 44 per cent of that in the United States in 1967, but it had risen to 62.3 per cent in 1972 and 70 per cent in 1977".

In deciding how to present the data, always ask the question "Do I need to use either a table or a graph or is the information so simple that it can be presented as part of the text?"

Graph 1



From The Washington Post, 1978.
Reprinted in Wainer (1984).

ADMIRALTY STATISTICS

Classification	*Year* 1914	*Year* 1928	*Increase or Decrease %*
Capital ships in commission	62	20	−67·74
Officers and men in RN	146,000	100,000	−31·5
Dockyard workers	57,000	62,439	+9·54
Dockyard officials and clerks	3,249	4,558	+40·28
Admiralty officials	2,000	3,569	+78·45

From Parkinson (1979)

3. See later for discussion of "chartjunk".

Tables

A table is more neutral than a graph in presenting data as it does not attempt to influence the senses by visual display. For this reason, a table should be used if there is no particular strong tendency that can be illustrated by a graph. A table is also useful if the titles of the categories are complicated, and it has the advantage that it can express exact figures rather than plots on a graph. Even so, it is important to keep the table as simple as possible. There is no need to include any information that is not essential for the telling of the story. Tables of reference should be included as an appendix.

A good example of a well-designed table is shown below. In his celebrated essay, C. Northcote Parkinson illustrated the thesis that bears his name with one well-designed table.

This has all the good characteristics of a well-chosen table and, most importantly, it fits comfortably into less than half a page and so illustrates the text that surrounds it, rather than replacing it.

There is not a lot more to say about construction of tables other than the three simple rules:

- keep them small;
- do not include information unless it is necessary to illustrate the argument in the text; and
- if it is felt necessary to present additional reference material, present it in an appendix.

There is one other point that should be borne in mind by people who are responsible for producing regular reports. Do not always use the same table. Information that was important when the table was first devised may no longer be of much interest. To continue to show it clutters the story. If the text is changing from month to month in order to keep up with events, it is almost certain that the table will have to change if it is still to be used to illustrate the text.

Graphs

(a) Some simple rules

The subject of graphs is more complicated and much has been written about it.[4] Points to remember when displaying data graphically are very similar to the those outlined above for tables. The main requirements are to make sure that they help to illustrate the point that is being made in the text, that they are simple and that they do not deliberately distort the facts. It is the capacity for graphs to deliberately mislead that makes the subject of graphical presentation so interesting.

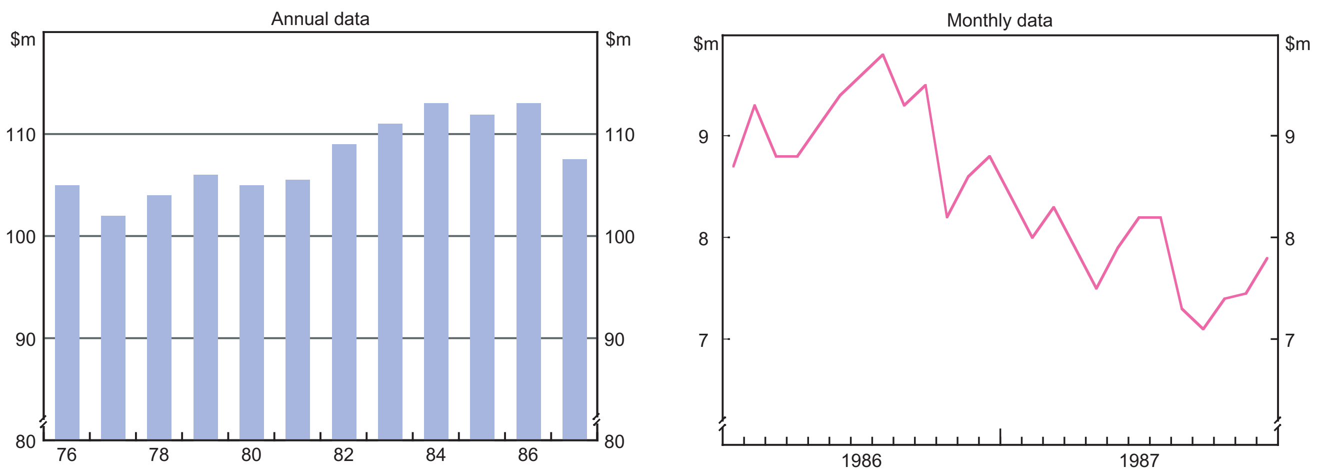
The following are some rules which should be followed to ensure that graphs are helpful to the reader:

(i) Give enough information to make the point.

This can be illustrated with an example. Suppose the question at hand is whether imports have stabilised or begun to rise again after falling for a year or so. One very unimaginative way of trying to illustrate the path of imports would be to plot a long series of annual figures, as shown in the left panel of Graph 2. This is unhelpful for two reasons. First, it gives information about a period which is no longer relevant (up to 12 years ago) and, second, it gives only one observation for the recent year, even though this is the period of most interest. The reader is probably interested in seeing whether there is a change of pattern within the year. A better way of helping the reader would be to present a couple of years of monthly data as in the right panel of Graph 2.

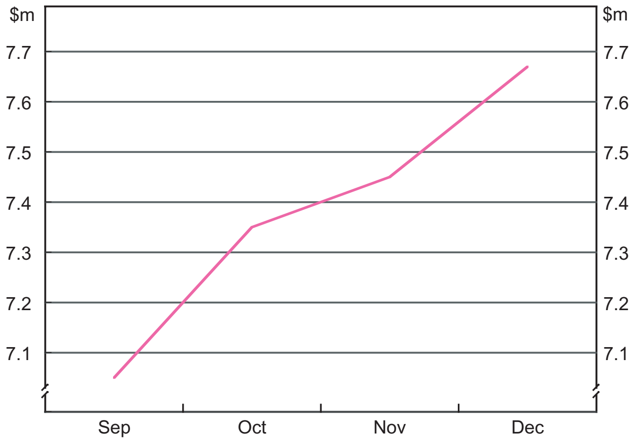
It is essential that there be enough past observations available to put the very recent experience in perspective. Graph 3 shows an example where past data have been omitted. It shows imports over a very short period - four months - and therefore makes the series look like it is rising dramatically. The graph is potentially very misleading. The financial press often omits past data in order to make recent changes look more dramatic.

Graph 2: Imports



4 The best humorous account is given by Huff (1954) and the best authoritative account is in Tufte (1983).

Graph 3: Imports ($ Million)



(ii) Do not exaggerate the vertical scale.

It is difficult to lay down hard and fast rules on the subject of vertical scales. To graph a series properly, the graphist needs to have an understanding of the magnitudes involved. It is only excusable to have a line moving steeply upwards if the variable being graphed is growing at an unusually fast rate. Most of the worst excesses of exaggerating the size of movements are perpetrated by graphists who have no background knowledge of the data they are using - e.g. graphic artists employed by newspapers.

One common reason why vertical scales end up being exaggerated is because too short a time period is plotted on the graph. This was the case for example in Graph 3, which clearly exaggerated the vertical scale. The problem of keeping the scale appropriate is usually helped by plotting a reasonable run of figures on the graph.

(iii) Keep the graph simple.

The main reason for using a graph is to illustrate a point simply. In the vast majority of cases, all that is needed is a series of bars or a line showing the time series observations. If lines are used, it is usually unwise to put more than three on one graph. For example, see the difficulties presented to the reader by Graph 4, which has seven lines (or is it eight?).

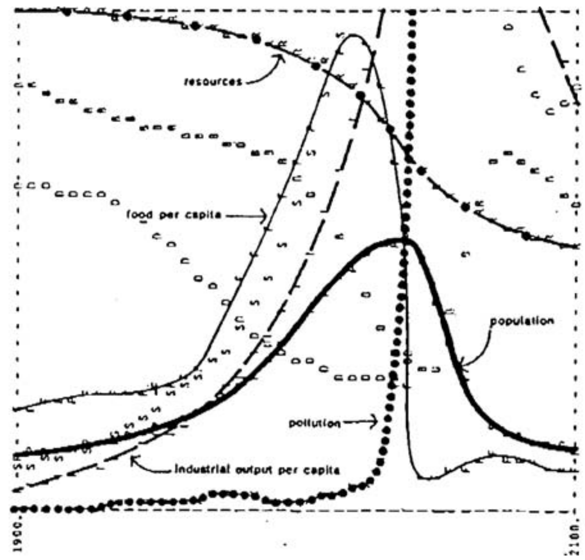
Figurative representation usually obscures the facts rather than helps the reader see the message. It goes without saying that in official memoranda, briefings, etc this sort of "chartjunk" should never be used.[5]

One common vice is to present graphs in a mock three-dimensional form. Graph 5 is an example. It presents only five pieces of information, but does so in a way that makes it appear much more profound; in the process, it makes it very hard for the reader to grasp the information.

(b) A digression - The purist view of the vertical axis

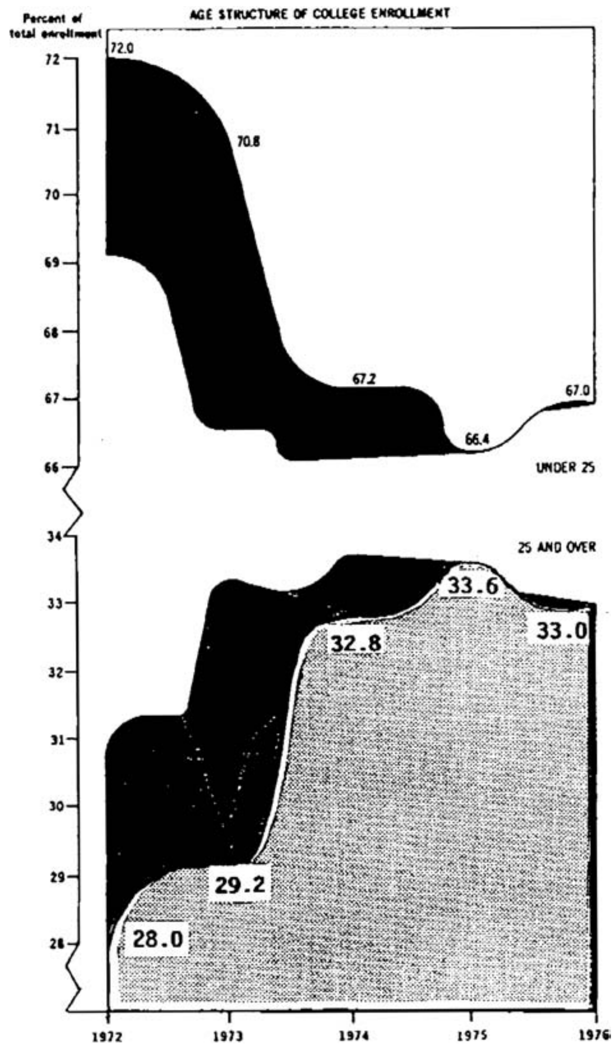
Two issues often exercise the minds of those looking for statistical purity: whether the vertical scale should be broken and whether log scales should be used.

Graph 4: World Model with Unlimited Resources



From Meadows et al (1972)

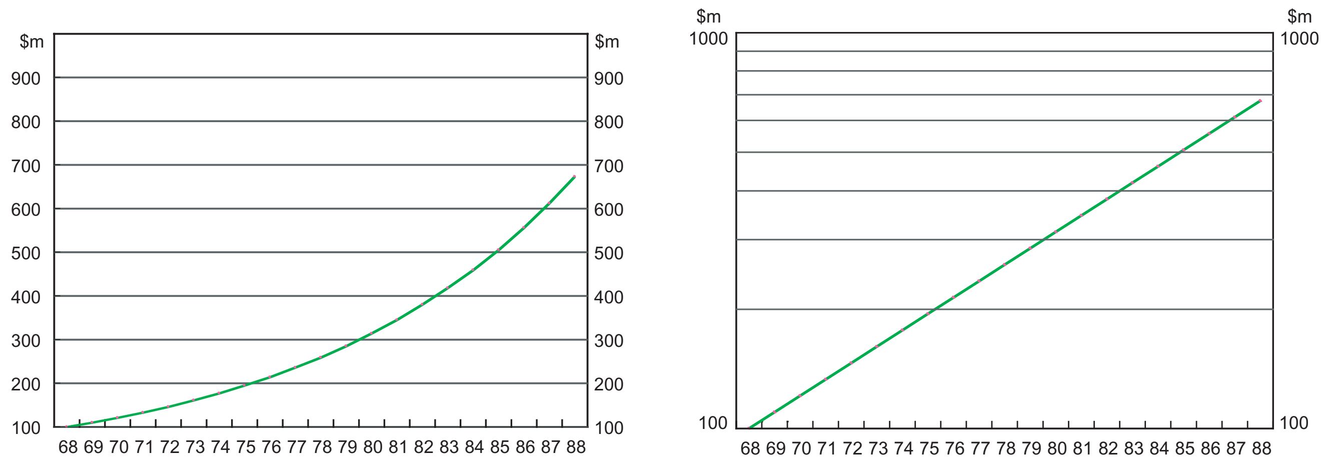
Graph 5



From "American Education". Reprinted in Tufte (1984).

5 Tufte (1984) coined the term "chartjunk" and devoted Chapter 5 of his book to illustrating the problem.

Graph 6



Purists often argue that the vertical scale should always start at zero and be continuous thereafter. This can present problems for variables of large magnitude, but which exhibit relatively small movements. Most people would now accept that it is permissible to break the scale as long as the break is shown and the scale on the vertical axis is clearly enumerated. There are other ways around this problem for the purist who is unwilling to break the scale – for example, the series can be shown in growth rates rather than levels.

The other question is whether to use a linear or a log scale on the vertical axis. It is well known that a variable with a constant rate of growth will rise exponentially. If its level is graphed on a linear scale, it will look as though its growth rate is increasing (left panel of Graph 6). If it is graphed on a log scale, the result will be an upward sloping straight line which clearly conveys the constancy of the growth rate (right panel of Graph 6).

Use of log scales add an extra dimension of complexity to the drawing of a graph. Fortunately, in most cases, it is not necessary to use them because the linear approximation is satisfactory. A log scale is necessary only if the graph covers a long period (as a rule of thumb, more than 10 years) and if the annual growth rate is quite high (again as a rule of thumb, more than 10 per cent). Even when the period to be graphed is long and the growth rate is high, a user who wanted to avoid log scales could do so by plotting growth rates rather than levels for the series.

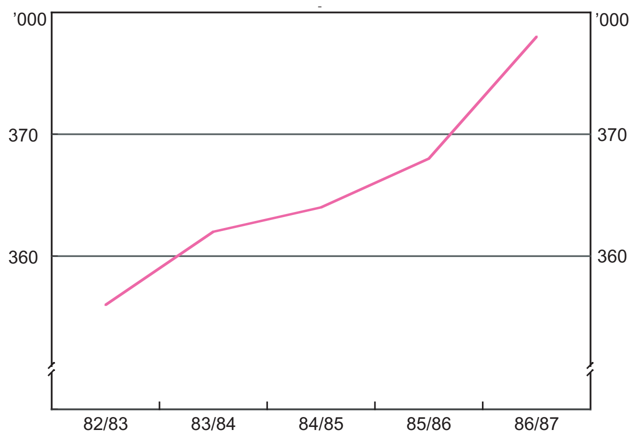
(c) A case study

Graph 7 below is reproduced from an official state government publication in Australia. It shows a steeply rising line representing the level of employment in the state public sector.

As is often the case, simply by looking at the graph, it is impossible to tell whether it is good or bad. It is rarely possible to tell the worth of a graph unless it is seen in relation to the text that it is illustrating. It is often also necessary to know more about the series in question before judgment can be made.

Graph 7 was presented underneath some text that said: “Employment in the public sector has also expanded with the rising share of the public sector in gross State product. The trend is reflected in the following graph.” The graph is clearly a serious misuse of graphical presentations because:

Graph 7



- it is not relevant to the proposition in the text. To be relevant, it would have to have shown public employment as a share of total employment - i.e. it would have to have shown that public employment grew more rapidly than private employment;
- the steepness of the line exaggerates the growth of public employment in this period. The average growth rate was 1-$^{1}/_{2}$ per cent per annum; and
- the growth rate is not only low absolutely, but it is lower than the growth of comparable private sector employment over the same period. In other words, had the authors shown the right variables on the graph, it would have contradicted the assertion made in the text.

Simple econometric relationships

Often there is a need to establish the relationship between variables. If there are only two variables in question, relationships can be discussed simply by drawing a graph or scatter diagram. Often, however, there is a need to go further; a graph will not determine whether the relationship is significant, nor will it quantify the relationship (i.e. give an estimate of the coefficient which links average movements in one variable to average movements in the other variable).

To make these extra steps, it is necessary to use econometric techniques such as correlation or regression analysis. These techniques have the added advantage of being able to establish relationships between more than two variables. They can also take account of lags in the relationships. Users should be aware of these extra features, though they are beyond the scope of the present paper and will not be elaborated here.

Econometric techniques can offer many advantages, but care should be taken in using them, as the following illustrates.

(a) Understanding the raw data

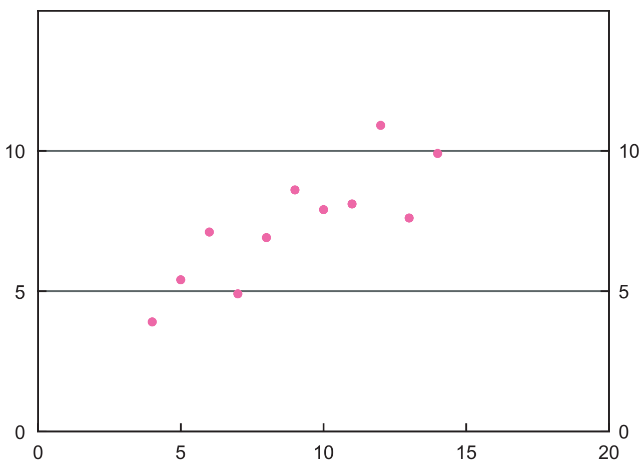
The basic principle with econometric work, as always, is to know your basic data. A fitted equation and accompanying statistical parameters (standard errors, correlation coefficients, etc) can provide a good summary of a relationship between variables, but they do not tell the user all that he needs to know. For example, take the equation below:[6]

Y = 3 + 0.5 X
Standard error = 0.118
r^2 = 0.67
N = 11

The equation is based on 11 observations of two variables, Y and X. It suggests that variable Y is equal to 0.5 X plus 3. The standard error on the coefficient on X is relatively low (0.118), so the user can be reasonably confident of its size and significance. The r^2 of 0.67 suggests that 67 per cent of the variation in variable Y is explained by variable X.

If one saw an equation such as this, it would be normal to expect the 11 observations to be loosely scattered around a generally upward trend such as in Graph 8. This suggests a reasonable relationship between Y and X.

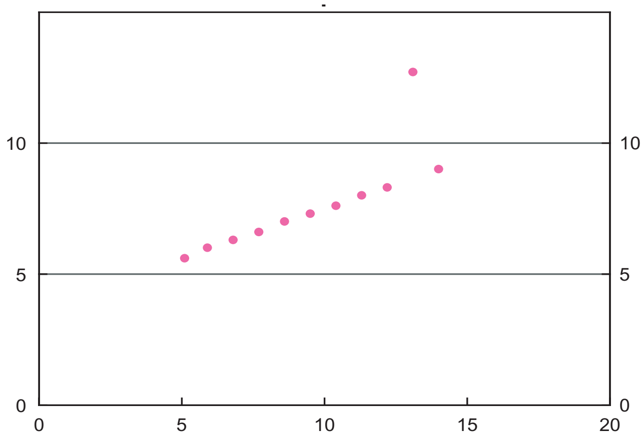
Graph 8



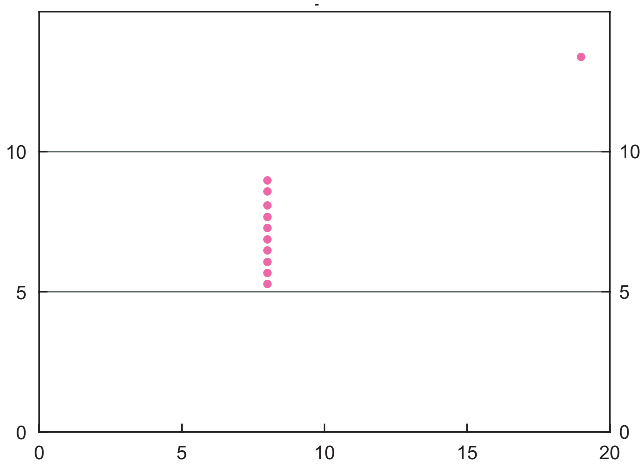
However, the results of the equation above are also exactly consistent with each of the scatter diagrams shown in Graphs 9, 10 and 11. Although the equation is the same, the implications of a movement in variable X, for variable Y are vastly different in each case.

It is extremely important to know what the raw data look like because they may tell the user something about the

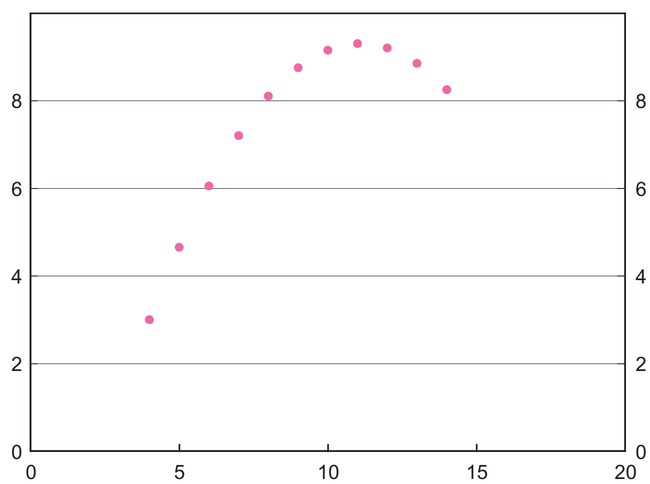
Graph 9



Graph 10

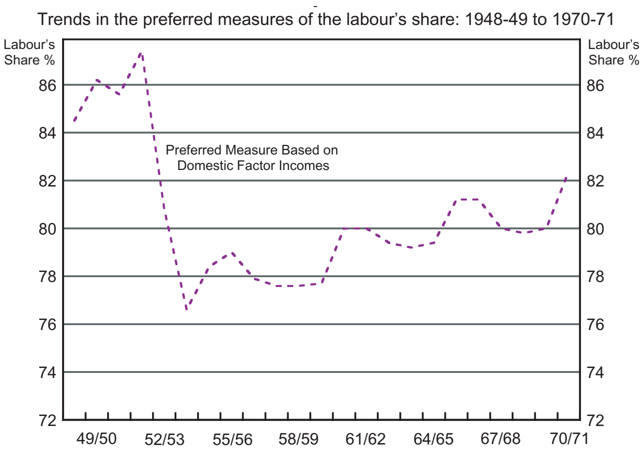


Graph 11



6 This example of equation and graphs is drawn from Anscombe (1973).

Graph 12



relationship that is not revealed by the equation. For the trained econometrician, there are a number of diagnostic statistics that should convey the same information as that obtained from graphs, but for the general user, graphs are best.

Another example is useful. Some years ago, a monograph appeared which contained an equation showing that the labour share of national product in Australia had fallen over the post-war period.[7] This seemed a surprising result to many observers who had noticed that labour's share was reasonably stable, if not rising. The authors of the monograph asserted that the fall in the labour share of 0.14 percentage points per annum was significant at the 95 per cent level, but there were "difficulties involved in assessing the significance of the downward trend because of the presence of autocorrelation". This is not surprising when we see the graph of the raw data.

All of the fall is accounted for by two years (1951/52 and 1952/53). For three-quarters of the period under review, this ratio was on a clear upward trend. This example shows the danger of imposing a trend on data that does not want one. Incidentally, this publication came out in 1975, by which time the ratio had risen further to a position above the top of the graph.

(b) Causation

This is a famous trap for beginners. Because two series are well correlated, it does not mean that one causes the other. Even if they are well correlated and one leads the other - i.e. movements in the dependent variable happen after movements in the independent variable - it still does not mean causation has been established. A good illustration of this is contained in the two letters to *The Times* reproduced below.

Letter of 4 April 1977

From Professor Ivor H. Wills

Sir, Professor Lord Kaldor today (March 31) states that "there is no historical evidence whatever" that the money supply determines the future movement of prices with a time lag of two years. May I refer Professor Kaldor to your article in *The Times* of July 13, 1976. It contains the following figures:

Excess money supply per cent		Increase in prices 2 years later per cent	
1965	4.7	2.5	1967
1966	1.9	4.7	1968
1967	7.8	5.4	1969
1969	4.0	6.4	1970
1969	1.3	9.4	1971
1970	7.8	7.1	1972
1971	11.4	9.2	1973
1972	23.4	16.1	1974
1973	22.2	24.2	1975

If one calculates the correlation between these two sets of figures the coefficient r=0.848 and since there are seven degrees of freedom the P value is less than 0.01. If Mr Rees-Mogg's figures are correct, this would appear to a biologist to be a highly significant correlation, for it means that the probability of the correlation occurring by chance is less than one in a hundred. Most betting men would think that those were impressive odds.

Until Professor Kaldor can show a fallacy in the figures, I think Mr Rees-Mogg has fully established his point.

Yours faithfully,

IVOR H. MILLS,
University of Cambridge Clinical School,
Department of Medicine,
Addenbrooke's Hospital,
Kills Road, Cambridge.
March 31.

7 Department of Labour and Immigration (1975).

Letter of 6 April 1977

From Dr G. E. J. Llewellyn and Mr R. M. Witcomb

Sir, Professor Mills today (April 4) uses correlation analysis in your columns to attempt to resolve the theoretical dispute over the cause(s) of inflation. He cites a correlation coefficient of 0.848 between the rate of inflation and the rate of change of "excess" money supply two years before.

We were rather puzzled by this for we have always believed that it was Scottish Dysentery that kept prices down (with a one-year lag, of course). To reassure ourselves, we calculated the correlation between the following sets of figures:

Cases of Dysentery in Scotland (thousands)*		Increase in Prices one year later (Per Cent)	
1966	4.3	2.5	1967
1967	4.5	4.7	1968
1968	3.7	5.4	1969
1969	5.3	6.4	1970
1970	3.0	9.4	1971
1971	4.1	7.1	1972
1972	3.2	9.2	1973
1973	1.6	16.1	1974
1974	1.5	24.2	1975

* *Annual Abstract of Statistics, 1976 Table 68.*

We have to inform you that the correlation coefficient is -0.868 (which is statistically slightly more significant than that obtained by Professor Mills). Professor Mills says that "Until … a fallacy in the figures [can be shown], I think Mr Rees-Mogg has fully established his point." By the same argument, so have we.

Yours faithfully.

G. E. J. LLEWELLYN, R. M. WITCOMB.
Faculty of Economics and Politics,
University of Cambridge,
Sidgwick Avenue,
Cambridge.
April 4.

The first letter by a Professor of Biology calculates a correlation between some figures for money supply and inflation and concludes that the provider of the figures (Mr Rees-Mogg, the editor of *The Times*) was right to assert that money supply *determines* the future movement of prices. Depending on your view of economics, this may be true or it may not be a reasonable proposition, but it certainly is not proved by running a correlation (or regression), as is wittily pointed out by the authors of the reply.

4. Conclusions

The main point to remember when using statistics is to exercise a fair amount of judgment and common sense. There are, however, some guidelines which, if followed, will help both the user and the reader get more out of the data.

- Know the data:
  - how it is collected
  - how reliable it is
  - what it looks like
- When using tables:
  - keep them as small and simple as possible
  - do not include information unless it is necessary
  - put complicated material in an appendix
- When using graphs
  - give enough information to make the point
  - do not exaggerate the vertical scale
  - keep the graph simple
- When using econometrics:
  - do not let econometric technique be a substitute for knowing the data
  - do not confuse correlation with causation.

Bibliography

Statistical works

F.J. Anscombe, "Graphs in Statistical Analysis", *American Statistician*, February 1973.

Darrel Huff, *How to Lie with Statistics*, Victor Gollancz, 1954.

Oskar Morgenstern, *On the Accuracy of Economic Observations*, Princeton University Press, 1963.

Edward R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Connecticut, 1983.

Howard Wainer, "How to Display Data Badly", *The American Statistician*, May 1984.

Sources for tables and graphs

Department of Labour and Immigration (Australia), *Labour's Share of the National Product*, Australian Government Publishing Service, Canberra, 1975.

Dennis Meadows et al, *The Limits to Growth*, Signet, 1972.

New South Wales Commission of Audit, *Focus on Reform: Report on the States Finances*, Government Printer, N.S.W., 1988.

C. Northcote Parkinson, *The Law*, Schwartz Publishing Group, 1979.

Appendix

Difference between "year-on-year" and "through-the-year" growth rates

Graph 1 below shows how "year-on-year" growth rates are calculated and the shortcomings of this measure. The shaded area shows the level of a variable in each quarter of two adjacent years. The variable in the left-hand panel is referred to as variable A and the one in the right-hand panel is variable B.

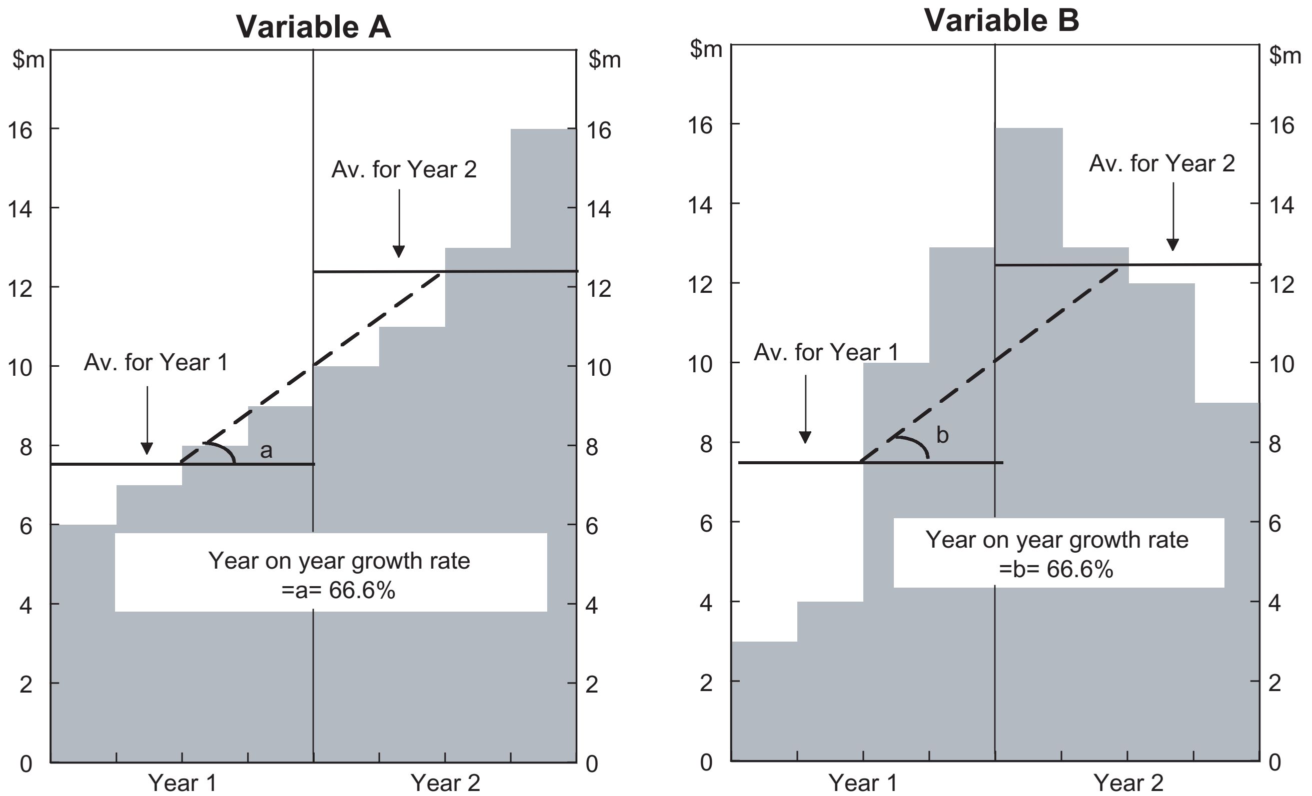
Starting with variable A, it can be seen that the series rises reasonably smoothly in each quarter of the two years. In order to arrive at a "year-on-year" growth rate, it is necessary to calculate the average level in year 1 and year 2 (these are shown by the horizontal black lines). The "year-on-year" growth rate is given by the average in year 2 over the average in year 1 (or by the slope of the dotted line, giving a growth rate of 66.6 per cent).[1] This growth rate gives a reasonable description of how the variable has behaved over the two years.

The graph has been drawn so that the "year-on-year" growth of variable B is exactly the same as that of variable A – i.e. the average level of variable B in each of the years is the same as it is for variable A. But, the path followed by variable B over the two years is very different. After rising through the first year, it falls through the second year. The "year-on-year" growth rate has obscured this fact.

In these circumstances, the alternative measure of calculating the growth rate – the "through-the-year" growth rate – would improve the interpretation. The "through-the-year" growth rate is calculated by dividing the level of the series at the end of the second year by the level of the series at the end of the first year. For variable A it makes a relatively small amount of difference, but for variable B the "through-the-year" growth rate would show a fall (the level of variable B in the fourth quarter of year 1 was 13 but in the fourth quarter of year 2 it was 9).

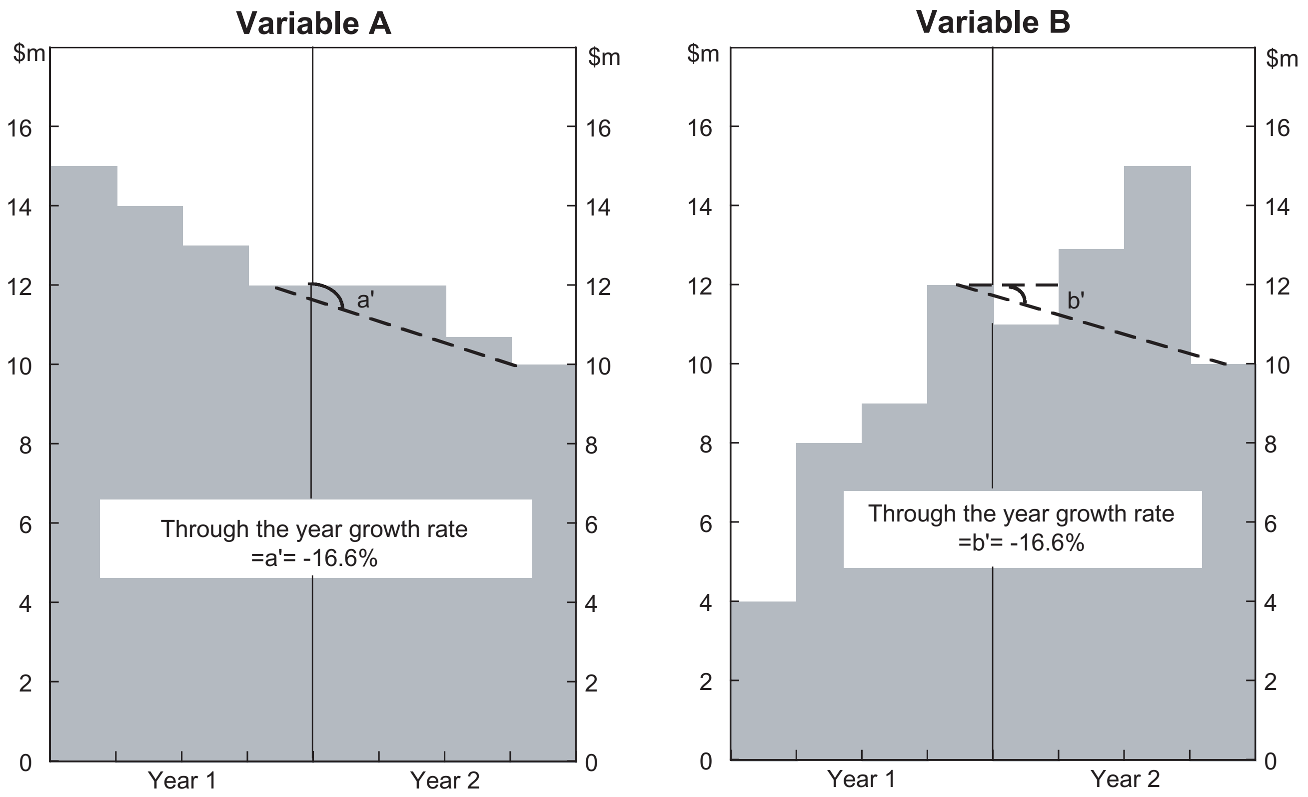
While in this case the "through-the-year" growth rate tends to give a more accurate picture, this is not always so. Graph 2 below shows a situation where it is the "through-the-year" growth rate that is misleading. Again, two variables are shown. Variable A is a series of quarterly observations which are tending to decline smoothly. The "through-the-year" growth rate, calculated by putting the last quarter of year 2 over the last quarter of year 1 is -16.6 per cent Variable B in the right-hand panel also has a "through-the-year" growth rate of -16.6 per cent. However, the general pattern of variable B shows a predominant upward tendency. The average level in year 2 is a lot higher than the average level in year 1 (the year-on-year growth rate is 48 per cent) and each quarter of year 2, except for the fourth quarter, is higher than the corresponding quarter of year 1. In the case of variable B,

Graph 1



1. Strictly speaking, it is the slope over the initial starting value. In this example, the variable A has risen by $5 million over one year and the starting point was $7.5 million.

Graph 2



the "through-the-year" growth rate is misleading because the end point was not representative of the year as a whole. In this type of situation, a "year-on-year" growth rate is better.

Both "year-on-year" and "through-the-year" growth rates show the average effect of number of past quarterly growth rates. The "through-the-year" growth rate is simply the product of the four latest quarterly growth rates. The "year-on-year" growth rate is more complicated as it is the **weighted** average of the quarterly growth rates. The weights attached to each quarter are shown in the next section.

Weighting patterns behind "period-on-period" growth rates

Any average of period growth rate is the weighted average of the growth rates of the sub-periods that make up the average period. For example, a "year-on-year" growth rate is the weighted average of the seven constituent quarters, a "quarter-on-quarter" growth rate is the weighted average of the five constituent months, etc.

The simplest way of demonstrating this is to look at the relationship between a half-yearly growth rate and the four quarterly levels that underly it.[2]

Some simple algebra[3] will show that the "six-month on six-month" growth rates are roughly equivalent to multiplying the quarterly growth rates by a factor of 2 (to convert from a quarterly to a six-monthly scale) and then applying weights of $^1/_4$, $^2/_4$ and $^1/_4$.

The "six-month on six-month" growth rate of a quarterly variable Xt is:

$$\frac{(x_t + x_{t-1}) - (x_{t-2} + x_{t-3})}{(x_{t-2} + x_{t-3})}$$

$$= \left(\frac{1}{x_{t-2} + x_{t-3}}\right)(((x_t - x_{t-1}) + (x_{t-1} - x_{t-2}))$$

$$+ ((x_{t-1} - x_{t-2}) + (x_{t-2} - x_{t-3})))$$

$$= \left(\frac{x_{t-1}}{x_{t-2} + x_{t-3}}\right)\left(\frac{x_t - x_{t-1}}{x_{t-1}}\right) + 2\left(\frac{x_{t-2}}{x_{t-2} + x_{t-3}}\right)\left(\frac{x_{t-1} - x_{t-2}}{x_{t-2}}\right)$$

$$+ \left(\frac{x_{t-3}}{x_{t-2} + x_{t-3}}\right)\left(\frac{x_{t-2} - x_{t-3}}{x_{t-3}}\right)$$

The second half of each of the three expressions above is the quarterly growth rate, and the first half is an expression that approximates $^1/_2$ if xt is not growing very fast, i.e. if the levels of xt-1, xt-2 and xt-3 are similar, i.e. if:

$$\left(\frac{x_{t-1}}{x_{t-2} + x_{t-3}}\right) \approx \left(\frac{x_{t-2}}{x_{t-2} + x_{t-3}}\right) \approx \left(\frac{x_{t-3}}{x_{t-2} + x_{t-3}}\right) \approx \frac{1}{2}$$

2. This has been chosen because there are fewer numbers to juggle than with quarterly to monthly, where there are six quarterly levels, or annual to quarterly, where there are eight quarterly levels.
3. This formal demonstration was provided by Rob Trevor of the Research Department of the Reserve Bank of Australia.

If so, then the "half-year on half-year" growth rate is approximated by:

$$2\left(\frac{1}{4}\left(\frac{x_t - x_{t-1}}{x_{t-1}}\right)+\frac{2}{4}\left(\frac{x_{t-1}-x_{t-2}}{x_{t-2}}\right)+\frac{1}{4}\left(\frac{x_{t-2}-x_{t-3}}{x_{t-3}}\right)\right)$$

Similar manipulations may be used to show that a "quarter-on-quarter" growth rate is equivalent to multiplying the monthly rates by a factor of 3 (to convert them from a monthly to a quarterly scale) and then smoothing them with a five-period moving average with weights $^{1}/_{9}$, $^{2}/_{9}$, $^{3}/_{9}$, $^{2}/_{9}$, $^{1}/_{9}$. The commonly used "year-on-year" growth rate is then equivalent to multiplying the "quarter-on-quarter" growth rates by a factor of 4 (to get the scale right) and then smoothing them with a seven-period moving average with weights $^{1}/_{16}$, $^{2}/_{16}$, $^{3}/_{16}$, $^{4}/_{16}$, $^{3}/_{16}$, $^{2}/_{16}$, $^{1}/_{16}$. The three main weighting schemes are summarised in Table 1.

Table 1: Weighting Patterns

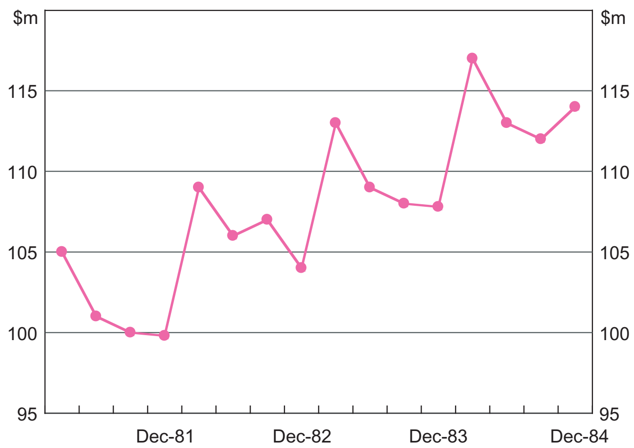
Quarterly from monthly growth rates									
Months:		1	2	3	4	5	6		
	3	(-	1/9	2/9	3/9	2/9	1/9)		
Half-yearly from quarterly growth rates									
Quarters:		1	2	3	4				
	2	(-	1/4	1/2	1/4)				
Annual from quarterly growth rates									
Quarters:		1	2	3	4	5	6	7	8
	4	(-	1/16	2/16	3/16	4/16	3/16	2/16	1/16)

Crude seasonal adjustment methods

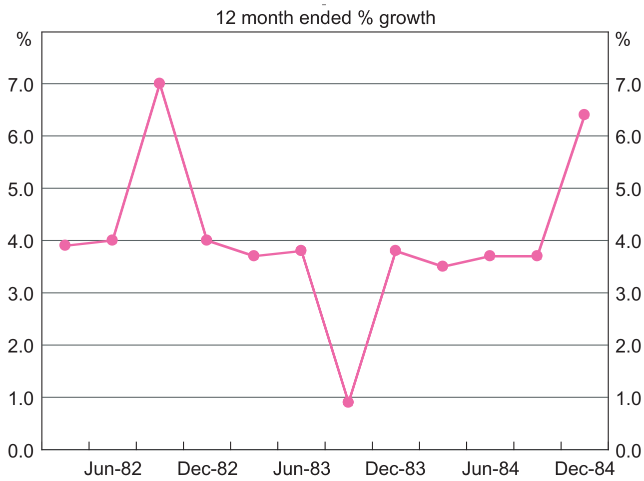
One method of adjusting for seasonality is to calculate "twelve-months-ended" growth rates, but errors of interpretation can arise. This is illustrated with the data in Graph 3, which shows the quarterly levels of a series over four years. There is a clear seasonal pattern with, for example, each March quarter considerably above the trend.

Graph 4 shows "twelve-months-ended" growth rates. The trend growth is around 4 per cent, with three observations being "off-trend". This "twelve-months-ended" growth rate will often correctly identify an observation which is "off" the seasonally-corrected trend. For example, the 7 per cent growth in the year to September 1982 can be seen as a 4 per cent trend plus 3 per cent above-trend growth in that quarter. However, the observation one year later – September 1983 – is, in fact, on-trend but the twelve-months-ended rate **falls** sharply. This fall does not reflect anything that happened in the September 1983 quarter, but simply reflects the higher September 1982 base for the calculation.

Graph 3



Graph 4



One simple method of presentation which overcomes this problem is shown in Graph 5. The data are graphed so that each quarter can be compared with the corresponding quarter of other years. The seasonal pattern is apparent. The above-trend observations in September 1982 and December 1984 are clearly shown, without the false reading for September 1983 given by the twelve-months-ended growth rate.

Graph 5