Presentation of Data : Graphs

*****Jeffery, C.J., Adam, P., 2000. Presentation of Data: Graphs.
Faculty of Life Sciences, UNSW, Sydney 2052
*Tel.: + 61 2 9385-1825
*Fax: + 61 2 9385-1086
*E-mail: cas**jeffery@hotmail.com**

Faculty of Life Sciences, University of New South Wales, Sydney, NSW 2052, Australia

1. Introduction

While this resource material on Graphs has a biological emphasis, it was prepared in response to the introduction of lists of generic skills, common to ***all*** Stage 6 syllabi in Science, for the New South Wales Preliminary Courses for 2000. It primarily aims to address the following generic skills as listed and defined in Board of Studies (1999):

P13 A student identifies appropriate terminology and reporting styles to communicate information and understanding in biology

13.1 Students present information by:

b) selecting and using appropriate media to present data and information
d) using symbols and formulae to express relationships and using appropriate units for physical quantities
f) selecting and drawing appropriate graphs to convey information and relationships clearly and accurately
g) identifying situations where use of a curve of best fit is appropriate to present graphical information

P14 A student draws valid conclusions from gathered data and information

14.1 Students analyse information to:

a) identify trends, patterns and relationships as well as contradictions in data and information
b) justify inferences and conclusions
c) identify and explain how data supports or refutes an hypothesis, a prediction or a proposed solution to a problem
d) predict outcomes and generate plausible explanations related to the observations
e) make and justify generalisations
f) use models, including mathematical ones, to explain phenomena and/or make predictions
g) use cause and effect relationships to explain phenomena
h) identify examples of the interconnectedness of ideas or scientific principles

Although graphing skills are required in all scientific disciplines our examples are related to Biology. In these examples we use a number of technical terms, and introduce the scientific (Latin) names for a number of species. The curriculum requires experience in scientific communication and part of scientific communication involves the proper use of scientific terminology. This is to increase the precision and reduce possible ambiguity, not to baffle with jargon. For example, a Latin binomial name for a species can only apply to one organism, whereas common, or vernacular, names may apply to several species, or usage may differ in different regions ("blue gum" in New South Wales - *Eucalyptus saligna* is different from "blue gum" in Tasmania – *Eucalyptus globulus*). It is important to get into the habit of using proper terminology – particularly in exams so that markers clearly understand what you have written!

2. Advantages of Graphs

- Data are instantly conveyed
- Data are presented clearly and simply
- Graphs can expose relationships and patterns
- Graphs can be used to emphasise information and relationships
- Graphs are valuable in presentations for emphasising messages
- Different graph styles can be used to highlight different aspects of data

If you want to determine particular relationships (eg. you may predict (hypothesise) that more herbivorous snails will be found on rocky shores with a greater percent cover of green algal turf than in areas with reduced cover) you will need to display data so that they are easy to understand. Hence, an understanding of **terms** and the **type of graph** to use will be crucial in making your data easy to interpret.

3. Terms

3.1 What is a variable?

Variables can be **qualitative** (eg. eye colour) or **quantitative** where observations can be measured or counted. Quantitative **discrete** data are obtained by counting (eg. numbers of crabs in a rockpool) whereas quantitative **continuous** data are measurements, usually within a range (eg. size of crabs). Furthermore, individuals being measured or counted can be sorted into separate groups – **categories** (eg. worms, crabs, snails, anemones, algae) or individuals can be divided into groups based on size – **size intervals** (eg. 0-1.9 cm, 2-3.9 cm etc) (Table 1).

Data can be collected to determine relationships between variables. For example, a fish farmer may want to know the relationship between body weight of fish and amount of food supplied in order to manage his costs; a pearl farmer may want to know how pollution will affect his pearl yield; and a marine scientist may want to know the relationship between survivorship of exotic marine larvae in ballast water after exposure to varying temperatures to provide information for measures to prevent invasion of foreign marine species (eg. in Tasmania the yellow seastar *Asterias amurensis* was introduced in ballast water of woodchip ships from Japan). To make sense of data, measurements should first be tabulated, then graphed appropriately.

Table 1 Types of Data

Light Intensity (Lux)	Length of Stem (cm)
0	12
2000	34
4000	40
6000	30
8000	20
Continuous	**Continuous**

Species	Number of Animals
Barnacle	520
Crab	2
Blue Snail	200
Whelk	10
Zebra Snail	50
Category	**Discrete**

Size of Barnacles (mm)	Number of Barnacles
0 - 1.9	20
2 - 3.9	50
4 - 5.9	6
6 – 7.9	4
8 – 9.9	2
Size Interval	**Discrete**

3.2 Statistical Tests

While ***secondary students studying Science in New South Wales schools are only required to present data in graphical form***, research scientists also use statistical tests to emphasise differences in results from different experimental treatments, particularly if graphs do not readily expose relationships. For example, while differences in treatments (eg. numbers of beetles surviving different dosages of pesticides) may not be obvious in bar graphs when you have not spread the *y*-axis scale line sufficiently, statistical tests will show this difference. Statistical procedures to manipulate data are available in many statistical packages such as SysStats and SPSS which can easily be purchased. Tests to use may include *t* tests, analysis of variance (ANOVA) and multiple comparisons of means such as Student Newman Keuls (SNK) analyses. For example, each of these procedures may highlight differences in responses of intertidal snails to different temperatures. Alternatively, when these snails are counted in several areas at the one rocky shore, statistical tests may emphasise that numbers of these snails will vary spatially.

It is important to remember that the computer program does not know when the question you are addressing, or the form of data you have collected, are inappropriate for a particular statistical procedure. A statistical package in inexperienced hands can therefore be a very dangerous weapon!!!!

3.3 Independent and Dependent Variables

When preparing graphs it is important that the **independent variable** is on the *x*-axis or **abscissa**, and that the **dependent variable** is on the *y*-axis or **ordinate** (Fig. 1). **Independent variables** may be factors which have been experimentally manipulated, or may be factors over which we cannot have an influence (eg. Time). The **dependent variable** is the outcome or response to the varying **independent variable**. The **independent variable** is therefore the variable that is hypothesised to have an effect on some other variable. That is, the **independent variables** on the *x*-axis predict or explain the **dependent variables** on the *y*-axis. For example, if tree height is measured over time

(Fig. 1), height will be shown to increase with time (probably in response to increased exposure over time to sunlight, carbon dioxide and water although none of these variables will have been measured). Not all measures of time, however, are independent variables. For example, when time to onset of germination (Days) is measured in plants in response to varying temperatures, Time should be plotted on the *y*-axis and will be **dependent**, whereas Temperature should be plotted on the *x*-axis and will be **independent**.

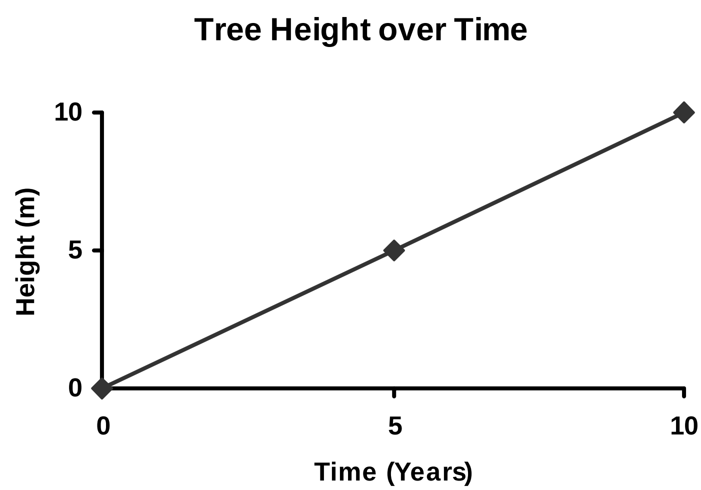


***Fig.1**
Height of eucalypts measured over ten years.

**If one were to calculate statistics it would be necessary to measure different trees at each time, not repeated measures of the same trees over ten years.*

4. Tabulating Data (see Tables 1, 2, 4, 5)

- Include a Title and Table number (eg. Table 2. Numbers of barnacle larvae caught in traps; *cyprid = final larval stage before settlement)*
- The independent (*x*) variable should always be tabulated in the far left column
- Size intervals or categories should also be tabulated in the far left column
- Each variable should be labelled at the head of each column of data
- Rows should be clearly labelled where necessary (eg. species name, size intervals)
- All data should be entered neatly and evenly in columns under appropriate headings for variables
- Each variable should be listed with its scientific unit of measurement
- Missing values should be recorded as a dash (**–**)
- Zero values should always be recorded
- The number of significant figures should be appropriate for the method of measurement (eg. 40.1, 3.289, and 28.42 should be recorded as 40.10, 3.29 and 28.42) (*Note that, in some cases, three decimal places may be appropriate!!*)

Table 2
Mean annual total numbers of cyprid larvae of the honeycomb barnacle *Chamaesipho tasmanica* captured in one set of traps (*n* = 3) at one place low on the shore at Cape Banks, Botany Bay, New South Wales from July to December, 1990 to 1993 (Jeffery, 1997; Jeffery and Underwood, 2000). (SE, the Standard Error of the Mean is a statistical measure of the variability in the data – see **5. Graphical Presentation of Data** and **7.4 Standard Errors**)

Time	Mean Number of Cyprids	S.E.
1990	173.00	26.00
1991	79.33	10.99
1992	74.00	5.13
1993	20.67	14.43

5. Graphical Presentation of Data

Graphs, rather than tables, are simple and powerful vehicles for conveying information and emphasising patterns and relationships from data. To enhance presentation of these data, however, certain guidelines must be followed (see Fig. 2).

- Include a Title and Figure number
- A figure caption must be included to describe what dependent variable/s are plotted on the *y*-axis against the independent variable on the *x*-axis (Note that often more than one dependent variable can be plotted)
- The variables on the *x* and *y* axes must be labelled (eg. Shell length)
- Units of measurement for each variable should be included when appropriate (eg. cm)
- Scales must always be marked on the *x* and *y* axes to illustrate relationships (Note, however, that scales may not need to be included on the *x*-axis for bar graphs)
- Markers (eg. ■ ,◆) should be used to indicate data points
- If several markers are used, a key or legend should be included to explain each one
- *Standard error (SE) bars may be used to illustrate how much each set of measurements (replicates) will vary from the mean (see **7.4 Standard Errors**)
- **When lines of best fit are used, the appropriate equation may be included on the graph or on a separate table (see **7.1 Line of Best Fit**)

Note that secondary students studying Science in New South Wales are not required to include Standard Error (SE) bars (see **7.4 Standard Errors).*
*** Note that secondary students studying Science in New South Wales are not required to calculate lines of best fit (see **7.1 Line of Best Fit**) – you should merely include a line of best fit through data points as estimated visually.*

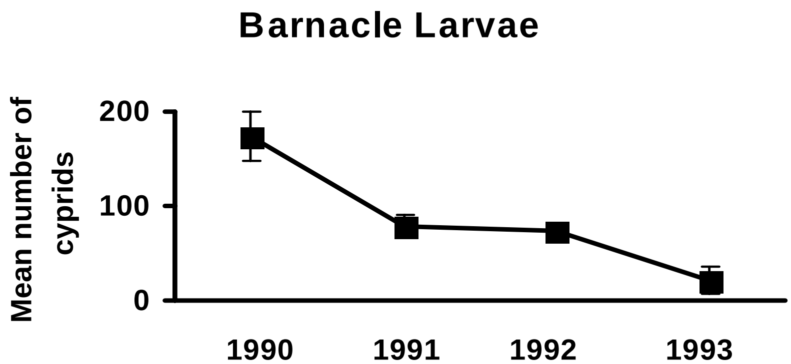


***Fig. 2**
Mean annual total numbers of cyprid larvae of the honeycomb barnacle *Chamaesipho tasmanica* captured in one set of traps (number of samples (*n*) = 3) at one place low on the shore at Cape Banks, Botany Bay, New South Wales from July to December, 1990 to 1993 (Jeffery, 1997; Jeffery and Underwood, 2000).

**This graph might be read to imply that very few larvae of barnacles will arrive in 1994 and that this population of the species is on the way to extinction. Many cyprids could actually arrive to settle and metamorphose into adult barnacles, but, without repeated recruitment from a planktonic larval stage, populations of barnacles would become extinct.*

6. Types of Graphs

There are many ways in which data can be presented graphically so as to aid interpretation. The type of data collected may depend on the question being addressed. For example, if numbers of particular species of snails, barnacles and crabs are counted randomly in 50 cm by 50 cm **quadrats** (= **replicates** = **samples**) on rock platforms (Table 1) obviously there are no dependent or independent variables and data should be represented as **bar graphs** (Fig. 3). Alternatively, if numbers of animals are recorded randomly at different general (unmeasured) levels of the shore (Low, Mid and Upper) or at actual (measured) distances on the rock platform from the sea at low tide (Mean Low Water), the numbers of individuals (dependent variable, *y*-axis) may be related to these general or actual distances on the shore from Mean Low Water (independent variable, *x*-axis). Therefore, if numbers per species are recorded at generally Low, Mid or Upper levels, data may be represented by **bar graphs** (Fig. 4); if numbers are recorded at actual distances data may be expressed as **line graphs** (Fig. 5); or if there are a lot of data points and it is necessary to determine a relationship between the dependent and independent variables **scatter graphs** (Fig. 6) would be appropriate. Counts should confirm initial observations and it would be expected that more blue snails would be recorded, and expressed graphically, with increased distance from Mean Low Water (Figs. 4, 5, 6).

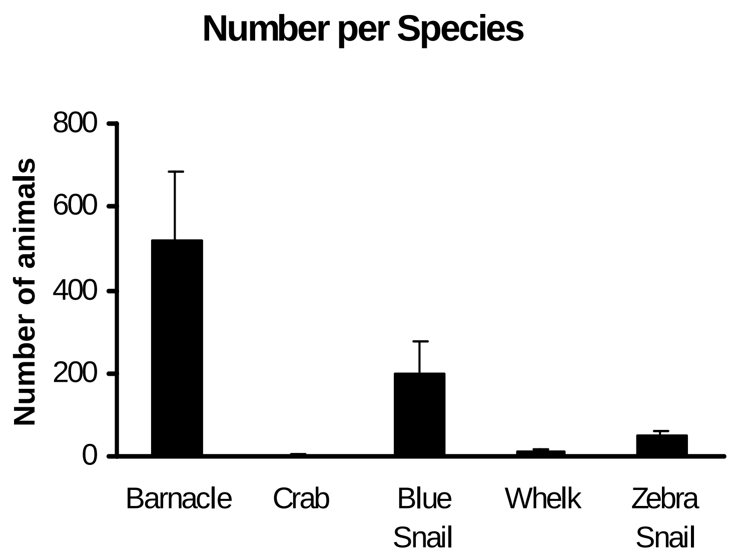


Fig. 3
Mean number of each species recorded on the rock platform in 50 cm x 50 cm quadrats (number of samples (*n*) = 3).

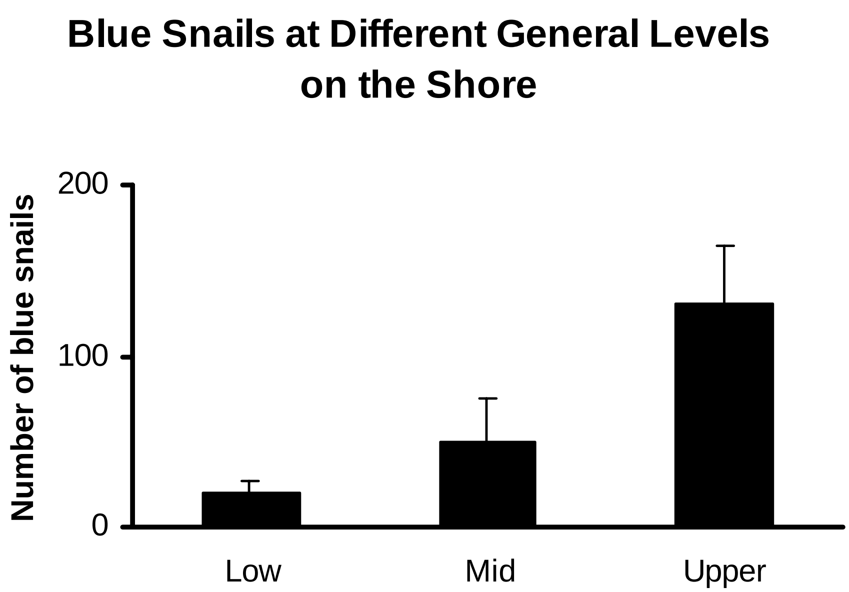


Fig. 4
Mean number of blue snails in 50 cm x 50 cm quadrats (number of samples (*n*) = 3) recorded at Low, Mid and Upper levels on the shore.

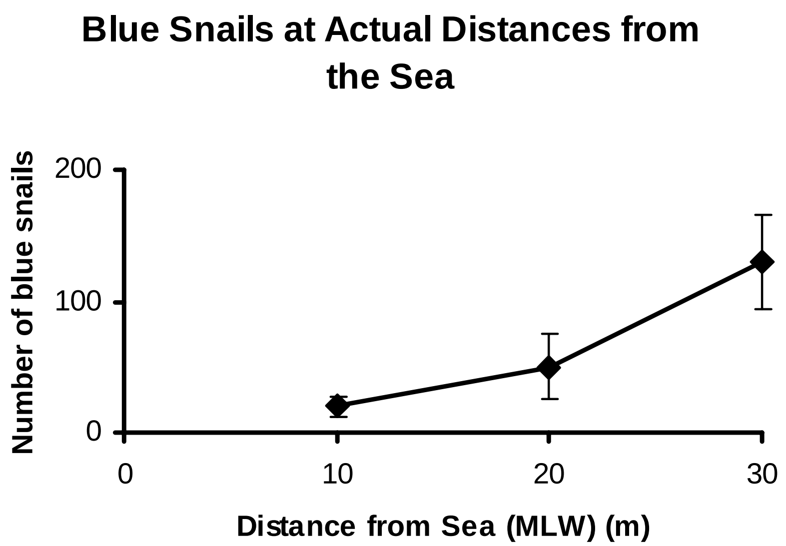


Fig. 5
Mean number of blue snails (number of samples (*n*) = 3) recorded in 50 cm x 50 cm quadrats at actual distances (10, 20, 30 m) from Mean Low Water (MLW).

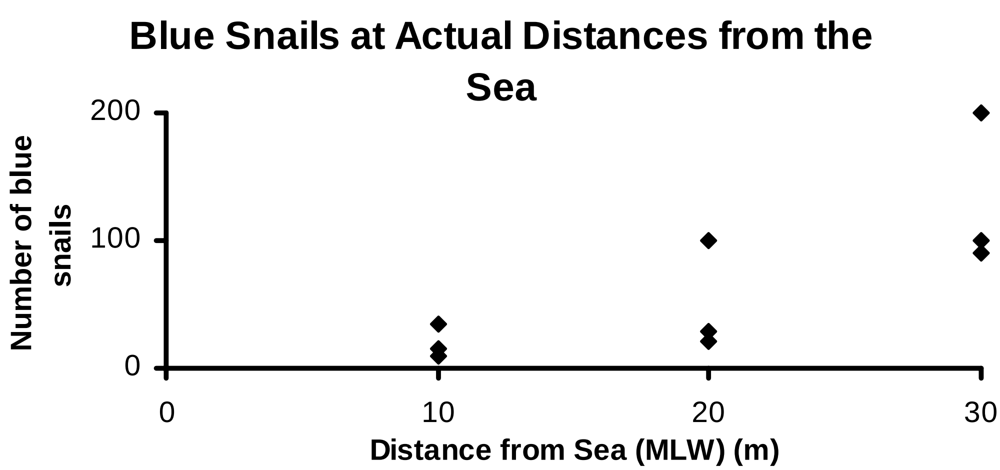


***Fig. 6**
Number of blue snails (number of samples (*n*) = 3) recorded in 50 cm x 50 cm quadrats at 10 m, 20 m and 30 m from Mean Low Water (MLW).

No Standard Error (SE) bars (see **7.4 Standard Errors) have been included because all data points, rather than means, have been entered.*

Furthermore, if the number of individuals of each species are recorded as a % of the total number of individuals counted on the rock platform, there would be a greater % of barnacles than any other species (Table 1) and this could be expressed as a **pie chart** (Fig. 7). Also, if barnacles are measured and classified into particular size classes or size intervals at any one time (Table 1), **size frequency histograms** (Fig. 8) will indicate at a glance the numbers of barnacles within a size range and the numbers of populations within a sample. If barnacles in a population are measured randomly over time, however, **line graphs** (Fig. 9) or **scatter graphs** (Fig. 10) may be the appropriate graphical representation. Hence, in order to see any trends or patterns in data it is important to decide which types of graphs are appropriate for the type of data.

Graphs which can be used include:

- Bar graphs (Figs. 3, 4, 19, 20)
- Line graphs (Figs. 1, 2, 5, 9, 12, 17, 18)
- Scatter graphs (Figs. 6, 10, 11, 13, 14, 16)
- Pie charts (Fig. 7)
- Frequency histograms (Fig. 8)

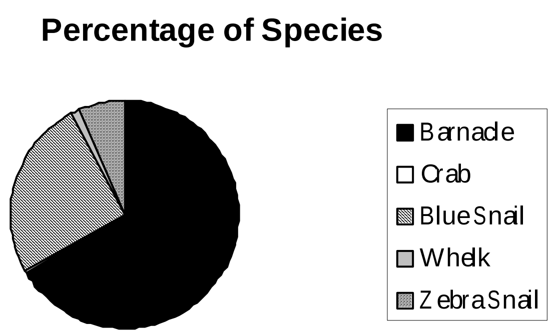


***Fig. 7**
Percentage of each species recorded on rock platform.

No Standard Error (SE) bars (see **7.4 Standard Errors) are included in pie charts. Note that one of the problems of % data is that changes in total number may be obscured because % must add up to 100 regardless of sample size.*

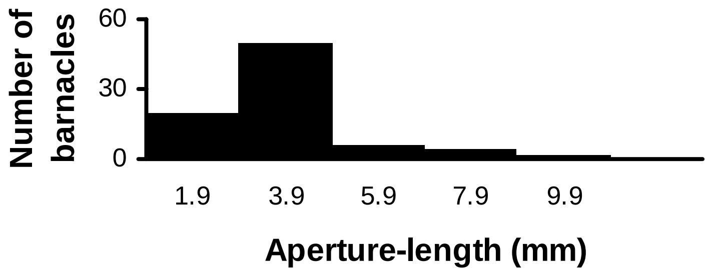


***Fig. 8**
Size frequency histogram of barnacles in 6 cm diameter quadrats on a rocky shore.

No Standard Error (SE) bars (see **7.4 Standard Errors) are included because all data are entered as separate data points (no means).*

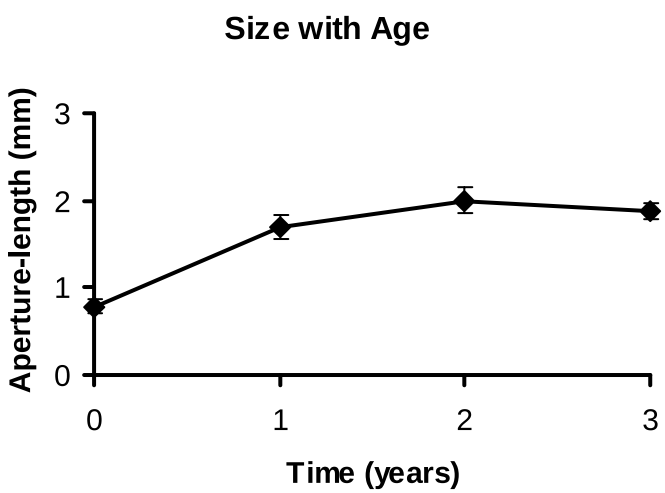


Fig. 9
Mean size of one cohort of juveniles (<1 mm at first measurement) of the honeycomb barnacle *Chamaesipho tasmanica* (number of samples (*n*) = 18) measured at the end of each year from 1989 to 1992 at each of Low, Mid and Upper levels of Site 4 at Cape Banks, Botany Bay, New South Wales (Jeffery, 1997; Jeffery and Underwood, 2001).

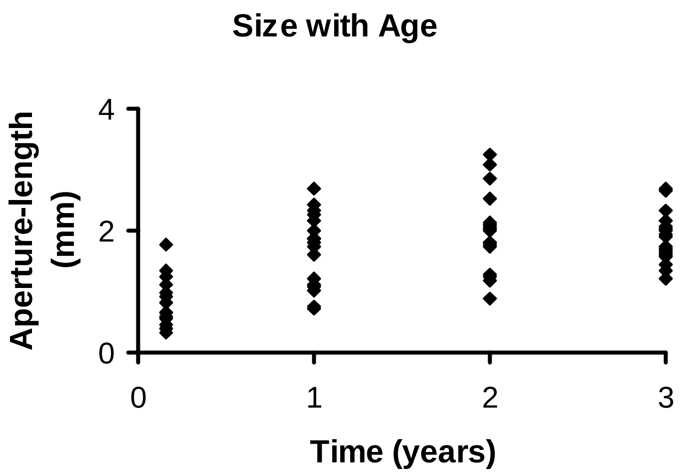


Fig. 10
Size of one cohort of juveniles (<1 mm at first measurement) of the honeycomb barnacle *Chamaesipho tasmanica* (number of samples (*n*) = 54) measured at the end of each year from 1989 to 1992 at Site 4 at Cape Banks, Botany Bay, New South Wales (Jeffery, 1997; Jeffery and Underwood, 2001).

When choosing types of graphs to express data, the following notes and checklists may help you. (Also see 7.4 Standard Errors and 8 Common Faults with Graphs):

6.1 Bar Graphs (Figs. 3, 4, 19, 20)

Bar graphs are simple expressions of data where the height of each bar represents the number of observations in a category. These observations/variables can be discrete (counts) or continuous (measurements). Bar graphs are useful for expressing non-related values of measurements for different categories. For example, they may be used to represent the numbers of snails, crabs or barnacles on a rocky shore (Fig. 3) or the metabolic rates of these different species at one temperature.

Characteristics of Bar Graphs:

- Figure numbered and titled
- Bars of equal width
- Bars not touching unless multiple sets of data included (eg. Treated versus Untreated data; numbers of fish in polluted creeks versus numbers in unpolluted creeks)
- Different shading or texture to represent different data sets
- Key or legend to indicate different data sets
- Non-numerical variable on horizontal *x*-axis
- Labels and units included on *x* and *y* axes
- Even scales on axes
- Standard Error (SE) bars included (see **7.4 Standard Errors**)

6.2 Line Graphs (Figs. 1, 2, 5, 9, 12, 17, 18)
Line graphs visually present a series of data by connecting individual data points. They are particularly useful for representing trends over time (Figs. 1, 2, 9, 12, 17, 18). Alternatively, where appropriate, data can be represented by a straight line (see **7.1 Line of Best Fit**) where data points are not joined, and statistical tests are used to measure how a set of data points fits a straight line.

Characteristics of Line Graphs:

- Figure numbered and titled
- Independent variable on horizontal *x*-axis
- Key or legend to indicate different data sets (usually different marker shape)
- Labels and units included on *x* and *y* axes
- Even scales on axes
- Scales increase in right direction
- Range of scales fit data
- Line of best fit drawn
- Standard Error (SE) bars included (see **7.4 Standard Errors**)

6.3 Scatter Graphs (Figs. 6, 11, 13, 14, 16)
Scatterplots can be used to express the relationship between continuous variables. If we have data on two variables (eg. number of adult female dugongs (seacows) and number of young calves) we may need to determine whether there is a relationship between them. If the two variables were related there should be some pattern that connects them and this could be expressed in a linear pattern or a straight line. (***Note that many relationships are curvilinear but they are harder to interpret***). **Scatter graphs** will reveal that a linear relationship exists between two variables. We would therefore expect that, if our data were graphed, an obvious relationship between the number of young (dependent *y*-axis variable) and the number of adult female seacows (independent *x*-axis variable) would be evident. A line can be drawn to represent this relationship. If the line slopes upwards, an increase in numbers of adult females would be positively associated with an increase in the numbers of juveniles. If the line slopes downwards, a decrease in numbers of adult females would be negatively associated with a decrease in the numbers of juveniles. While this graph and trendline may visually express the relationship between numbers of adults and young dugongs, the **actual** relationship can be determined mathematically by statistical analyses such as **correlation analysis** or **regression analysis** (see **7.1 Line of Best Fit**). (Note that a third variable could be incorporated in the scatter graph, but different plotting symbols should be used to differentiate between them. For example, we might want to determine whether the numbers of male dugongs are also associated with numbers of females).

Characteristics of Scatter Graphs:

- Figure numbered and titled
- Independent variable on horizontal *x*-axis
- Key or legend to indicate different data sets (usually different marker shape)
- Labels and units included on *x* and *y* axes

- Even scales on axes
- Scales increase in right direction
- Range of scales fit data
- Line of best fit drawn

Note that Standard Error (SE) bars (see ***7.4 Standard Errors****) are not included in scatter graphs: data are usually entered as separate entities although data can also be entered as means. When data are entered as means, this information is included in the figure caption.*

6.4 Pie Charts (Fig. 7)

Pie charts represent proportions by displaying data as pie slices. That is, each set of data is expressed as a percentage of a whole. For example, after we have counted numbers of snails, barnacles and crabs inhabiting a rock platform, we may want to show the relative frequency of each species (Fig. 7).

Note that Standard Error (SE) bars (see ***7.4 Standard Errors****) cannot be displayed in pie charts.*

6.5 Frequency Histograms (Fig. 8)

In **frequency histograms** the height of each bar represents the number of individuals in each size or class interval (see **7.2 Measurement**). Histograms are drawn without gaps between the bars because the *x*-axis is used to illustrate the size or class intervals. For example, we may need to visually represent the size frequency distribution of a population of barnacles (Fig. 8; Table 1) or we may need to compare the range of blood pressure of elderly people at a retirement home with that of a population of young athletes (obviously two histograms would be necessary when comparing the two age groups).

Characteristics of Frequency Histograms:

- Histogram has title and figure is numbered
- Frequency entered on vertical *y*-axis
- Size intervals entered on horizontal *x*-axis
- Class (size) intervals don't overlap
- Size intervals increase evenly
- Labels and units included on *x* and *y* axes
- No gaps between bars
- Bars shaded

Note that Standard Error (SE) bars (see ***7.4 Standard Errors****) cannot be included because data are entered as separate entities rather than as means.*

7 Presenting Data
7.1 Line of Best Fit
7.1.1 What is Line of Best Fit?
The purpose of fitting a line to a set of data points is:
- To seek a generalised relationship amongst variable data
- To describe hypotheses, that is, to make predictions from graphs

Therefore, **lines of best fit** are lines that best represent the data points around this line. For example, if the height and girth of several trees were measured (Fig. 11), it would be meaningless to join all data points because the curved line would give the impression that height was fluctuating whereas measurements were taken from many trees. Hence, line of best fit indicates that there is a trend for tree height to increase with increased tree girth (Fig. 11).

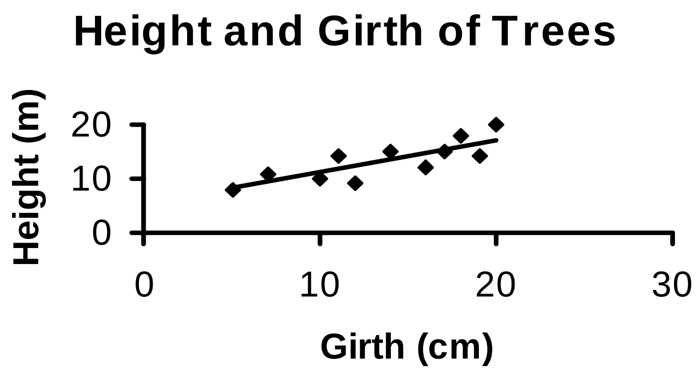


Fig. 11
Relationship between height and girth of trees.

In certain circumstances it is not appropriate to use line of best fit and data must be represented as joined points. For example, because arrival of larvae of the barnacle *Chamaesipho* is episodic and pulses of cyprid larvae arrive only once or twice a year in spring (Jeffery, 1997; Jeffery and Underwood, 2000), if a straight line were to be drawn between all points in Figure 2, interpolation (see **7.3 Interpolation and Extrapolation**) from this line between data points would give an inaccurate representation of the number of cyprids arriving at that time of the year.

Similarly, it would be legitimate to connect points when measuring atmospheric carbon dioxide over time (Fig. 12). While a line fitted to this set of data points would best describe the trend of increase in carbon dioxide, joining the points would make seasonal differences between summer and winter evident (Fig. 12). This seasonal trend would not be evident if line of best fit were used. Obviously you will need to use common sense to decide whether linking points is appropriate, but in general, the line of best fit is the best choice.

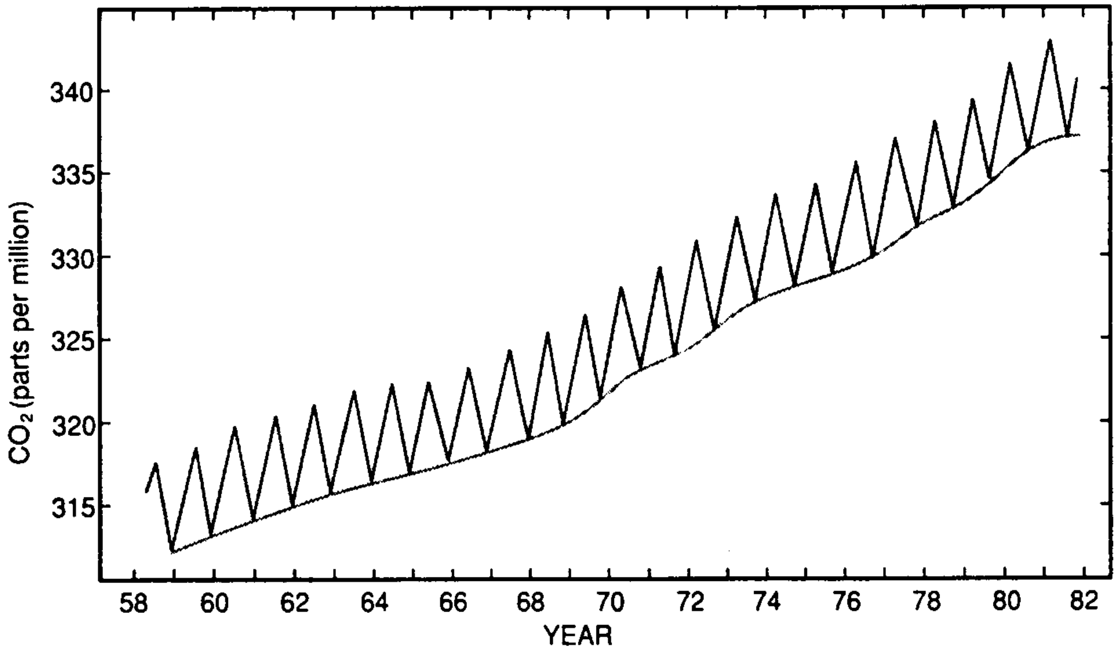


Fig. 12
Monthly mean values of atmospheric carbon dioxide at Mauna Loa, Hawaii 1958-1981. The annual "sawtooth" indicates the winter decline in deciduous and evergreen photosynthesis in the northern hemisphere (Walker, 1992).

7.1.2 Creating Line of Best Fit

Secondary students studying Science in New South Wales are ***not*** required to ***calculate*** lines of best fit but they should draw ***by eye*** each line of best fit through data points for each graph. It may help. However, ***to understand*** how lines of best fit can be calculated. For example, statistical analyses such as **regression analysis** and **correlation analysis** can be used to calculate lines which best fit data points around that line.

Linear regression and **correlation** are related, but different, tests. Linear regression finds the line that best predicts Y from X whereas correlation quantifies how well X and Y vary together. Note, however, that, after regression analysis, X and Y variables may also found to be correlated. When choosing which analysis to use, consider the following points:

- If you control X (eg. time, dose, concentration) select linear regression
- Only choose regression if you can clearly define which variable is independent X and which is dependent Y because regression finds the line of best fit by predicting Y from X

The **correlation coefficient** *r* was developed to measure how a set of data points fits a straight line and to determine whether there is a relationship between variables. Therefore *r* measures the strength of the association between X and Y. That is, the correlation coefficient *r* measures the degree of linear relationship or how well data are grouped around a line. A positive (+) *r* value is associated with an upward sloping line, whereas a negative (-) *r* value is associated with a downward line. That is, the dependent *y*-axis variable may increase (+) or decrease (-) with the independent *x*-axis variable. Therefore, the correlation analysis measures the degree to which a change in the independent variable will result in a change in the dependent variable: the closer the value of *r* to +/- 1, the stronger the relationship. The correlation analysis also calculates the *P* value or probability that the relationship is due to chance. Hence, although there is a relationship between the numbers of cyprid larvae of the honeycomb barnacle that arrive (independent variable) and then settle (dependent variable) on the rocky shore at Cape Banks (Fig. 13; correlation coefficient $r = 0.57$, probability $P < 0.001$, sample size $n = 43$), this relationship is not very strong. While the probability $P < 0.001$ shows that a relationship exists between larval supply and the numbers of larvae that settle, the relatively low *r* value indicates that other processes, apart from larval supply, influence settlement.

A high *r* correlation (eg. +/- 0.9) does not always mean that the independent variable **causes** the dependent variable to fluctuate; *r* values merely indicate trends. For example, although there may be a high correlation between number of tick bites in Sydney and number of people rescued in heavy surf at the beach, this may be due to a third variable. That is, high temperatures in summer may cause both an increase in numbers of people swimming (and drowning!) and an increase in numbers of ticks in the bush. Similarly, when height of brothers and sisters is plotted, there may be a tendency for tall girls to have tall brothers and short sisters to have short brothers but this may actually be determined by another variable such as their father's height.

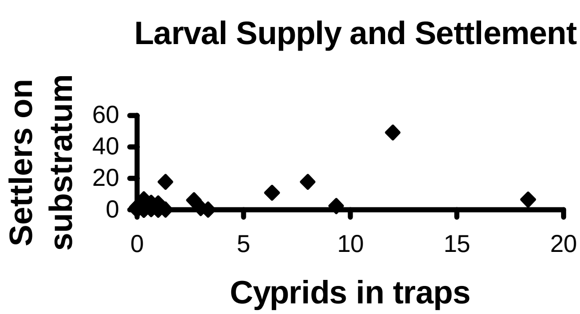


Fig. 13
The relationship between daily mean numbers of cyprid larvae and daily mean numbers of new settlers ($r = 0.57$, $P < 0.001$, $n = 43$) at one area low on the shore at Cape Banks, Botany Bay from late July to early December 1991. Each point represents the mean daily capture of cyprids in three traps and the mean daily settlement in three patches on the substratum next to these traps (Jeffery, 1997; Jeffery and Underwood, 2000).

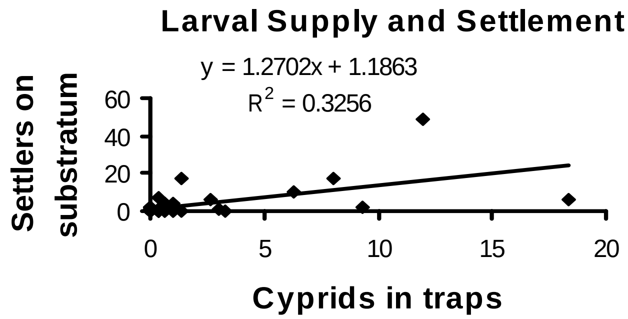


Fig. 14
The relationship between daily mean numbers of cyprid larvae and daily mean numbers of new settlers at one area low on the shore at Cape Banks, Botany Bay from late July to early December 1991. Each point represents the mean daily capture of cyprids in three traps and the mean daily settlement in three patches on the substratum next to these traps (Jeffery, 1997; Jeffery and Underwood, 2000).

A mathematical model, **regression analysis**, was also developed to obtain a line of best fit through a set of data points, but causation is implied with this analysis. That is, Y the dependent variable, increases or decreases as a function of X, the independent variable. The numerical measure r^2, the **coefficient of determination**, describes fit (see Fig. 14). It has a value that ranges from zero to one, and is the fraction of the variance in the two variables that is shared. For example, if $r^2 = 0.33$ (see Fig. 14), then 33% of the variance in Y can be explained by variation in X. That is, 33% of the variance is shared between X and Y. Therefore, a line fitted through Y values, by regression, is the line of best fit and is calculated with the following formula:
$Y = a + bX$
Where X is the independent variable, Y is the dependent variable, *a* is the intercept and *b* is the slope of the line.
That is, *a* (**the intercept**) is the value of Y when X is zero (*in some cases a is also zero*). For a fixed value of X (the independent variable), the population of Y (the dependent variable) is normally distributed with equal variances across Xs.

Note that, while both correlation (Fig. 13) and regression (Fig. 14) analyses were performed on the same data, the X and Y variables were found to be weakly correlated. That is, while larval supply influences settlement of arriving larvae, other processes are also operating (Figs. 13, 14).

7.2 Measurement

Units of measurement used depend on the type of data being collected. For example, it is standard practice to measure aperture-lengths, rather than shell-lengths, of barnacles (Fig. 15) to represent size. Adult *Chamaesipho* typically inhabit rocky shores as joined sheets of individuals making it very difficult to measure shell-lengths. Research, however, has shown a very strong relationship between shell-length and aperture-length of this barnacle (Fig. 16; Jeffery, 1997; Jeffery and Underwood, 2001). Hence the use of aperture-length as a unit of measure for this species is warranted (Figs. 8, 9, 10).

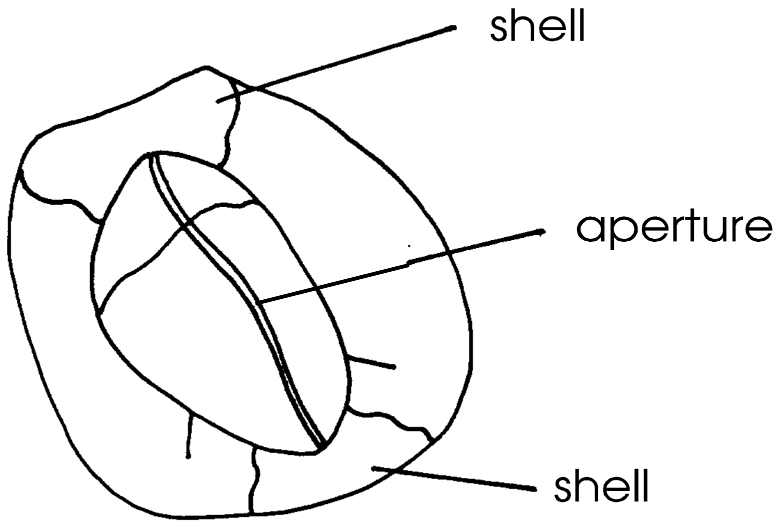


Fig. 15
Four-day old juvenile of the honeycomb barnacle *Chamaesipho tasmanica*. Aperture-length of this individual is 0.5 mm (Jeffery, 1997).

7.3 Interpolation and Extrapolation

Interpolation is where a value is estimated within the range for which data are available. **Extrapolation** is where values are estimated outside the range for which data are available.

Because variables are continuous we can **interpolate** from standard curves to find values that we haven't measured. For example, if we have only measured reaction rate of human salivary amylase at 0^o, 10^o and 40^oC, we will be able to determine the rates at 20^o and 30^oC because they fall between two known values. To do this we need to draw lines perpendicular to the *x*-axis (Temperature is the **independent variable**) from 20^o and 30^oC on the *x*-axis up to the data line and then across and perpendicular to the *y*-axis where the appropriate reaction rates can be read.

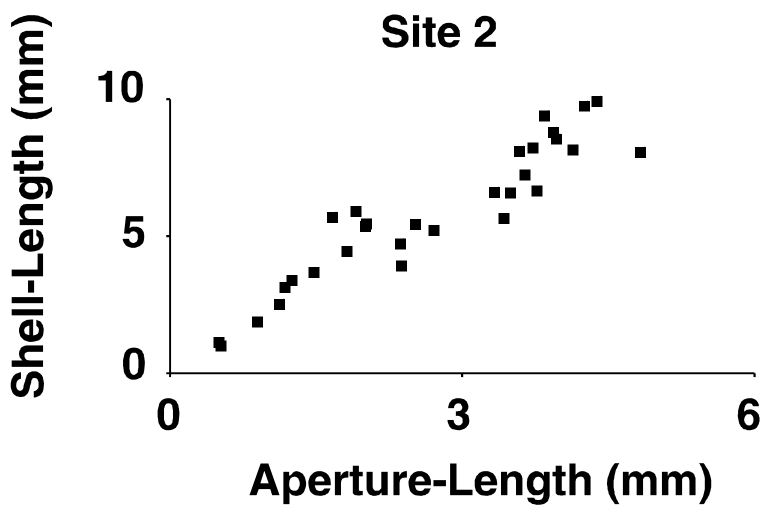


Fig. 16
Relationship between shell-length (mm) and aperture-length (mm) of *Chamaesipho* at Site 2 at Cape Banks, Botany Bay, New South Wales. The correlation ($r = 0.93$, $P < 0.001$, $n = 30$) was calculated on pooled raw data from ten barnacles randomly sampled from black and white negatives photographed in October, 1989 from each of Low, Mid and Upper levels of the shore (Jeffery, 1997; Jeffery and Underwood, 2001).

Figs. 17 and 18 show the area of saltmarsh and mangrove at Towra Point, estimated from old maps and aerial photographs (Mitchell and Adam, 1989). Data indicate that mangroves increased in area whereas saltmarsh decreased in area from 1942 to 1986 (Fig. 17; Mitchell and Adam, 1989). Prior to 1942, however, only one map, in 1882, distinguished between saltmarsh and mangrove. Using these data, a line of best fit was drawn to represent only the increase in mangroves over time (Fig. 18). **Interpolation** would suggest that the area of mangroves in 1910 would have been about 100 ha (1 sq km) (Fig. 18). **Extrapolation** suggests that in 1810 there were no mangroves at all (Fig. 18) although this may not be true. Hence much more caution is needed with extrapolation than interpolation. That is, the line of best fit in Fig. 18 provides an empirical description of data but we do not, as yet, have a model (model = cause; hypothesis = prediction) to explain the relationship between saltmarsh and mangrove so that extrapolation so far beyond the available data must be treated with great caution.

7.4 Standard Errors

7.4.1 What is Standard Error?

There is natural variability within populations so that not all individuals will respond or behave similarly when exposed to the same variables or treatments. For example, if a few gregarious barnacles are mashed up with seawater and equal volumes of this crude extract are painted on a rocky shore at several places close together where lots of cyprid larvae of the honeycomb barnacle are known to arrive to settle on the substratum, and where all other factors appear similar, different numbers of new settlers will be recorded at each place. In fact, settlement may be poor in some spots but aggregations of settlers may be recorded in others (Jeffery, 1997). That is, more larvae will choose to settle in some areas than others despite the presence of the same stimulus. Hence, not all larvae of this barnacle will settle in response to the presence of cues from adults even though the population is generally gregarious. Because of this variation among individuals in a population, **mean** or average values are used to represent the overall response of the whole population but the **standard error** (SE) is calculated to represent this variation. Therefore it is important to always replicate treatments or sampling so that a general picture of the population can be recorded and then the variation within this set of samples can be represented by the standard error (Table 3).

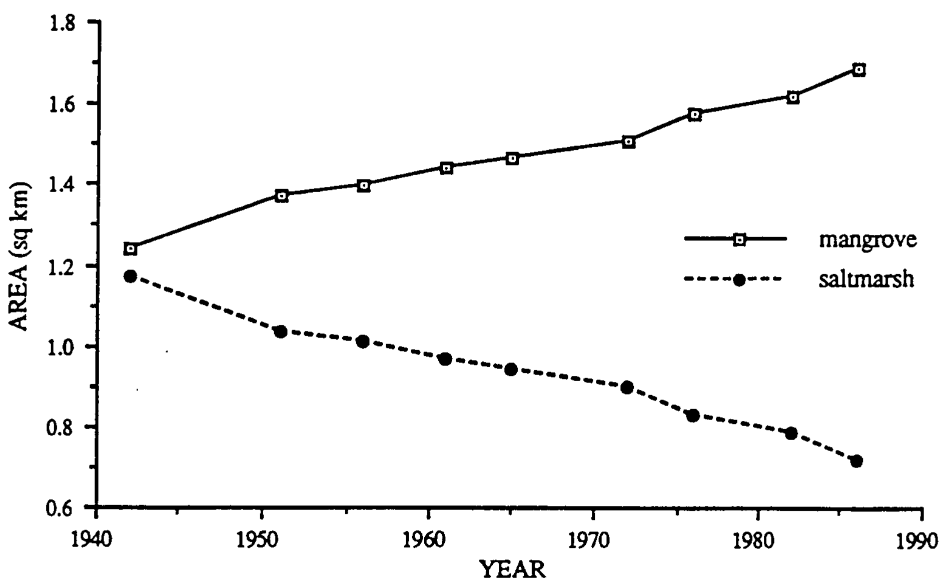


Fig. 17
Area of saltmarsh and mangrove on Towra Point, 1942-1986 (Mitchell and Adam, 1989).

Although the variance is the mathematically correct or absolute expression of how samples vary from the mean, standard deviations and standard errors are often used to express variability. Previously, the standard deviation was the convention for expressing the variance within populations but standard error, which incorporates size of sample, is now standard practice. Standard error is an estimation of the standard deviation of the means.

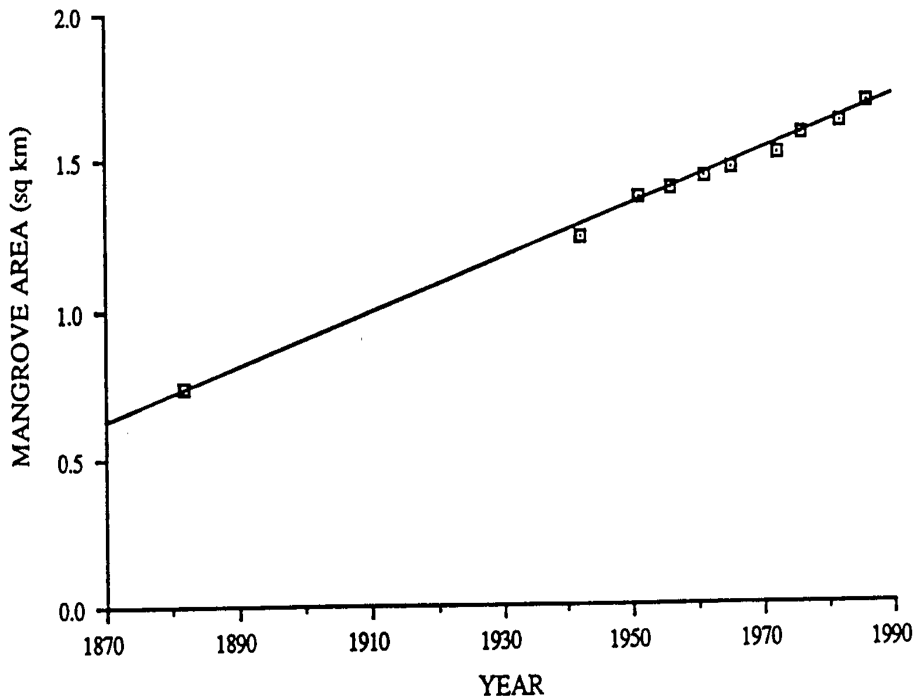


Fig. 18
Area of mangroves on Towra Point from 1882 (Mitchell and Adam, 1989).

7.4.2 Calculating the Standard Error

To find the **mean** ($\bar{x}$) number of barnacle cyprids caught in three traps at one place low on the shore at Cape Banks, Botany Bay in 1990 (Table 3), the total number of larvae (X) must be divided by the number of traps (*n*).

The mean, $\bar{x} = X/n = 173.00$ (That is, 519/3)

The **variance** s^2 is an estimation of how much each sample or replicate varies from the mean and is represented by the **standard deviation** *s* or, most likely, by the **standard error** SE = $s/\sqrt{n}$

The **variance** of the number of barnacle cyprids caught in three traps at one place low on the shore at Cape Banks, Botany Bay in 1990 (Table 3) can therefore be calculated using the formula:

$s^2 = \Sigma(X-\bar{x})^2/ n-1$ (That is, $(175-173)^2 + (127-173)^2 + (217-173)^2/ 2) = 2028$

And the **standard deviation**, which is the square root of the variance:

$s = \sqrt{\Sigma(X-\bar{x})^2/ n-1}$ (That is, $\sqrt{2028} = 45.03$)

And the **standard error**:

SE = $s/\sqrt{n}$ (That is, 45.03/$\sqrt{3}$ = +/- 26.00)

***Table 3**
Total numbers of cyprid larvae of the honeycomb barnacle *Chamaesipho tasmanica* captured in one set of traps (n = 3) at one place low on the shore at Cape Banks, Botany Bay, New South Wales from July to December, 1990 to 1993 (Jeffery, 1997; Jeffery and Underwood, 2000).

Time	**Trap1**	**Trap2**	**Trap3**	**Mean Number of Cyprids**	**S.E.**
1990	175	127	217	173.00	26.00
1991	78	99	61	79.33	10.99
1992	67	84	71	74.00	5.13
1993	4	29	29	20.67	14.43

**Note that, although raw data have been included here to help you to understand the concept of standard error, usually raw data are included in the Appendix and only means are recorded in tables.*

Therefore, in 1990 the mean number of cyprids of the honeycomb barnacle varied within one set of traps (n = 3) at one place on the shore by +/- 26.00 (Table 3). In other words, the numbers of larvae were estimated to vary within the range 147 – 199.
Because standard errors are an estimation of how much variation occurs within a population, obviously there was much more variation within sets of traps in 1990 than in any other years (Fig.2; Tables 2, 3).

Note that secondary students studying Science in New South Wales are not required to include Standard Error (SE) bars. However, because there is always variation within sets of samples (replicates) or within populations, if standard error SE bars were to be included in line graphs and in bar graphs they would represent this variability. Because scatter graphs often have individual data rather than means, SE bars can/should not be included. Obviously SE bars could not be used in frequency histograms where the whole data are included across a range of size intervals, nor in pie charts.

7.4.3 Inserting Standard Errors on Bar Graphs in *EXCEL

EXCEL is a widely used computer program that permits manipulation of tabular data and calculation of statistics

a) Separate Bars = One Data Series (Fig. 19)

Table 4 Data Setup for One Data Series

	Mean	**SE**

Low	276.0	*37
Mid	71.5	*57
Low	1.0	*3
Mid	5.5	*87

**Note that these are hypothetical, not calculated, SE's*

Highlight first two columns and four rows: Low, Mid and Means. Insert Chart.
Chart title, *x*-axis title, *y*-axis title. No grid, **no** legend. Finish.
Adjust scales, patterns, labels, tick bars on *x* and *y* axes.
Plot area colour. Use Control key to alter bar colours.
SE: Select all bars (Series 1).
Click on Y error bars, select top bar only, select SE, click on + Custom box.
Need to highlight SE's on data sheet and the relevant formula and data will come up in Custom box eg. =Sheet1!$C$4:$C$7.

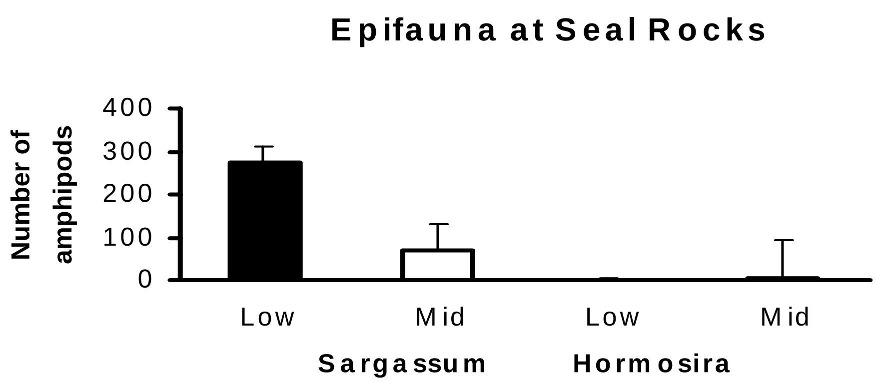


Fig. 19
Number of amphipods recorded on two species of algae (number of samples (*n*) = 3) at Seal Rocks.

b) Paired Bars = Two Data Series (Fig. 20)

Table 5 Data Setup for Two Data Series

	Mean	**Mean**	**SE**	**SE**
Low	276.0	1.0	*37	*3
Mid	71.5	5.5	*57	*87

**Note that these are hypothetical, not calculated, SE's*

Highlight first three columns and two rows: Low, Mid and Means. Insert Chart.
Chart title, *x*-axis title, *y*-axis title. No grid, **show** legend. Finish.
Adjust scales, patterns, labels, tick bars on *x* and *y* axes.
Plot area colour. Use Control key to alter bar colours.

SE: Select either Series 1 or Series 2 bars.
Click on Y error bars, select top bar only, select SE, click on + Custom box. Need to highlight SE's on data sheet and the relevant formula and data will come up in Custom box eg. =Sheet1!$D$36:$E$36.

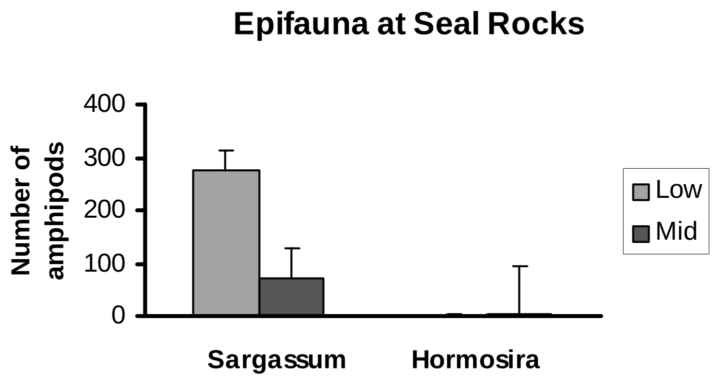
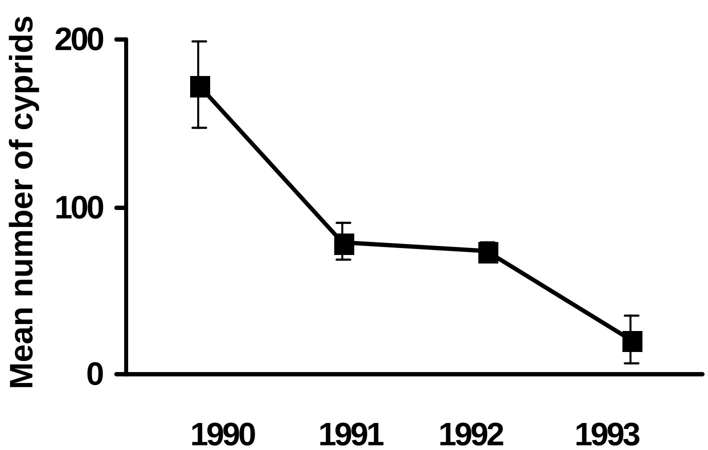


Fig. 20
Number of amphipods recorded on two species of algae (number of samples (*n*) = 3 at Seal Rocks.

8. Common Faults with Graphs

- No title – Fig. 21 (*correct figure = Fig. 2*)
- No figure caption – Fig. 21 (*correct figure = Fig. 2*)
- No variable labels – Fig. 22 (*correct figure = Fig. 5*)
- No units – Fig. 23 (*correct figure = Fig. 16*)
- No scale markings – Fig. 23 (*correct figure = Fig. 16*)
- Inappropriate scale choice – data points should be evenly represented – Fig. 24 (*correct figure = Fig. 5*)
- Incorrect choice of independent (*x*-axis) and dependent (*y*-axis) variables – Fig. 25 (*correct figure = Fig. 1*)
- Incorrect unit of measure – Fig. 25 (*correct figure = Fig. 1*)
- No legend when necessary – Fig. 26 (*correct figure = Fig. 20*)

Fig. 21 (*correct figure = Fig. 2)*: **no title; no figure caption.**



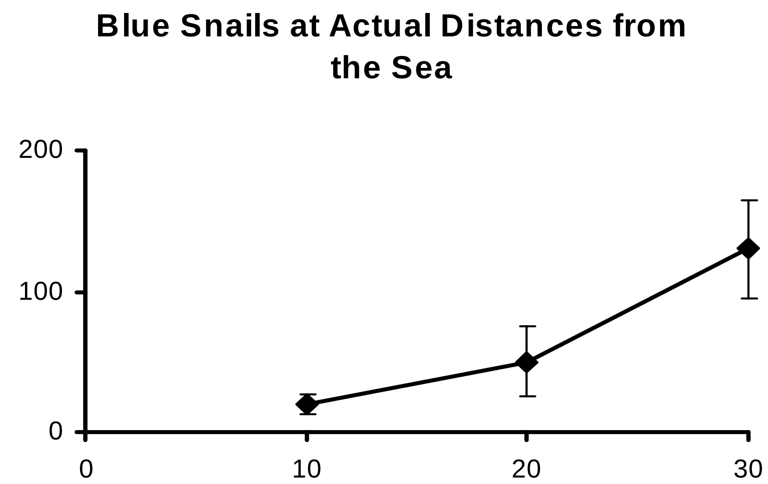


Fig. 22 (*correct figure = Fig. 5*): **no variable labels.**
Mean number of blue snails (number of samples (*n*) = 3) recorded in 50 cm x 50 cm quadrats at actual distances (10, 20, 30 m) from Mean Low Water (MLW).

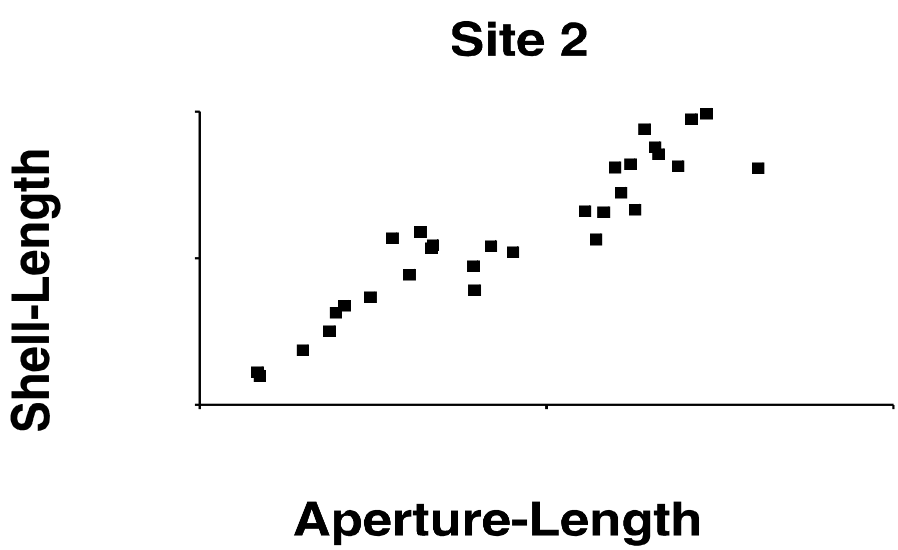


Fig. 23 (*correct figure = Fig. 16*): **no units; no scale markings.**
Relationship between shell-length (mm) and aperture-length (mm) of *Chamaesipho* at Site 2 at Cape Banks, Botany Bay, New South Wales. The correlation ($r = 0.93$, $P < 0.001$, $n = 30$) was calculated on pooled raw data from ten barnacles randomly sampled from black and white negatives photographed in October, 1989 from each of Low, Mid and Upper levels of the shore (Jeffery, 1997; Jeffery and Underwood, 2001).

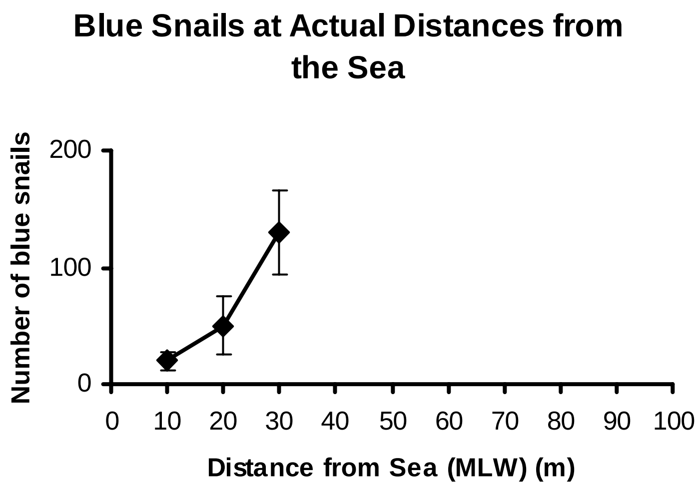


Fig. 24 (*correct figure = Fig. 5*): **inappropriate scale choice.**
Mean number of blue snails (number of samples (*n*) = 3) recorded in 50 cm x 50 cm quadrats at actual distances (10, 20, 30 m) from Mean Low Water (MLW).

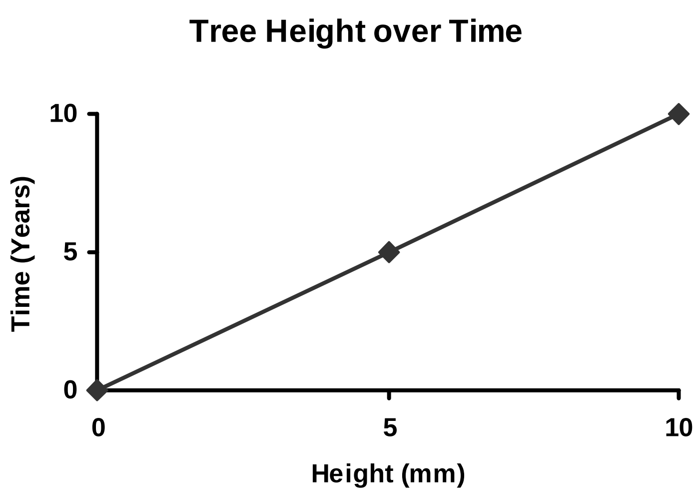


Fig. 25 (*correct figure = Fig. 1*): **incorrect choice of independent (*x*-axis) and dependent (*y*-axis) variables; incorrect unit of measure.**
Height of eucalypts measured over ten years.

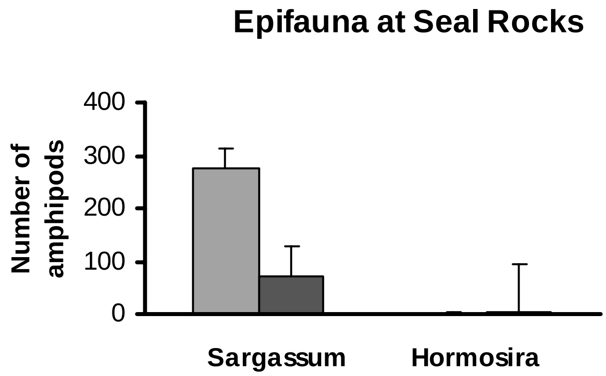


Fig. 26 (*correct figure = Fig. 20*); **no legend when necessary.**
Number of amphipods recorded on two species of algae (number of samples (*n*) = 3 at Seal Rocks.

9. References

Board of Studies, New South Wales, 1999. Stage 6 Syllabus, Biology, Preliminary and HSC Courses, Board of Studies, NSW, Sydney.

Hook, G. and Ingram, B., 1989. Data handling skills in biology, Longman Cheshire, Melbourne.

Jeffery, C.J., 1997. The ecology of the rocky shore intertidal barnacle *Chamaesipho tasmanica* in New South Wales, PhD thesis, University of Sydney.

Jeffery, C.J. and Underwood, A.J., 2000. Consistent spatial patterns of arrival of larvae of the honeycomb barnacle *Chamaesipho tasmanica* Foster and Anderson in New South Wales. J. Exp. Mar. Biol. Ecol. 252, 109-127.

Jeffery, C.J. and Underwood, A.J., 2001. Longevity determines sizes of an adult intertidal barnacle. J. Exp. Mar. Biol. Ecol. 256, 85-97.

Mitchell, M.L. and Adam, P., 1989. The decline of saltmarsh in Botany Bay. Wetlands, 8 (2), 55-60.

Walker, D., 1992. Energy, Plants and Man (second edition), Oxygraphics, Brighton.

www.math.virginia.edu/~der/usem170.html
www.statsoft.com/textbook/stgraph.html
http://surfstat.newcastle.edu.au/surfstat/main.html