

R Module 2



Univariate Descriptive Statistics

Certificate in EnviroStats (Non-Award)

This document is part of an online Certificate in EnviroStats (Non-Award) by the University of Canberra. Course enquiries can be directed to the address below. Expressions of interest in the course can be made online through:

<http://aerg.canberra.edu.au/envirostats>

Copies of this publication are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: bernd.gruber@canberra.edu.au, georges@aerg.canberra.edu.au

Copyright © 2011 Arthur Georges, Bernd Gruber [Converted the manuscript from SAS to R]

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, including electronic, mechanical, photographic, or magnetic, without the prior written permission of the author.

R is an open source statistical programme. It is developed by:

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Correct citation:

Georges, A. (2002). *Biometry: Statistics for Ecology and Natural Resource Management. Module 2: Univariate Descriptive Statistics*. Flexible Delivery Development Unit, Centre for the Enhancement of Learning, Teaching and Scholarship (CELTS), University of Canberra, ACT 2601, Australia. ISBN: 1 740880269

SPONSORED BY:**Materials development team:**

Author:	Arthur Georges, 2002, 2006, Bernd Gruber 2011
Instructional designer:	Peter Donnan, 2002
Editor:	Loretta Barnard, 2002
Graphic Design:	Peter Delgado, 2002
Desktop Publishing:	Kristi McDonald, 2004 Sue Bebbington, 2004
FDDU Project Manager:	Deborah Veness, 2002

Dynamic Web Page Design:	TCNI Software Solutions PO Box 47 LATHAM ACT 2615 Australia
--------------------------	--

First prepared in January, 2002 for Semester 1, 2002.
Reprinted January 2003 for Semester 1, 2003.
Reprinted January 2004 for Semester 1, 2004.
Reprinted November 2004 for Semester 1, 2005.
Revised and reprinted, June 2006
Revised and reprinted, June 2007
Revised and converted to R, January 2011

Published by Technology & Educational Design Services

(TEDS)
University of Canberra
ACT 2601, AUSTRALIA

Content

Content	3
Lesson 1: Key Concepts	5
Sample and summarise	5
The frequency tabulation	6
Histograms and bargraphs	8
Statistics	9
Averages	9
Measures of variability	10
Statistics of shape	12
Where have we come?	13
Lesson 2: Application Notes	14
Levels of measurement	14
Nominal level	14
Ordinal level	14
Interval level	15
Ratio level	15
Levels of measurement and descriptive statistics	16
Normality	19
Analysing Normal data	19
Analysing non-normal data	20
Where have we come?	22
Lesson 3: Step-through Examples	23
Example 2-1: Burton's Bush Rat	23
Start a R-session	24
Prepare the data	25
Frequency tabulation	27
Barchart	28
Barcharts with subgroups	30
Multiple plots	32
Source	33
Example 2-2: Chesapeake blue crabs	34
Start a R-session	35
Prepare the Data	35
Frequency Tabulation	36
Barchart	38
Enhanced cross tables	39
Regrouping nominal data	42
Source	43
Example 2-3: Water Quality of Lake Burley Griffin	44
Start a R-session	45
Prepare the data	45
Summary statistics	46
Graphical Presentation	54
Report the results	55
Non-Normal data	56
Report the results	58
Example 2-4: Blue crab sizes	59
Start a R session	59
Prepare the data	59
Summary Statistics	60
Report the results	65
Normal data	65
Source	69
Where have we come?	70
Lesson 4: Some Challenging Exercises	71

Exercise 2-1: Trawl Catch Statistics	71
Exercise 2-2: Water Chemistry of Lake Carcoar	73
Exercise 2-3: Inflows to Burrinjuck Dam	75
Analysis of inflows	75
Analysis following transformation	76
Where have we come?	77
References	77

Lesson 1: Key Concepts

Sample and summarise

Descriptive statistics have an important role to play in science. When specific problems are addressed in science, data needs to be collected, analysed and presented in a concise form so that others may benefit from what has been found.

In the biological sciences, work is often undertaken on **populations** of organisms. Almost invariably, these populations are too large to study in their entirety. For example, an ornithologist could not possibly hope to capture all of the individuals of the bird species he chose to study. Studies usually require that subsets of entire population be taken. These subsets are called **samples**.

Most of the advantages of sampling are fairly obvious. It is cheaper and quicker to obtain information from a sample than from an entire population, which may be very large. More comprehensive data can be obtained by studying a relatively small sample thoroughly, rather than a large population superficially. Sampling may be the only means of obtaining data if the process of measurement is destructive. The biologist examining ovaries to judge the proportion of female seals that breed each year will surely refuse to slaughter all individuals to gather data. Whatever the reason for taking a sample, it is important for the sample to be **representative** of the population from which it is drawn. Only then will it be possible to extend your findings to the entire population.

One of the first steps taken in any analysis is to summarise the information contained in the sample. This is necessary because blandly perusing raw data in samples of even moderate size (Table 2-1) is unlikely to provide great insight. Nor is it worthwhile to present entire raw data set in a report or manuscript, except perhaps in an appendix, because the reader cannot be expected to glean the trends that the author considers important, directly from the raw data. Few readers would bother.

The objective of this module is to introduce descriptive analyses appropriate for a single variable.

The frequency tabulation

The **frequency tabulation** is a very popular method for summarising data because even very large data sets can be condensed to a manageable form without substantial loss of information. Consider the data on the lengths of shoots of *Banksia ericifolia* shown in Table 2-1.

Table 2-1.
Lengths (in cm)
of 500 shoots of
the shrub
Banksia ericifolia
from heaths in
the
Commonwealth
Territory of
Jervis Bay.

28.4	29.1	25.7	26.2	25.8	30.2	26.2	35.8	33.4	29.0	31.6	33.8	22.6	34.8
25.1	36.8	29.9	38.3	27.1	32.3	29.0	27.7	28.1	28.9	21.8	25.0	23.2	26.8
29.9	36.3	26.0	21.2	19.8	36.7	21.1	34.6	29.6	32.6	27.2	34.2	26.7	27.6
25.8	23.2	28.8	38.2	32.7	38.7	33.2	24.5	21.6	28.6	19.4	27.4	43.7	25.5
35.0	25.1	24.7	29.5	24.9	29.2	19.5	20.1	30.3	38.9	28.2	26.2	29.4	22.4
36.4	31.8	31.0	39.4	28.8	31.8	28.7	37.0	25.5	19.3	44.0	38.0	28.6	36.5
29.1	21.1	30.4	31.2	38.0	39.0	19.3	27.6	19.1	32.5	26.8	39.9	36.1	41.5
33.2	26.5	38.1	14.9	33.2	27.8	24.7	24.9	25.0	33.1	24.1	19.7	19.1	31.4
26.9	22.5	25.5	33.0	19.4	26.8	24.6	37.5	19.8	43.7	38.1	30.8	34.5	22.8
34.2	33.6	42.5	19.0	25.8	34.0	34.4	42.0	35.4	31.5	40.6	19.2	24.9	33.5
32.0	38.4	29.1	29.4	29.3	26.8	32.4	25.2	28.5	29.8	22.8	17.1	29.6	33.3
31.7	22.4	21.7	20.1	21.6	23.5	33.2	33.0	29.6	36.9	26.8	38.1	29.8	21.2
23.6	16.2	27.3	33.3	21.6	30.2	22.5	33.0	38.3	29.5	34.9	30.3	26.0	24.5
31.1	31.7	31.6	41.7	25.9	32.3	35.7	31.6	26.0	26.6	26.3	19.6	22.0	40.2
29.1	15.8	22.1	23.2	25.4	28.4	20.2	25.5	26.9	32.7	28.8	39.6	31.9	31.9
29.8	27.1	36.9	32.7	24.8	18.0	40.2	28.0	26.8	41.9	15.8	16.2	33.6	34.8
31.5	27.4	37.2	30.6	32.2	34.8	28.2	31.3	34.3	32.0	33.3	30.1	20.5	37.3
20.6	27.2	25.0	26.1	34.5	44.9	40.5	32.1	40.4	35.7	33.9	29.3	28.1	34.3
29.3	24.7	37.0	36.9	34.8	29.8	22.8	32.3	34.8	29.6	33.6	22.8	31.6	34.7
24.0	30.5	31.6	28.7	20.8	14.6	23.4	26.3	31.9	32.5	34.3	25.7	36.0	37.7
32.4	36.4	24.1	33.1	26.3	35.7	26.4	34.7	27.5	39.6	16.5	30.2	23.8	23.7
30.5	30.1	21.3	27.1	19.0	25.4	36.5	22.6	25.5	30.0	34.4	30.6	32.7	30.4
29.2	30.2	20.8	30.3	27.8	32.9	28.2	20.6	33.6	22.2	37.5	30.0	24.2	18.8
26.1	29.7	32.0	22.2	29.2	21.5	31.4	43.1	35.9	14.9	24.6	26.2	33.4	29.9
38.8	21.9	25.6	29.7	29.9	32.5	30.4	29.2	40.9	14.1	22.1	20.0	24.3	28.6
22.4	19.9	34.8	33.4	28.0	29.1	27.2	18.8	36.2	27.8	20.2	21.9	27.0	21.9
29.9	21.8	33.1	30.4	30.8	33.9	27.1	27.6	37.2	30.9	31.4	41.9	24.7	25.8
28.3	34.3	34.0	29.0	30.9	24.4	29.0	25.4	30.5	31.1	33.9	15.7	40.5	29.9
26.3	38.3	24.8	23.5	29.3	37.8	29.9	28.6	27.4	29.9	33.5	17.0	34.1	30.9
24.3	20.7	22.5	39.5	32.0	27.6	36.3	22.0	28.4	19.1	25.8	25.7	33.9	43.0
16.8	43.9	27.9	44.4	29.7	23.0	26.8	43.4	29.4	26.7	16.5	22.1	23.0	39.4
27.8	33.1	34.9	20.5	25.4	10.0	28.2	31.0	10.6	28.4	16.5	22.3	17.6	24.2
21.9	27.0	26.5	29.2	24.9	18.4	24.1	28.3	29.0	29.1	18.8	36.7	24.7	23.2
26.2	32.6	22.3	31.7	37.1	35.6	19.5	26.9	24.8	19.2	25.1	22.1	37.9	28.3
29.9	42.1	36.6	25.5	34.2	22.4	40.5	21.2	32.3	31.5	34.2	34.5	39.2	29.3
29.3	31.6	23.2	32.1	20.3	27.8	22.9	32.5	24.5	36.5				

There are 500 measurements, quite a formidable data set. By inspection, the minimum shoot length is 10.0 and the maximum is 44.9 cm. These values define the sample **range**. We now need to subdivide the range into intervals or classes, each of equal size. It is generally advisable to round the minimum value down and the maximum value up to appropriate values when deciding on class intervals. In this case it seems sensible to divide the range 10 to 45 cm into seven intervals each 5 cm wide.

If we count the number of shoots that lie in each of the seven intervals, we have the basis for frequency tabulation. Such a tabulation is shown in Table 2-2. The frequency column was obtained by counting the number of measurements that lie within each class.

The percent frequency column was obtained by representing each count as a percentage of the total count. The cumulative frequency and cumulative percentage frequencies were obtained by progressively summing the corresponding frequencies.

Table 2-2.
A frequency tabulation of shoot lengths for the shrub *Banksia ericifolia* from Jervis Bay

Length	Frequency	Percent	Cumul. Freq	Cumul.Percent Frequency
10<x≤15	6	1.2	6	1.2
15<x≤20	35	7.0	41	8.2
20<x≤25	93	18.6	134	26.8
25<x≤30	155	31.0	289	57.8
30<x≤35	130	26.0	419	83.8
35<x≤40	57	11.4	476	95.2
40<x≤45	24	4.8	500	100.0

Frequency tabulations provide summaries of data sets without substantial loss of information. In this case there has been minimal information lost — for example, the average shoot length calculated directly from the frequency tabulation (using the class midpoints) of 28.85 is in close agreement with the figure of 28.97 calculated from the raw data. A reader of a paper containing a frequency tabulation would have access to almost as much information as if the entire data had been published, yet the table takes up far less space and would take up little more room if based on 5 million measurements rather than only 500.

Frequency tables included in reports or papers do not generally contain all four types of frequencies shown in the above tabulation. The most usual columns to include are:

- Class intervals
- Raw frequencies
- Percentage frequencies

Percentage frequencies, although the most commonly used, can be very misleading if the sample size is small. For small samples, frequency tabulations should contain raw frequencies only. When dealing with a large sample, you may choose not to include the raw frequencies, but you must provide the total sample size upon which the percentage frequencies were based.

Histograms and bargraphs

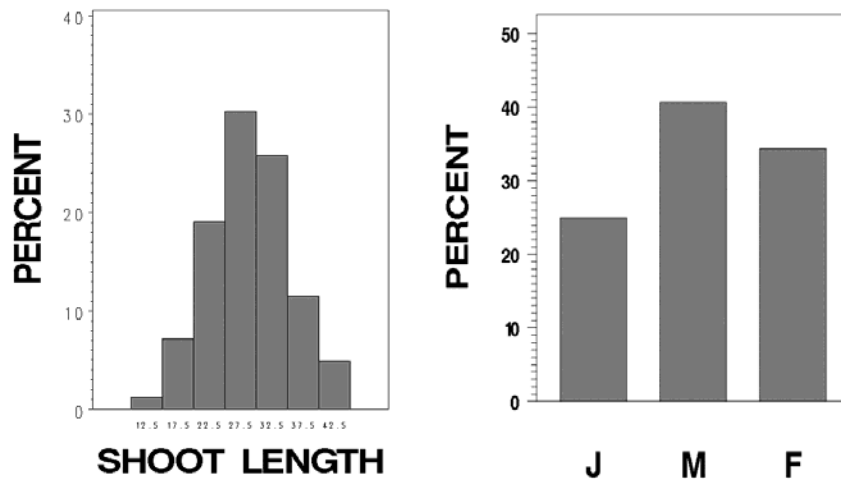
Frequency tabulations can be represented as a graph of frequency (raw or percentage) against the measurement variable. The information contained in the frequency table of shoot lengths of *Banksia ericifolia* is graphically represented by the **histogram** shown in Figure 2-1. The shape of the frequency distribution is usually of greater interest than the absolute heights of each column, so histograms are usually constructed from percentage frequencies.

The shoot lengths of *Banksia ericifolia* are examples of continuous measurements — they are free to take any whole or fractional number within their range. For discrete measurements such as hair colour or sex class, frequency tabulations can be represented as bargraphs. Bargraphs are similar in construction to histograms except that the vertical columns used represent class frequencies are spaced to indicate that the data are discrete (Figure 2-2).

Modern statistical packages have great versatility in graphical output. Experimentation with options described in the manuals will enable you to produce frequency polygons, composite histograms, horizontal bargraphs, block diagrams and pie charts.

Figure 2-1 (left).
A histogram showing the distribution of lengths for shoots of *Banksia ericifolia* ($n=500$).

Figure 2-2 (right).
A bargraph showing the relative proportions of unsexed juveniles (J), mature males (M) and mature females (F) in a population of pig-nosed turtles at Pul Pul Billabong in Kakadu National Park ($n=32$).



Statistics

The term **statistic** does not only refer to the academic discipline that deals with measurement, summary, inference and hypothesis testing. The term refers also to single-valued characteristics of samples. The sample mean, median, standard deviation and skewness are all statistics. They each describe some aspect of the sample.

Although the frequency tabulation (or histogram) summarises a data set with little loss of information, as a description it lacks sufficient definition to satisfy most researchers. Various summary statistics are usually presented in addition to or in place of the frequency tabulation or histogram.

Commonly used statistics fall into one of three categories — **averages** or statistics location, statistics of dispersion or **variability**, and statistics of **shape**.

Averages

An average is a single value that summarises the position of the sample measurements with respect to the range of values possible for the measurements. When asked how big the quoll *Antechinus stuartii* is, it would be appropriate to answer with the mean adult body length of 5.3 cm. This value might not give an accurate indication of the size of a particular individual *Antechinus*, but it does give a good indication of the size of *Antechinus* in general.

Several averages are available. The most common is the **Arithmetic Mean** calculated by summing all the measurements in a sample and then dividing by the total number of measurements, as follows:

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

The **Median** is the value that has 50% of values greater than it and 50% less than it. To calculate the median, you must rank the measurements in order of magnitude, then select a value that equally divides the number of measurements in the sample.

The **Mode** is the most commonly occurring value in a sample. For continuous data that have been grouped, it is sensible to define it as the midpoint of the class interval in a frequency tabulation that contains the most values. Interpolation formulae are available to improve the answer.

Other averages are in use, such as the harmonic mean and the geometric mean. These have specific applications, and they will not be dealt with further here.

Measures of variability

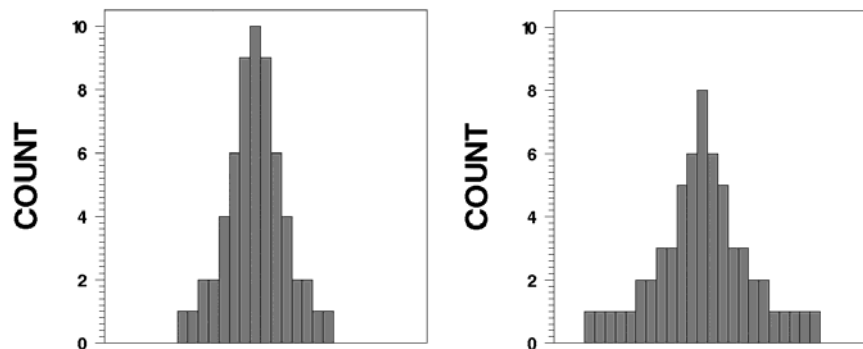
It is possible to imagine two histograms with the same mean but which differ in the spread of measurements about their means (Figure 2-3). Clearly an average does not provide an adequate summary of a data set. Other statistics are required, among which are statistics of variability or spread of measurements about the mean.

A simple measure of variability is the range. It is the difference between the largest and the smallest values in a sample. The range is obviously affected by even a single outlying value and for this reason is only a rough estimate of the variability of all observations in a sample. The range is also of limited value because it cannot be used easily in sampling theory, upon which so much of statistics depends. By far the most important measure of variability is the **Standard Deviation**.

$$S_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n-1}}$$

The deviations of all sample observations from the sample mean are squared to remove negative signs, then summed. The summed squared deviations or **Sums of Squares** are then divided by $n-1$. The square-root is then taken to ensure that the units of the standard deviation are the same as those of the measurement variable. The standard deviation is a sort of average absolute deviation of measurements from their mean, though quite distinct from the mean absolute deviation defined in many textbooks but no longer widely used.

Figure 2-3.
Two histograms
with identical
means but
differing in
variability about
those means.



The **Variance** or **Mean Square** is simply the standard deviation squared.

The **Coefficient of Variation** is used to compare the variability of two samples that have widely differing means. In such cases it is useful to measure the variability in relative terms by dividing the standard deviation by the mean. The result is the coefficient of variation, usually expressed as a percentage.

$$CV = \frac{S_Y}{\bar{Y}} \cdot 100\%$$

The coefficient of variation has the added advantage in that it has no units, and can be used to compare the variability of two very different measures, for example, to compare variability in body weight and in the concentration of a hormone in the blood.

The **Inter-Quartile Range** is another useful measure of variability. Recall that the median is the value below which 50% of all values in the sample lie. We might similarly define the 1st Quartile as the value below which 25% of all values in the sample lie and the 3rd Quartile as the value below which 75% of values lie. The Inter-Quartile Range is the difference between the first and third Quartiles and is a measure of variability that is superior under some circumstances to the standard deviation.

The 5th and 95th **percentiles** cut off 5% of the most extreme values in the distribution of values for the sample. The 1st and 99th percentiles may be similarly defined. As with Quartiles, a list of percentiles may be far more informative than the standard deviation for summarising the spread of values about the mean or median, especially when the spread is asymmetrical.

The **Evenness Index** J' (Pielou, 1966) is an appropriate measure of the variability among nominal level measurements such as hair colour or sex class.

$$H' = \frac{n \log n - \sum_{i=1}^k f_i \log f_i}{n}$$

$$J = \frac{H'}{\log k}$$

where f_i represents the class frequencies, k is the number of classes and n is the total number of objects measured. The derivation and relative merits of several measures of variability for nominal level data are discussed by Zar (1984, Chapter 4).

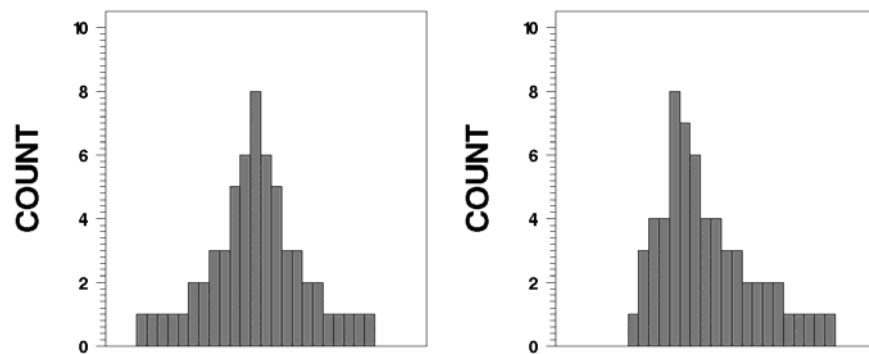
Statistics of shape

It is possible to have two sample frequency distributions with the same arithmetic mean (ie they are in the same overall position on the horizontal axis) and with the same standard deviation (ie variability about the mean is the same on average) but with different overall shape (Figure 2-4).

There are two statistics useful for describing shape. **Skewness** is another name for asymmetry which means that one tail of the frequency distribution is drawn out more than the other. A skewness of zero implies a symmetrical shaped histogram, a negative value implies skewness to the left, and positive value implies skewness to the right. The skewed histogram of Figure 2-4 is skewed to the right.

Kurtosis is a measure of how "peaked" (**leptokurtic**) a frequency distribution is or how "flattened" (**platykurtic**) it is. A negative value indicates platykurtosis, and a positive value indicates leptokurtosis.

Figure 2-4. Two histograms with the same mean and the same standard deviation, but differing in shape.



Where have we come?

In this lesson, we have covered the basics of descriptive statistics for data comprised of a single variable. You should now appreciate

- The distinction between a **sample** data set and the **population** from which it is drawn.
- The value of the **frequency tabulation** for presenting data in a concise form with minimal loss of information, and that frequency tabulations can be presented in graphical form as **histograms** and **barcharts**.
- The range of statistics available for measuring **central tendency**, and the distinction between them in theory and practice. You have the arithmetic mean, the median and the mode.
- The range of statistics for measuring **dispersion** of sample values about their mean, and the distinction between them. You have the standard deviation, the range, the interquartile range and percentiles, the coefficient of variation and the evenness index, each with their own particular application.
- The range of statistics for measuring the **shape** of a sample distribution. You have the skewness statistic and the kurtosis statistic, and should appreciate the distinction between the two.

Lesson 2: Application Notes

Levels of measurement

There are two considerations which, more than any other, will determine your choice of descriptive statistics. The first relates to the type of measurements you take, because not all descriptive statistics are appropriate for all types of measurement. The second is whether or not your sample of measurements is taken from a normally distributed population.

In its broadest sense, measurement is the process of abstracting a value from each item or entity that is the subject of study. For organisms we might choose to measure the variables length, weight, and body temperature or the variables sex, colour, presence or absence of evidence of lactation, and species. Measurement of a property means assigning a value, not necessarily numerical, to represent it.

Measurements can be made at different levels of precision, or **Levels of Measurement** as they are so called. The level of measurement often dictates the choice of valid summary statistics, and the statistical analyses that follow, and so are described in some detail below.

Nominal level

Measurements are at the **nominal level** if the items being measured are simply assigned to one of several classes that may be denoted by numbers or alphabetic codes.

Measurements at the nominal level place units in categories, nothing more. The order in which the categories are presented is quite arbitrary. Examples include sex, species in a rainforest assemblage, genetic phenotypes among second generation offspring, colour of the scales on the dorsal surface of *Crocodylus johnstoni*, identity number etc.

Ordinal level

Measurements are at the **ordinal level** if, in addition to having all the properties of nominal measurements, they can be ordered on magnitude

Sperm in the epididymides of a seasonal breeder may be recorded as absent, sparse, common, abundant or very abundant. Military ranks can be considered ordinal measurements. A General is greater in military rank than a Captain who is greater in rank than a

Sergeant. Moh's scale of hardness is another good example of measurement at the ordinal level.

Note that in all of these cases we can say that one measurement is greater than another, but we cannot say by how much it is greater. Is the difference between absent and sparse comparable to the difference between abundant and very abundant? Does subtracting a Captain from a General have any meaning? A diamond with a Moh's hardness of 10 scratches corundum with a Moh's hardness of 9 and corundum scratches topaz with a Moh's hardness of 8. This does not mean that as corundum is one unit harder than topaz, then diamond is in absolute terms one unit harder than corundum. In fact any mineralogist will tell you that is untrue. If hardness is measured as the force per unit area necessary to produce a permanent deformation in a polished surface (the Knoop Scale), then diamond has a hardness of 7000 compared to 2100 for corundum and 1340 for topaz. Measurement at an ordinal level as on Moh's scale, allows us to say that one item has more of a property than another, but not how much more.

Interval level

Measurements are at the *interval level* if, in addition to having all the properties of ordinal measurements, they can be used to determine by how much one measurement differs from another.

At the interval level, the measurements can not only be ranked, but differences between measurements have a real meaning. Obviously measurements at the interval level must be represented by numerical values, unlike measurements at the nominal and ordinal levels. Consider measurement of temperature in degrees centigrade. The difference between 30°C and 60°C is equal to the difference between -10°C and 20°C and half the difference between 50°C and 110°C.

However, an object heated to 40°C does not have twice as much heat in it than when it was initially at 20°C. This becomes clear when you realise that the centigrade scale has its zero value set arbitrarily at the freezing point of water. On the Fahrenheit scale, 40°C and 20°C correspond to 104°F and 68°F respectively. The former temperature (104°F) is no longer double the latter (68°F). Centigrade and Fahrenheit temperature scales are not measured in relation to true zero and so are measured at the interval level.

Ratio level

Measurements are at the *ratio level* if, in addition to all the properties of interval measurements, they are measured relative to a true zero point, as opposed to an arbitrary zero point.

Zero represents absence of the property being measured. One can validly present ratios of the measurements in the knowledge that those ratios are absolute and not dependent on the scale chosen for measurement. Examples of ratio level measurements include height in cm, weight in grams, counts, plant densities in plants/m².

Levels of measurement and descriptive statistics

Not all of the descriptive techniques covered in these notes can be applied to all types of data. The summaries appropriate for describing **nominal data** are frequency tabulations, bargraphs, the mode and the Evenness Index. One might choose to record the colour of head scutes of the freshwater crocodile by assigning the colours to one of seven nominal classes. It is not meaningful to talk of average scute colour (say) in the sense of calculating an arithmetic mean, yet the number of crocodiles with scutes of each colour can be readily tallied, graphed and the modal class determined. The standard deviation cannot be calculated for scute colour, but the Evenness Index will provide a measure of the spread of individuals across the range of scute colour classes.

The summaries appropriate for describing **ordinal data** are frequency tabulations, bargraphs, the mode, the median and percentiles. Recall that the median is the value for which half the measurements are smaller and half are larger. It is because the ordinal scale is ranked that one can calculate the median and percentiles. The arithmetic mean should not be calculated for ordinal data no matter how respectable the result may appear. This is because the intervals between units on the scale are not necessarily of equal size and are, in all respects apart from their order, arbitrary. On ordinal data, the operations of arithmetic (addition, subtraction) are not valid, and so it is not valid to calculate the arithmetic mean. Nor is it valid to calculate the standard deviation, variance or coefficient of variation, even though the computations may be possible. The inter-quartile range replaces them as an appropriate measure of variability for ordinal data.

With the exception of the coefficient of variation, all of the descriptive statistics described in these notes may be used to summarise **interval data**.

The distinction between interval and **ratio data** is often glossed over in statistical texts; in fact the two levels of measurement are often discussed together as if there were no practical differences between them. To dispel this view, consider an example. A researcher has decided that variability in environmental factors is more important in stimulating germination of a plant species than are the absolute values of ambient temperatures and humidity. She wishes to know which of the two factors, temperature or humidity, can be considered the more variable. Note that the temperature is measured on the

interval scale whereas humidity is measured on the ratio scale (zero humidity means zero moisture). Her data are shown in Table 2-3.

Table 2-3.
Temperature and humidity measurements collected in an investigation of plant germination.

	Temperature °C	Humidity %
	21.0	72.5
	22.5	74.0
	27.5	87.5
	28.0	88.0
	24.5	79.5
	26.0	81.0
	23.0	78.0
	27.5	79.5
MEAN	25.0	80.0
SD	2.65	6.02
N	8	8
CV	10.6%	7.5%

To determine which factor is more variable, it is of little value to compare the standard deviations directly because the mean humidity of 80% is much greater in magnitude than the mean temperature of 25°C. We must compare the coefficients of variation. The coefficient for temperature of 10.6% is greater than the coefficient for humidity of 7.5%, so we conclude that temperature is the more variable of the two. However, had the researcher chosen to measure temperature in °F rather than °C, she would have obtained figures of 77°F for the mean, 4.76°F for the standard deviation and only 6.2% for the coefficient of variation. Now humidity is the more variable.

This is clearly unsatisfactory. The results of an analysis should not depend on an arbitrary decision to measure temperature in degrees Fahrenheit rather than degrees centigrade. The anomaly arises because temperature is not measured with respect to a true zero; it is measured on the interval scale. The coefficient of variation can only be validly calculated for data measured at the ratio level.

Table 2-4. The applicability of various summary techniques to data measured at each of four levels of measurement.

Technique	Data Type			
	Nominal	Ordinal	Interval	Ratio
Frequency Tabulations	YES	YES	YES	YES
Bargraphs	YES	YES	YES	YES
Mode	YES	YES	YES	YES
Evenness Index	YES	YES	YES	YES
Median	no	YES	YES	YES
Quartiles and Percentiles	no	YES	YES	YES
Interquartile range	no	no	YES	YES
Histograms	no	no	YES	YES
Frequency Polygons	no	no	YES	YES
Arithmetic Mean	no	no	YES	YES
Standard Deviation	no	no	YES	YES
Variance	no	no	YES	YES
Coefficient of Variation	no	no	no	YES

Level of measurement depends on the method of measurement, not on the property being measured. Most science students would be familiar with measurement of temperature on the Kelvin Scale or Absolute Scale. 0°K is the temperature at which molecules stop vibrating and there is literally no heat. 0°K is a true zero and an object at 20°K has twice as much heat as at 10°K . Temperature in degrees Kelvin is measured at the ratio level. If we choose instead to record temperature in $^{\circ}\text{C}$, relative to the arbitrarily chosen freezing point of water, we are measuring it at the interval level. Alternatively we might choose to record it as freezing, cold, cool, neutral, warm, hot, boiling — an ordinal level of measurement (Table 2-5).

Table 2-5. Temperature measured at each of the four levels of measurement

Ratio	Interval		Ordinal	Nominal
	$^{\circ}\text{C}$	$^{\circ}\text{F}$	Touch	Treatment
273	0	32	Freezing	C
283	10	50	Cold	A
293	20	568	Cool	D
313	40	104	Hot	E
353	80	176	Very Hot	B

The same property is measured, but at four different levels and a higher level can be converted to a lower level, but with some loss of information. The consequence of this is that statistics that are valid at a low level of measurement can always be applied to measurements at a higher level (Table 2-4).

Normality

Analysing Normal data

Most natural groups of objects show variation. Humans differ in height, even if of the same sex, race and age. In many instances, measurements of similar objects vary about their mean according to a well defined function, the Normal or Gaussian distribution function. The second major consideration in deciding upon appropriate descriptive statistics is whether or not the data can be adequately modelled by a normal distribution.

The normal distribution has the following characteristics:

- It is symmetric about its mean, median and mode. Hence a normal distribution has a skewness of zero.
- It is bell-shaped, with a kurtosis of zero.
- It is a continuous curve defined for ordinate values from minus infinity to plus infinity.
- It is completely defined by its mean and standard deviation. That is, if you know the mean and standard deviation of the normal curve, you can calculate its exact equation.
- 95% of observations fall in the range defined by the mean plus or minus 1.96 standard deviations and 99% fall in the range defined by the mean plus or minus 2.58 standard deviations.

A normal distribution can be expected to be a good model of the distribution of data if, in general, the value taken by a measurement of one particular object is influenced by a myriad of small and independent factors. Hence one would expect the birth weights of female infants of Caucasian descent born full-term to be normally distributed. One would not expect the weights of individuals in a sample of people taken at random to be normally distributed because there are a few variables, which have an over-riding influence on a person's weight. Age is one, sex is another. The distribution of body weights in a sample will depend very much on the distribution of ages of the people in the sample.

Similarly, we might expect counts of a plankton species in column samples taken at the same time and place to be normally distributed provided the average count is around 50. If the average count is around four, however, a normal distribution is not likely because counts in individual column samples can vary below the mean by only four places (a negative count is not possible) but is not limited in how far it can vary above the mean. Such a distribution would be skewed to the right.

A normal distribution with a mean and standard deviation equal to those of the sample, has been fitted to the frequency distribution of shoot lengths of *Banksia ericifolia* (Figure 2-5). Clearly the normal distribution is an adequate model for these data, as the fit is very good. A practical consequence of this Normality is that the data on shoot lengths of *Banksia ericifolia* can be adequately summarised by the sample mean and standard deviation alone, because they are all that is required to specify exactly the corresponding normal distribution. With these two statistics, we can re-construct a close approximation of the frequency tabulation for the original data, and as we have seen, this tabulation contains almost as much information as the original data itself.

Summary statistics for a sample drawn from a normally distributed population would usually include the range of values encountered (10.0 to 45.9 cm), the arithmetic mean (28.97 cm), the standard deviation (6.37 cm) and the size of the sample from which these statistics were calculated ($n = 500$). All other information, including the frequency tabulation, the mode, median, percentiles, sample skewness and kurtosis would be superfluous.

Analysing non-normal data

For data that do not conform to the theoretical normal distribution, the situation is more complex. No longer will the mean and standard deviation suffice in order to reconstruct the frequency distribution of the raw data. No longer would we expect only 5% of values to lie outside the mean plus or minus 1.96 standard deviations. A more detailed description of the characteristics of non-normal data is required.

Figure 2-5 (left). A histogram of the shoot lengths of *Banksia ericifolia* showing the close agreement with a Normal distribution with the same mean and variance.

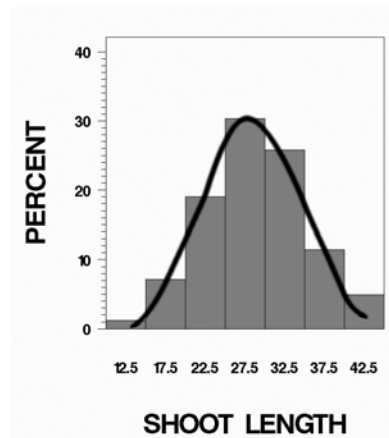
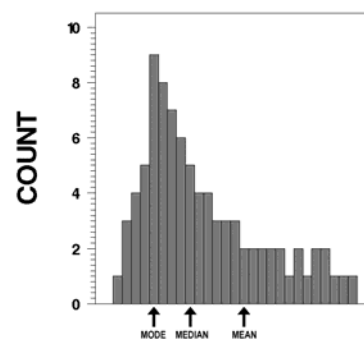


Figure 2-6 (right). A distribution that is skewed to the right showing the difference between mean, median and mode.



Substantial differences between the mode, median and arithmetic mean are apparent when a skewed distribution is considered (Figure 2-6). Clearly the three averages can have distinctly different values. Which one is the most appropriate average?

The mean is markedly affected by outlying observations whereas the median and mode are not. Consider for example the salaries of staff employed by a large company (Figure 2-6). The very high salaries of the few senior executives would shift the arithmetic mean (the centre of gravity) toward a completely unrepresentative value. On the other hand, the median would be little affected by a few large salaries and may indeed be more representative of the typical salary. Quadrupling the salaries of the ten most highly paid executives would have a marked effect on the value of the mean, but would not alter the mode and median at all.

This difference between the mean and median has important practical consequences for analysis of data that contains aberrant outlying values, because of errors at the time of measurement or during transcription in preparing the data. Such errors, if they go unnoticed, can seriously affect an analysis based on the mean and standard deviation, less so an analysis based on the median and other percentiles.

The median may be preferred over the mean in cases where it is difficult or impossible to measure the entire sample. A case in point is the time it takes a particular poison to kill half of the animals in the experiment. Such a value represents the median time to die and it is preferable to the mean which may be incalculable if some animals fail to die.

Similar considerations apply to measures of dispersion. The interquartile range or comparison of percentiles may provide the best summary of the dispersion of values in a sample taken from a very skewed population, and indeed they are preferred over standard deviations for data on river flow which are typically very skewed. In a preliminary analysis to detect outliers (aberrant values), one might choose to omit from subsequent analyses, all values that are more than four standard deviations from their mean. Only 0.01% of values would be expected to be as extreme, if the sample is drawn from a normal distribution, so values more extreme are highly suspect. The catch is that this approach is only appropriate for normal distributions. For skewed data, a definition based on percentiles is preferred. We might choose to reject, or at least check, data that are greater than the 99th percentile or less than the 1st percentile.

Most modern statistical packages perform various tests to determine if your data are likely to have been drawn from a normally distributed

population. These will be introduced when presenting the step-through examples.

Where have we come?

The objective of the Application Notes is to introduce the nuances of applying descriptive statistics in practice. In completing this lesson, you should appreciate

- The various levels at which measurements can be taken -- Nominal, Ordinal, Interval and Ratio -- and the distinction between them. You should also appreciate the central role level of measurement plays in your choice of descriptive statistic.
- The distinction between normal and non-normal data, and how this governs the descriptive statistics you use and how you report the results of your analysis.

Lesson 3: Step-through Examples

Example 2-1: Burton's Bush Rat

This is a sample analysis of nominal level data

In a study of the reproduction of the small native rodent *Melomys burtoni*, Bob Begg and his colleagues at the Conservation Commission of the Northern Territory laid out a grid of 96 trap stations at Cobourg Peninsula. A total of 143 individuals were captured over 116 weeks (Begg *et al*, 1983).

Female *Melomys* were classified into one of four reproductive groups:

- **Juveniles:** rats in this category had an imperforate vagina, indicating that they had not previously mated, their nipples were not clearly visible and they weighed less than 50 grams.
- **Non-breeding adults:** this group included all non-pregnant perforate females and imperforate adults (body weight ≥ 50 g).
- **Pregnant:** as this was scored by palpation, only females in the more advanced stages of pregnancy were scored as positive.
- **Lactating:** the nipples of rats in this category were elongated, swollen and surrounded by rings of bare skin. Milk was expressed when the nipples were squeezed. These animals had young in the nest.

The data, which are clearly at the nominal level of measurement, are held in the file MELOMYS.DAT and are arranged as follows:

```
MELOMYS APR F 2
MELOMYS APR F 2
MELOMYS APR F 3
MELOMYS APR F 3
MELOMYS APR F 2
MELOMYS APR F 2
MELOMYS APR F 2
MELOMYS APR F 2
MELOMYS APR F 1
```

etcetera.

The first variable is the species name, the second is the month in which the animal was examined, the third is the sex of the animal, and the fourth variable is its reproductive status. As reproductive status is a measurement made at the nominal level, it cannot be validly analysed in terms of means, standard deviations and other

such statistics. This is true even though the classes are represented by the digits 1 to 4 and the calculation of such statistics is manually possible. Instead, we have at our disposal, procedures that yield frequency tabulations and barcharts, from which the modal class might be determined.

Throughout this series your action is only required when you encounter instructions in one of these text boxes.

When you encounter code it will be typed in *Courier New*. Code that is in **bold** and is led by a greater than symbol ">" has to be entered by you, whereas code that appears in normal type without the greater than symbol is the output of the R console.

Start a R-session



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui).



Note

First thing to do once you started R, is to set your working directory to the folder where you place your data file. If your data is in C:\Name\R\Data\ then you will need to set the working directory to this folder.

Set your working directory to the folder where your data are located using the `setwd()` function.

```
> setwd("C:\\Name\\R\\Data\\")
```



Note

You can specify the path either using double backslash "\\" (see the code above) or using a single forward slash "/" (this is a leftover from UNIX...).

```
> setwd("C:/Name/R/Data/")
```


Prepare the data

First familiarise yourself with the data. Read it into a data frame object and peruse its form (a mixture of character and numeric data) and structure (a series of fixed field columns).

We must read in the data using the `read.table()` function.

```
> MELOMYS <- read.table("MELOMYS.DAT", header=FALSE)
```

Simply typing the object name at the R Console will reveal the contents of the object:

```
> MELOMYS
      V1 V2  V3 V4
1  MELOMYS JAN FALSE 1
2  MELOMYS JAN FALSE 1
3  MELOMYS JAN FALSE 2
.      .   .   .
.      .   .   .
.      .   .   .
257 MELOMYS DEC FALSE 4
258 MELOMYS DEC FALSE 4
259 MELOMYS DEC FALSE 4
```

Scroll upwards to check the beginning of the data set.



Allways check and confirm that your data (MELOMYS.DAT) were read in correctly. [e.g use the functions `dim()`, `summary()`, `str()`, `edit()`, `head()`, `tail()` to explore, check the help how to use them e.g. `?dim`, `?str` etc.]

Note that the data columns don't have meaningful names at this stage. The data file did not have headings for each column (that's why we used `header=FALSE`). We will use the `names()` function to give the columns required names (remember you can check the help file to see how the names function works `?names`).

```
> names(MELOMYS) <- c("SPECIES", "MONTH", "SEX", "REPCODE")
```

We can now see that the data frame contains column names:

```
> head(MELOMYS) #this shows only the first (six) entries of
the data set
  SPECIES MONTH  SEX REPCODE
1 MELOMYS  JAN FALSE      1
2 MELOMYS  JAN FALSE      1
3 MELOMYS  JAN FALSE      2
4 MELOMYS  JAN FALSE      2
5 MELOMYS  JAN FALSE      2
6 MELOMYS  JAN FALSE      2
```

```
> tail(MELOMYS) #and this the last entries
      SPECIES MONTH  SEX REPCODE
254 MELOMYS   DEC FALSE      3
255 MELOMYS   DEC FALSE      3
256 MELOMYS   DEC FALSE      4
257 MELOMYS   DEC FALSE      4
258 MELOMYS   DEC FALSE      4
259 MELOMYS   DEC FALSE      4
```

Because the reproductive status codes are 1 to 4 and are not very informative, it is also best to generate a new variable with value labels. We use REPRODUCTIVE.STATUS as a heading for this new variable in place of the rather obscure REPCODE. Be advised to use a period (.) or underscore (_) between two words, as spaces are fiddly to deal with.

```
> MELOMYS$REPRODUCTIVE.STATUS[MELOMYS$REPCODE==1] <-
"JUVENILE"
> MELOMYS$REPRODUCTIVE.STATUS[MELOMYS$REPCODE==2] <- "NON-
BREEDING ADULT"
> MELOMYS$REPRODUCTIVE.STATUS[MELOMYS$REPCODE==3] <-
"PREGNANT"
> MELOMYS$REPRODUCTIVE.STATUS[MELOMYS$REPCODE==4] <-
"LACTATING"
```

We can check that the new variable REPRODUCTIVE.STATUS has indeed been added to our data frame either by using the `names()` function, or listing the data frame again:

```
> names(MELOMYS)
[1] "SPECIES"           "MONTH"             "SEX"
[4] "REPCODE"          "REPRODUCTIVE.STATUS"
```

To save space we will leave it to you to list the data frame and check its contents (you can use the same function to check the data as suggested above).

You can use the `fix()` function to peruse the data at this point to see if it has been read as intended, and make change if needed. However, we strongly discourage the use of `fix()` and advise that you ensure your code correctly reads in the data, and give columns the appropriate labels etc. This way, you will be able to repeat your analyses much more easily as by using the `fix()` function the change of data is not recorded and you maybe have forgotten about it, when you restart your analysis the next morning. Type in the following to open the Data Editor window:

```
> fix(MELOMYS)
```

It is save to use the `edit()` function as strangely this function does not allow to change your data in the edit window. (As a matter of fact you can change the data using `edit()`, but it will not be saved.)

Close the Data Editor window when you have finished. If you are sure that the data frame is as required, attach it using the `attach()` function so that we need only type the variable name of interest (e.g. `MONTH`), rather than the data frame and variable name (e.g. `MELOMYS$MONTH`).

```
> attach(MELOMYS)
```

Frequency tabulation

Now let us analyse the data to determine the numbers of rats falling into each reproductive class. The following code will yield an appropriate frequency tabulation.

```
> table(REPRODUCTIVE.STATUS)
REPRODUCTIVE.STATUS
JUVENILE      LACTATING NON-BREEDING ADULT      PREGNANT
          49             30             133             47
```

Within the data frame `MELOMYS`, the variable `REPRODUCTIVE.STATUS` is classified as a character. We can check this using the `class()` function:

```
> class(REPRODUCTIVE.STATUS)
[1] "character"
```

Or using `summary()`.

```
> summary(REPRODUCTIVE.STATUS)
Length      Class      Mode
 259 character character
```

At this stage R does not recognize the new variable is a categorical one, it is only a number of character strings. Therefore we need to tell R that `REPRODUCTIVE.STATUS` is a factor. Furthermore when using the `table()` function, we often want data to be tabulated into classes in a set order, otherwise R will default to alphabetical as seen above. We can specify the required factor levels and their order by converting a variable to a factor using the `factor()` function. For example:

```
> MELOMYS$REPRODUCTIVE.STATUS <-
  factor(MELOMYS$REPRODUCTIVE.STATUS,
  levels=c("JUVENILE", "NON-BREEDING
  ADULT", "PREGNANT", "LACTATING"))
```

We can check that REPRODUCTIVE.STATUS has been converted to a factor as required [after first de- and reattaching the modified data frame]:

```
> detach(MELOMYS) #necessary to update the change in
> attach(MELOMYS) #REPRODUCTIVE.STATUS
> class(REPRODUCTIVE.STATUS)

[1] "factor"
```

If we now repeat the `table()` command as above, the frequency tabulation will be in the order of the factor levels that we have specified.

```
> table(REPRODUCTIVE.STATUS)

REPRODUCTIVE.STATUS
      JUVENILE NON-BREEDING ADULT      PREGNANT
LACTATING
30          49          133          47
```



Move to your text editor (e.g. Notepad) and type in the above steps. Save the program then submit it for execution.

The summarized results are shown in Table 2–6.

Table 2-6.
A frequency tabulation showing the relative representation of four reproductive classes for female *Melomys burtoni* from Cobourg Peninsula.

Reproductive status	Frequency	Per cent	Cumulative frequency	Cumulative per cent
Juvenile	49	18.9	49	18.9
Non-breeding adult	133	51.4	182	70.3
Pregnant	47	18.1	229	88.4
Lactating	30	11.6	259	100.0

Barchart

Some would prefer to see these summary data presented in graphical form. This can be done by applying the `barplot()` function to our tabulated data:

```
> barplot(table(REPRODUCTIVE.STATUS))
```

To be able to see all labels you may have to resize plotting window. Another possibility is to specify the names that should be plotted below the bars using the `names.arg` parameter of the `barplot()` function. Make sure the order of the labels is the same as the order of the levels.

```
> levels(REPRODUCTIVE.STATUS)
```

```
[1] "JUVENILE"          "NON-BREEDING ADULT" "PREGNANT"
[4] "LACTATING"

>
barplot(table(REPRODUCTIVE.STATUS), names.arg=c("Juvenile", "
NB-Adult", "Pregnant", "Lactating"))
```

Should we want the height of the bars on the barchart to represent percentages, we would divide our tabulated data by the total number of observations prior to calling the `barplot()` function. For example:

```
> percent.status <-
table(REPRODUCTIVE.STATUS) / length(REPRODUCTIVE.STATUS)
```

Check the content of `percent.status` to see what we have created. Actually the line of code above gives you a flavour how cryptic and sleek R can be, when several functions are performed in one line of code. A good idea to learn and understand how R works is to run functions separately. First type in the `length()` function,

```
> length(REPRODUCTIVE.STATUS)
[1] 259
```

which gives you the total number of observations. The function `table()` gives you four numbers, namely the number of observation for each reproductive status,

```
> table(REPRODUCTIVE.STATUS)
REPRODUCTIVE.STATUS
      JUVENILE NON-BREEDING ADULT      PREGNANT      LACTATING
           49           133           47           30
```

and some labels. The final function here was to divide the four numbers by a single number. As R is vectorized and reuses the 259 as often as it needed to do the required number of division, R divides each number of the table by the total number of observation, which is exactly what we wanted.

```
> table(REPRODUCTIVE.STATUS) / length(REPRODUCTIVE.STATUS)
REPRODUCTIVE.STATUS
      JUVENILE NON-BREEDING ADULT      PREGNANT      LACTATING
0.1891892    0.5135135    0.1814672    0.1158301
```

Finally we can make a barplot for these for numbers adjusting the labels below the bars to fit in the plotting area. You can check further arguments of `barplot()`, e.g. find out how to change the color of the bars to green.

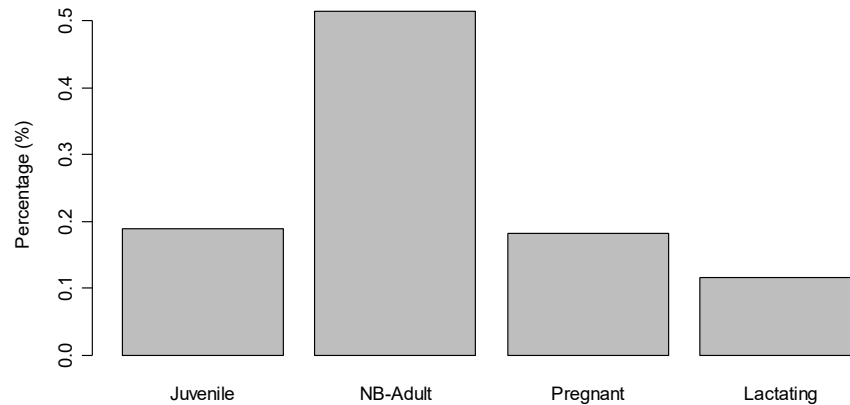
```
> barplot(percent.status, names=c("Juvenile", "NB-
Adult", "Pregnant", "Lactating"), ylab="Percentage (%)")
```



Submit the above program for execution.

The resulting barchart is shown in Figure 2-7. The modal class, non-breeding adults [NB-adult], is clearly evident.

Figure 2-7.
A barchart showing the relative percentages of female *Melomys burtoni* in each of four reproductive classes.



Barcharts with subgroups

Next it might be of interest to see if the relative proportions of animals in each reproductive class change as the year progresses, as most animals have a seasonal breeding season, at least in the temperate zone. What of the tropics? First we need to convert MONTH to a factor with levels in the order that we wish to chart the data:

```
> MONTH <- factor(MONTH, levels=
  c("JAN", "FEB", "MAR", "APR", "MAY", "JUN",
    "JUL", "AUG", "SEP", "OCT", "NOV", "DEC"))
```



Note

Be sure to use only uppercase characters for the months if that is how they are in the raw data frame. Make sure all quotes are balanced.

For a shortcut you could have used the `month.abb` object, which is a predefined object in R and has all the abbreviated names of the month stored into a vector. To complicate things again, you would have to convert all characters into upper cases using `toupper()`. You can check this by typing `month.abb` and `toupper(month.abb)` separately in the R-console.

```
> MONTH <- factor(MONTH, levels= toupper(month.abb))
```

Next we tabulate the data for each level of REPRODUCTIVE.STATUS:


```
> table(REPRODUCTIVE.STATUS, MONTH)
```

MONTH

REPRODUCTIVE.STATUS	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
JUVENILE	2	0	1	4	5	7	6	4	2
NON-BREEDING ADULT	19	8	6	9	17	11	12	9	16
PREGNANT	9	1	4	3	4	5	3	0	3
LACTATING	3	0	2	2	3	4	1	4	2

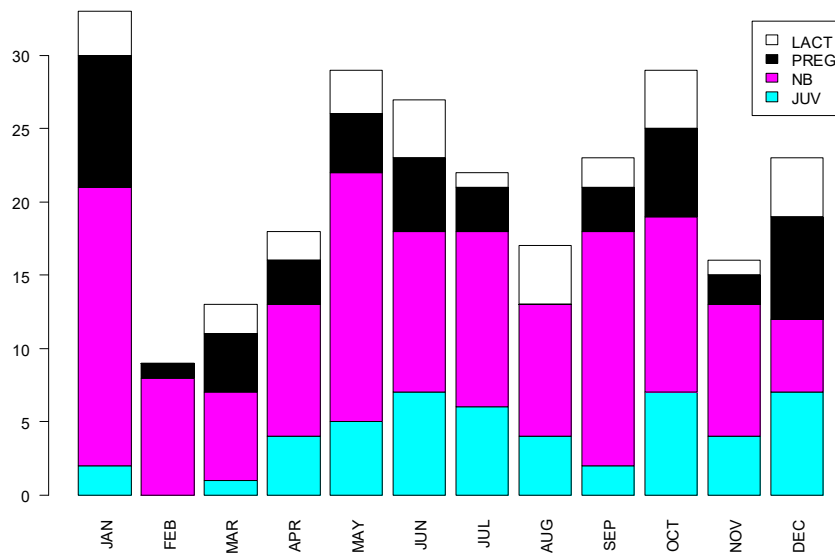
Finally, we call the `barplot()` function, supplying it with the frequencies and a range of arguments requesting bars are not plotted side-by-side, but staggered (`beside=F`), specifying bar labels are perpendicular to the axis (`las=2`), and specifying the colours to use, and the legend labels:

```
> barplot(table(REPRODUCTIVE.STATUS, MONTH), beside=F,
las=2, col=c(5,6,1,0), legend=c("JUV", "NB", "PREG", "LACT"))
```

 Submit the above program for execution. Try altering the argument "beside=F" to "beside=T"

The results are shown in Figure 2-8. With the exception of February when the sample size was very low, juvenile animals are present throughout the year. The same is true for pregnant and lactating females, so the conclusion of Begg *et al* (1983) that breeding occurs throughout the year in *Melomys burtoni* appears well supported.

Figure 2-8. Seasonal variation in the relative proportions of female *Melomys burtoni* in each of four reproductive classes (see legend).



Multiple plots

There may be differences between the wet season (November to March) and the dry season (May to September) but it is difficult to see clearly with the data presented in the form shown in Figure 2-8. Higher order graphics parameters within R are set using the `par()` function. For example, to ask for two plotting panels side by side we enter the following:

```
> par(mfrow=c(1,2)) #sets up two plotting areas [one row,
two columns]
```

This essentially asks for "one row and two columns." Then we need to specify the dry season. A possible way is to create a new variable called `season`, which is a factor and has two levels (wet and dry) in accordance to the month when the individuals were sampled.

To first plot the data of the dry season:

```
> season <- ifelse(MONTH=="MAY" | MONTH=="JUN" |
MONTH=="JUL" | MONTH=="AUG" | MONTH=="SEP" , "DRY", NA)
```

The `ifelse` command checks each entry in a vector if the condition (`MONTH=="MAY"` etc.) is true and creates the entry "DRY" in the new vector if this is the case otherwise it creates an NA. Now we are set up to first plot the data of the dry season:

```
> barplot(table(REPRODUCTIVE.STATUS[season=="DRY"]),
names=c("Juvenile", "NB-Adult", "Pregnant", "Lactating"),
las=2, ylim=c(0,80), main="Dry season")
```

Next the wet season.

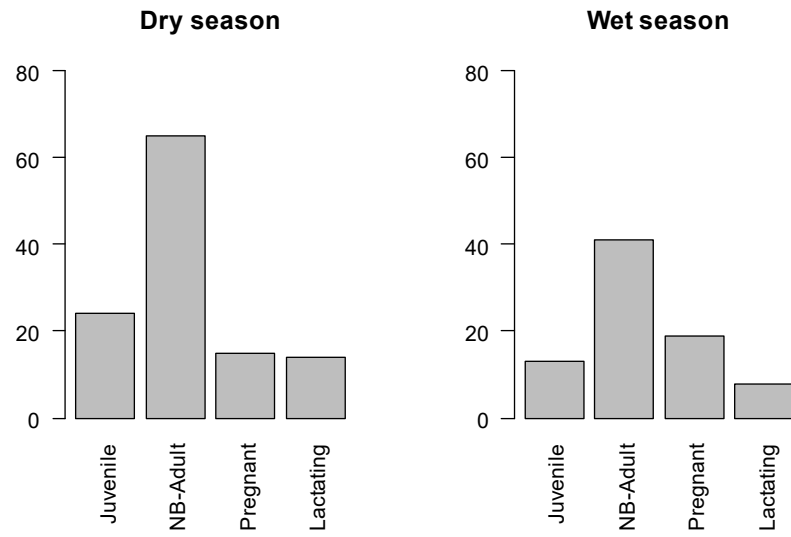
```
> season <- ifelse(MONTH=="NOV" | MONTH=="DEC" |
MONTH=="JAN" | MONTH=="FEB" , "WET", NA)
> barplot(table(REPRODUCTIVE.STATUS[season=="WET"]),
names=c("Juvenile", "NB-Adult", "Pregnant", "Lactating"),
las=2, ylim=c(0,80), main="Wet season")
```



Submit the above program for execution.

The output is as shown in Figure 2–9. The differences between the wet and dry seasons, though slight, are now more clearly evident. *Begg et al* (1983) concluded that breeding of the *Melomys burtoni* at Cobourg Peninsula occurred throughout the year, with increased incidence in the wet season. Juveniles entered the populations over an extended period, though recruitment was low.

Figure 2–9.
Relative proportions of *Melomys burtoni* in each of four reproductive classes in the wet season compared with those of the dry season.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Source

Begg, R, Walsh, B, Woerle, F. and King, S. (1993). *Ecology of Melomys Burtoni, the grassland Melomys (Rodentia : Muridae) at Cobourg Peninsula, N.T.* Australian Wildlife Research 10:259-267.

Example 2-2: Chesapeake blue crabs

This is a sample analysis of nominal level data

The Chesapeake Bay is the largest, most productive estuary in the United States of America, providing a natural habitat for more than 2,700 migratory and resident wildlife species. It supports important commercial fisheries that supply millions of kilograms of seafood annually, and year-round recreational fisheries for species such as striped bass, blue crabs and bluefish which are a multi-million dollar industry. These Chesapeake Bay resources are studied, monitored, and managed in an effort to conserve them for future generations.

The Virginia Institute of Marine Science (VIMS) conducts a Juvenile Trawl Survey project that has been an integral part in this process for over 40 years. The project began sampling in 1955 and continues in similar fashion today. The primary objective of the trawl survey is to monitor trends in seasonal distribution and abundance of juvenile fish of about twenty important finfish and invertebrates.

Currently, the survey includes waters from the mouth of the Chesapeake Bay up to the freshwater interface at the fall line of the James, York, and Rappahannock Rivers. Samples from about 60 stations are collected every month of the year by the research vessel Fish Hawk. At each station, a 30 foot wide shrimp trawl is towed for five minutes. Once on board, the catch is sorted by species, the number of fish of each species is counted, a large proportion of the fish are measured, and all are released. Each month, 20 to 50 thousand fish, crabs, and other invertebrates are processed. About 70 species are commonly caught, though a total of 223 species have been identified over the last 40 years.

We queried the VIMS online database and downloaded a subset of data on the commercially valuable blue crab *Callinectes sapidus*. The data includes information on 934 crabs, including their sex, reproductive status and carapace width in mm. We are interested in the reproductive status of the caught individuals. In which state is the majority of individuals and does this change over the course of the year? The data for sex and maturity are clearly at the nominal level of measurement.

The data are held in the file BLUECRAB.DAT and are arranged as follows:

APR	F	1	72
MAY	F	1	99
MAY	F	1	21
JUN	F	2	59
JUN	F	2	46

The first column is the month in which the sample was collected, second column is the sex of the crab (M or F), the third column is the reproductive status of the crab (1: Undetermined; 2: Immature; 3: Mature), and the fourth column is the carapace width in mm.

As reproductive status is a measurement made at the nominal level, it cannot be validly analysed in terms of means, standard deviations and other such statistics. This is true even though the classes are represented by the numerical values 1 to 3 and the calculation of such statistics is manually possible. Instead, we have at our disposal, procedures that yield frequency tabulations and bar charts, from which the modal class are to be determined.

Start a R-session



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui).

Prepare the Data

First familiarise yourself with the data by importing it and peruse its form (a mixture of character and numeric data) and structure (a series of fixed field columns).



Use the `read.table()` function to read the data file "BLUECRAB.DAT" into a data frame object in R. Do not forget to set your working directory [`setwd("c:\\...where the\\data is...")`] Peruse the data frame to check the data has been read in correctly (use functions such as `dim()`, `names()`, `head()`, `tail()`).

```
> crab <- read.table("BLUECRAB.DAT", header=TRUE)
```

Note this time the data already had headings and therefore we used the option `header=TRUE`. You can check this using the `names()` function.

Now let us analyse the data to determine the numbers of crabs falling into each reproductive class. Because the reproductive status codes are 1 to 3 and are not very informative, it is also best to give them value labels via indexing (see Workbook 1: *Getting Started*).

```

> crab$rep.status[crab$rep.status==1] <- "IMMATURE"
> crab$rep.status[crab$rep.status==2] <- "MATURE"
> crab$rep.status[crab$rep.status==3] <- "UNDETERMINED"

```

There are other possibilities to convert numbers to factors. If you are keen read about it help pages of the `?factor()` and `?label()` functions.



Submit the above commands for execution (e.g. by using the hotkey ALT-p within Tinn-R).

You can peruse the data at this point to see if it has been read as intended. You could check the number of rows in the data frame:

```

> nrow(CRAB)
[1] 1161

```

Frequency Tabulation

A complete frequency tabulation of the reproduction status consists of the counts of each level of the reproductive status, the percentage of these counts, the cumulative frequency and the cumulative percent. Think about what quantities do we need to calculate these values.

The following code will yield an appropriate frequency tabulation.

```

> attach(crab)

```

This inserts the crab object into the search path and saves us to type the data.frame name each time. Be aware from now on you work on a copy of the original data.frame! (see Workbook 1 for details).

```

> table(rep.status)
rep.status
  IMMATURE      MATURE UNDETERMINED
      388         475         298

```

This gives us the counts of each reproductive level. To calculate the percentage of each level we need to know the number of all valid observations.

```

> n.observation <- length(rep.status)

```

Another way would be to sum over all entries of the frequency table counts.

```

> n.observation.alt <- sum(table(rep.status))

```

Now we can calculate the percentages

```
> table(rep.status)/n.observation
rep.status
  IMMATURE      MATURE UNDETERMINED
0.3341947    0.4091301    0.2566753
```



Note

This only works because R **recycles** the value in n.observation and divide each value of the table object.

To calculate the cumulative frequency, we use the `cumsum()` function.

```
> cumsum(table(rep.status))
[1] 388 863 1161
```

And finally the cumulative percentages:

```
> cumsum(table(rep.status))/n.observation
  IMMATURE      MATURE UNDETERMINED
0.3341947    0.7433247    1.0000000
```



Submit the above commands for execution.

The summarised results are shown in Table 2–7.

Table 2-7. A frequency tabulation for relative representation of three reproductive classes of blue crabs *Callinectes sapidus* from Chesapeake Bay.

Reproductive status	Frequency	Percent	Cumulative frequency	Cumulative percent
IMMATURE	388	33.42	388	33.42
MATURE	475	40.91	863	74.33
UNDETERMINED	298	25.67	1161	100.00



Extra task

Try to combine the results of the frequency analysis into a single R object `freq.table` using the function `cbind()`. Lookup how the function works in the help pages typing `?rbind`.

Barchart

Some would prefer to see these summary data presented in graphical form. This can be done with the `barplot()` function.

```
> barplot(table(rep.status))
```

Should we want the height of the bars on the barchart to represent percentages, we need to divide our tabulated data by the total number of observations prior to calling the `barplot()` function.

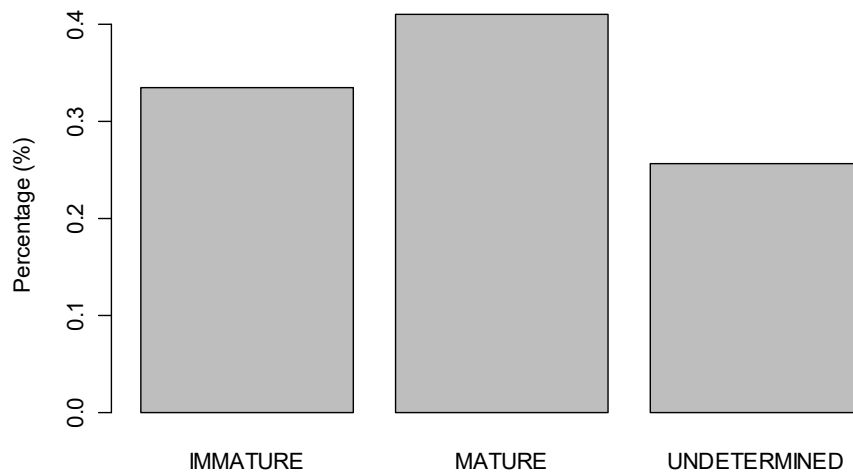
```
> percent.status <-table(rep.status)/n.observation
> barplot(percent.status, ylab="Percentage [%]")
```



Submit the above commands for execution.

The resulting bar chart is shown in Figure 2–7. The modal class, mature females, is clearly evident.

Figure 2–7.
A barchart showing the relative percentages of blue crabs in each of three reproductive classes.



Extra task

Change the colors of the bar chart to "lemonchiffon". R has 657 named colours, try `colours()`.

Enhanced cross tables

Next we are interested to see if the relative proportions of crabs in each reproductive class change as the year progresses, as most animals have a seasonal breeding season, at least in the temperate zone.

So we have a closer look at the variable `month`. Just type `month` into the console to check how it looks.

```
> month
 [1] APR APR APR APR APR APR APR APR APR APR APR APR
 [13] APR APR APR APR APR APR AUG AUG AUG AUG AUG AUG
 [25] AUG AUG AUG AUG AUG AUG AUG AUG AUG AUG AUG AUG
.....
[1129] SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP
[1141] SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP
[1153] SEP SEP SEP SEP SEP SEP SEP SEP SEP SEP
Levels: APR AUG DEC FEB JUL JUN MAY NOV OCT SEP
```

Or try

```
> table(month)
month
APR AUG DEC FEB JUL JUN MAY NOV OCT SEP
 18 265  50   6 199 119  56 119 167 162
```

So `month` seems to be a factor, but obviously no crabs were caught in some months (e.g. January and March). We still would like to have a barplot for all months of the year. Therefore we need to convert `month` to a factor with additional levels, which are also in the right order (not in the alphabetical order as this is the default of R)

```
> month <- factor(month, levels=c("JAN", "FEB", "MAR",
  "APR", "MAY", "JUN", "JUL", "AUG", "SEP", "OCT", "NOV", "DEC"))
> table(month)
```

This look better!

```
month
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
  0   6   0  18  56 119 199 265 162 167 119  50
```

Next we tabulate the data for each month and level of `rep.status`:

```
> frt.rep.month <- table(rep.status, month)
```

Check the new created object `frt.rep.month` by typing it into the console directly.

```
> frt.rep.month
```

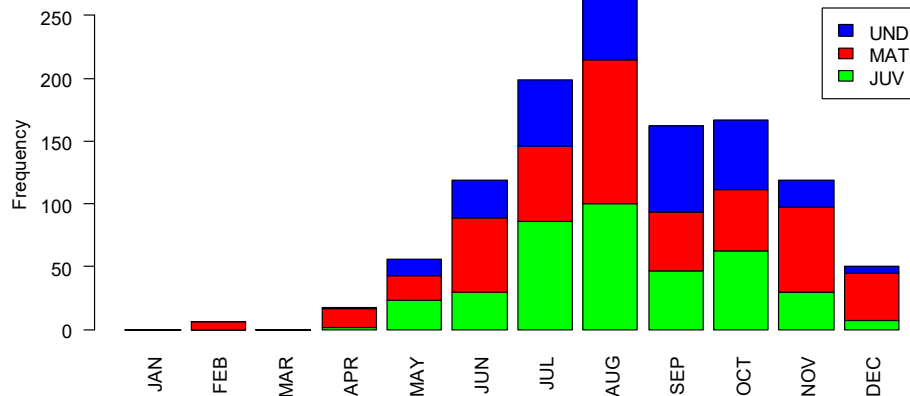
rep.status	month											
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
IMMATURE	0	0	0	2	23	30	86	100	47	63	30	7
MATURE	0	6	0	15	20	59	60	115	47	48	67	38
UNDETERMINED	0	0	0	1	13	30	53	50	68	56	22	5

Finally, we call the `barplot()` function with a range of arguments to make the plot a bit nicer. We specify bar labels are perpendicular to the axis (`las=2`), and the colours (`col=`) to use. In addition we plot a legend for the color coding.

```
> barplot(frt.rep.month, beside=F, las=2,
col=c("green","red","blue"), legend=c("JUV","MAT","UND"),
ylab="Frequency")
```

The results are shown in Figure 2-8. With the exception of January through March when the sample size was zero, immature crabs are present throughout the year.

Figure 2-8.
Seasonal variation in the relative proportions of blue crabs in each of three reproductive classes (see legend).



Visually it is difficult to decide whether the reproductive status changed between months, therefore we would like to calculate the ratio of mature to immature over months. This is a classic case for indexing our data. Let's have a look again at our table

```
> frt.rep.month
```

rep.status	month											
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
IMMATURE	0	0	0	2	23	30	86	100	47	63	30	7
MATURE	0	6	0	15	20	59	60	115	47	48	67	38
UNDETERMINED	0	0	0	1	13	30	53	50	68	56	22	5


```

      IMMATURE      0  0  0  2  23  30  86 100  47  63  30
7
      MATURE        0  6  0 15  20  59  60 115  47  48  67
38
      UNDETERMINED  0  0  0  1  13  30  53  50  68  56  22
5

```

To calculate the ratio of immature to mature over the seasons, we could calculate this for each month separately.

For April we would type

```

> 2/15
[1] 0.1333333

```

Resulting in a ratio of 0.133.

Now we could go on to do this for each month separately. A much quicker way is to use the vector capabilities of R and divide all 2x12 numbers at once. First we need to somehow access the 12 values of counts for the immature crabs.

```

> frt.rep.month[1,]
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
  0   0   0   2  23  30  86 100  47  63  30   7

```

The squared brackets is a way in R to index values in an object. [1,] means select the first row and then all the columns of this object. Instead we could have used [1,1:12] which means take all values of the first row and columns 1 to 12. Then we do the same for the mature values. Here we have to index the second row.

```

> frt.rep.month[2,]
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
  0   6   0  15  20  59  60 115  47  48  67  38

```

We get the ratio simply by dividing these 2 x 12 values.

```

> frt.rep.month[1,] / frt.rep.month[2,]
      JAN      FEB      MAR      APR      MAY      JUN
JUL      AUG      SEP      OCT      NOV      DEC
1.4333333 0.8695652 1.0000000 1.3125000 0.4477612 0.1842105

```

It is easier to read if we use the `round()` function.

```

> round( frt.rep.month[1,] / frt.rep.month[2,] ,3 )
JAN  FEB  MAR  APR  MAY  JUN  JUL  AUG  SEP  OCT
NOV  DEC
NaN   0.000 NaN   0.133 1.150 0.508 1.433 0.870 1.000 1.312
0.448 0.184

```



Submit the above commands for execution.



Extra task

Create a barplot like 2.8 that shows the counts of different sex classes and calculate the sex ratio for each month.

Regrouping nominal data

There may be differences between the winter season (November to March) and the summer season (May to September) but it is difficult to see clearly with the data presented in the form shown in Figure 2-8. To create such a graph we need to regroup our `month` variable into winter and summer season. One way to do this is to create a new variable called `season`, which is based on `month`.

```
> season <- ifelse(month=="NOV" | month=="DEC" |
  month=="JAN" | month=="FEB" | month=="MAR", "winter", NA)
```

The `ifelse` function is a convenient way to repeat an instruction for all values in a vector. So the `ifelse` statement above says literally: Go through all values of `month` and check if it is a winter month. If this is the case then write `"winter"` in the `season` vector, if not then write `"NA"` (R's default for missing value). Now we have to redo the whole thing for `"summer"`. Note that this time we write `season` as the last argument of the `ifelse` function as we do not want to override the previously generated winter season.

```
> season <- ifelse(month=="MAY" | month=="JUN" |
  month=="JUL" | month=="AUG" | month=="SEP", "summer", season)
```

A nice way to test if this was the right thing to do, we tabulate `season` against `month`.

```
> table(season, month)
```

```
      month
season JAN FEB MAR APR MAY JUN JUL  AUG  SEP  OCT NOV DEC
summer  0   0   0   0  56 119 199 265 162   0   0   0
winter  0   6   0   0   0   0   0   0   0   0 119  50
```

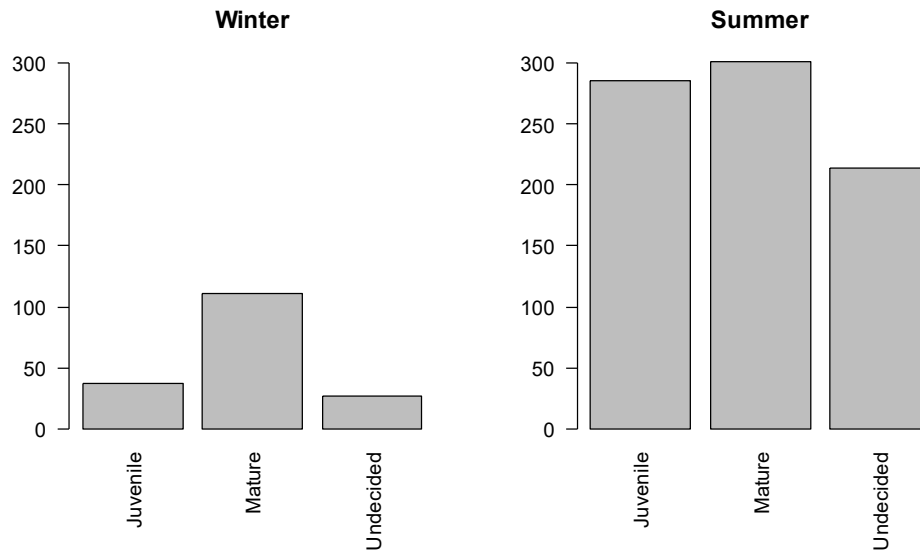
```
> barplot(table(rep.status, season), beside=T)
```



Submit the above commands for execution.

The output is as shown in Figure 2–9. The differences between the winter and summer seasons are now more clearly evident. Would you conclude that breeding occurred throughout the year, with increased incidence in the summer. Juveniles entered the populations over an extended period, though recruitment was low.

Figure 2–9.
Relative proportions of blue crabs in each of three reproductive classes in the summer compared with those of the winter.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Source

The length frequency data on blue crabs were kindly provided by the Virginia Institute of Marine Science, Juvenile Fish and Blue Crab Trawl Survey. The web-based data retrieval system appears online [<http://www.fisheries.vims.edu/vimstrawldata>]

Example 2-3: Water Quality of Lake Burley Griffin

This is a sample analysis of ratio level data, normal and non-normal.

As part of the requirements for the third year unit Special Studies in Science at the University of Canberra, Kurt Hammerschmidt, aided by staff of the Lakes Ecology Unit of the Parks and Conservation Service, collected ten replicate samples of water from each of the sites in Lake Burley Griffin. Turbidity (ntu) was measured once and total filterable phosphorus (mg/l) was measured twice for each sample. The data are held in the disk file KURT.DAT, and the form of the data is shown below.



Note

Missing values are represented by a period (.).

	site	tfp1	tfp2	turbid
SITE	01	0.045	0.052	43
SITE	01	0.053	.	28
SITE	01	0.067	0.073	43
SITE	01	0.063	0.066	28
SITE	01	0.066	0.073	42
SITE	01	0.66	0.073	42
SITE	01	0.66	0.073	42
SITE	10	0.058	0.057	19
SITE	10	0.063	0.065	16
SITE	10	0.063	0.061	15
SITE	10	0.061	0.063	11
SITE	10	0.059	0.061	15

The first field contains the site number, the second and third fields contain the two determinations of total filterable phosphorus, and the last field contains the turbidity measurements.

Kurt was interested to summarise these measurements for Lake Burley Griffin. He also wanted to learn something of the distribution of each measurement as this may influence decisions made later in analyses. Many analysis options require that the data are normally distributed.

Start a R-session



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)

Prepare the data

The first step in a R analysis is to read data from your data file folder KURT.DAT and convert it to a form suitable for use by R, in this case a data frame. At the same time we wish to combine the two determinations for total filterable phosphorus. The appropriate code would look like this:

```
> setwd("c:\\where your\\data\\is") #set the working
directory
> rt <-read.table("KURT.DAT", na.strings=".", header=T)
```

The resulting R data frame kurt should contain four variables—**site**, **tfp1**, **tfp2** and **turbid**. You can check the column headings using the `names()` function:

```
> names(KURT)
[1] "SITE" "TFP1" "TFP2" "TURBID" "TFP"
```

Please make sure you check if the data at has been read as intended (e.g. `head()`, `tail()`, `edit()`, `dim()`):

```
> head(kurt)
      site tfp1 tfp2 turbid
1 SITE-01 0.045 0.052    43
2 SITE-01 0.053    NA    28
3 SITE-01 0.047 0.053    43
4 SITE-01 0.063 0.066    28
5 SITE-01 0.066 0.073    42
6 SITE-01 0.041 0.062    43
```

Once finished we create a new variable, which is the mean between **tfp1** and **tfp2** for each measurement. To add a new variable to the **data.frame** simply invent a new name such as **tfpmean** and link it to the **data.frame** using the `$` symbol.

```
> kurt$tfpmean <- (kurt$tfp1+kurt$tfp2)/2 # Create a new
variable
attach(KURT)
```

The resulting R data frame `KURT` should now contain five variables—`site`, `tfp1`, `tfp2`, `turbid` and `tfpmean`. You can check the column headings using the `names()` function:

```
> names(kurt)
[1] "SITE"    "TFP1"    "TFP2"    "TURBID"  "TFP"
```

It is a good idea to check if the calculation of the new variable was successful. So use `head(kurt)` to check the first six entries of `tfpmean`.

```
> head(kurt)
      site tfp1 tfp2 turbid tfpmean
1 SITE-01 0.045 0.052    43  0.0485
2 SITE-01 0.053    NA    28     NA
3 SITE-01 0.047 0.053    43  0.0500
4 SITE-01 0.063 0.066    28  0.0645
5 SITE-01 0.066 0.073    42  0.0695
6 SITE-01 0.041 0.062    43  0.0515
```



Type in the above steps starting with the `setwd()` statement, and submit the program to the R Console for execution. Remember to ignore the ">" at the start of each line where it occurs.

Summary statistics

Now that the data are read in, we can use `summary()` to compute some basic descriptive statistics. Means, standard deviations, standard errors, minimums, maximums and sample sizes are of interest, and the appropriate step is as follows:

```
> summary(kurt)
```



Enter the above command to the R Console for execution.

The results of the analysis should look like those shown in Box 2–1.

Box 2–1.
Summary
statistics for
total filterable
phosphorus and
turbidity in
Lake Burley
Griffin,
Canberra,
produced by
PROC MEANS.

```
      site      tfp1      tfp2      turbid
SITE-01:10  Min.   :0.04100  Min.   :0.04200  Min.   : 9.00
SITE-02:10  1st Qu.:0.06350  1st Qu.:0.06300  1st Qu.:15.00
SITE-03:10  Median :0.06700  Median :0.06700  Median :18.00
SITE-04:10  Mean    :0.06699  Mean    :0.06756  Mean   :21.73
SITE-05:10  3rd Qu.:0.07100  3rd Qu.:0.07150  3rd Qu.:26.00
SITE-06:10  Max.    :0.08500  Max.    :0.08800  Max.   :43.00
(Other):40  NA's    :1.00000  NA's    :1.00000  NA's   : 3.00
      tfpmean
Min.   :0.04850
1st Qu.:0.06350
Median :0.06725
Mean   :0.06738
3rd Qu.:0.07037
```

```
Max. :0.08500
NA's :2.00000
```

We want to check the new variable `tfpmean` more closely. So we simply type its name to have it printed.

```
> tfpmean
```

Into the console. We get an error

```
Error: object 'tfpmean' not found
```

This is because the headings of a `data.frame` are not known to R unless we attach the object to the `search.path` using the `attach()` function so R knows where to look for it.

```
> attach(kurt)
> tfpmean
 [1] 0.0485      NA 0.0500 0.0645 0.0695 0.0515 0.0570
0.0560 0.0745 0.0765 0.0795 0.0850 0.0805
 [14] 0.0740 0.0845 0.0800 0.0825 0.0815 0.0680 0.0740
0.0720 0.0670 0.0665 0.0660 0.0665 0.0675
 [27] 0.0650 0.0640 0.0540 0.0700 0.0670 0.0675 0.0680
0.0650 0.0675      NA 0.0625 0.0605 0.0560
 [40] 0.0620 0.0690 0.0690 0.0650 0.0665 0.0625 0.0680
0.0660 0.0695 0.0665 0.0675 0.0675 0.0650
 [53] 0.0675 0.0705 0.0740 0.0640 0.0695 0.0685 0.0710
0.0680 0.0750 0.0795 0.0765 0.0765 0.0720
 [66] 0.0745 0.0765 0.0780 0.0755 0.0750 0.0655 0.0625
0.0640 0.0685 0.0660 0.0695 0.0680 0.0675
 [79] 0.0695 0.0645 0.0635 0.0640 0.0625 0.0685 0.0645
0.0595 0.0620 0.0635 0.0645 0.0690 0.0595
 [92] 0.0625 0.0630 0.0600 0.0600 0.0575 0.0640 0.0620
0.0620 0.0600
```

Now we can have a closer look and a more detailed analysis on just the variable `tfpmean`. Before we do any further analysis, we can check for missing cases by summing how many of the entries the `complete.cases()` function returns a `FALSE` result when the variable `TFP` is passed to it:

```
> sum(complete.cases(tfpmean)==FALSE)
 [1] 2
```



Note

The above example is a good example how several R functions can be combined elegantly. So let's do this step by step from the inner function to the outer functions.

Type

```
> complete.cases(tfpmean)
```

this gives you a vector with `TRUE` and `FALSE` entries for each entry in `tfpmean`.

So everywhere where is a complete case we get a TRUE otherwise a FALSE. Now we could sum all TRUE cases as TRUE is internally coded as 1 and FALSE as 0.

```
> sum(complete.cases(tfpmean)) # results in 98
```

As we are interested in the number of FALSE cases we need to have the opposite vector so each TRUE should be a FALSE and vice versa. This is achieved by comparing the vector by FALSE.

```
> complete.cases(tfpmean) ==FALSE
```

And sum these TRUEs up.

```
> sum(complete.cases(tfpmean) ==FALSE) #results is 2
```



Extra task

To check these case we could use the which function. Write the following and try to understand each step by decomposing the command from the inner to the outer function as done above.

```
>kurt[ which( complete.cases(tfpmean)==FALSE ) ]
```

Hence we see that tfpmean has 2 missing cases. For small datasets such as the one we are analyzing here, we could have seen these missing observations by inspecting the data visually, however for much larger datasets this will not be possible.

It is easy to calculate some basic statistical measures such as the mean, median, sum, standard deviation, variance and range. As the variable contains missing values, we need to specify the argument for removing the missing values when calculating the desired statistics:

```
> mean(tfpmean,na.rm=T)
```

```
[1] 0.06738265
```

```
> median(tfpmean,na.rm=T)
```

```
[1] 0.06725
```

```
> sum(tfpmean,na.rm=T)
```

```
[1] 6.6035
```

```
> sd(tfpmean,na.rm=T)
```

```
[1] 0.007125098
```

```
> var(tfpmean,na.rm=T)
```

```
[1] 5.076702e-05
```

```
> range(tfpmean,na.rm=T)
```

```
[1] 0.0485 0.0850
```


Finding the mode using R is a little tricky. First we tabulate the data using the function `table()`

```
>table(tfpmean)

0.0485  0.05 0.0515  0.054  0.056  0.057 0.0575 0.0595
0.06
      1      1      1      1      2      1      1      2
3
0.0605 0.062 0.0625  0.063 0.0635  0.064 0.0645  0.065
0.0655
      1      4      5      1      2      5      4      4
1
 0.066 0.0665  0.067 0.0675  0.068 0.0685  0.069 0.0695
0.07
      3      4      2      7      5      3      3      5
1
0.0705 0.071  0.072  0.074 0.0745  0.075 0.0755 0.0765
0.078
      1      1      2      3      2      2      1      4
1
0.0795  0.08 0.0805 0.0815 0.0825 0.0845  0.085
      2      1      1      1      1      1      1

```

, by eye we see that the mode is at 0.0675 (this value occurs 7 times) We can use `which.max()` to find where the maximum number of observations are:

```
> which.max(table(tfpmean))

0.0675
      22

```

This tells us that the mode is 0.0675, which occurs at the 22nd observation in the table. We can confirm this by checking the result of using `table(TFP)` and checking how many observations occurred:

```
> table(tfpmean)[22]

0.0675
      7

```

Now we move on to calculating some other common statistics that are useful for describing data. The coefficient of variation is found by dividing the standard deviation by the mean:

```
> sd(tfpmean,na.rm=T)/mean(tfpmean,na.rm=T)

[1] 0.1057408

```

The uncorrected sum of squares (Uncorrected SS) is the sum of the squared values for TFP:

```
> sum(tfpmean[is.na(tfpmean)==FALSE]^2)

[1] 0.4498858

```

The corrected sum of squares (Corrected SS) is the sum of the squared deviations of each value from the mean value for `tfpmean`.

```
> sum((tfpmean[is.na(tfpmean)==FALSE]-
mean(tfpmean,na.rm=T))^2)
[1] 0.004924401
```

The standard error of the mean can be calculated as the standard deviation divided by the square root of the sample size (in this case 98):

```
> sd(tfpmean,na.rm=T)/sqrt(98)
[1] 0.0007197435
```

The quartiles can be found as the default output from the `quantile()` function:

```
> quantile(TFP,na.rm=T)
      0%      25%      50%      75%     100%
0.048500 0.063500 0.067250 0.070375 0.085000
```

We can undertake some standard tests such as the *t*-test and Wilcoxon sign-rank test to compare the mean of `tfpmean` with a chosen value, say zero.

```
> t.test(tfpmean, mu=0)

      One Sample t-test

data:  tfpmean
t = 93.6204, df = 97, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.06595416 0.06881114
sample estimates:
mean of x
0.06738265
```

```
> wilcox.test(tfpmean, mu=0)

      Wilcoxon signed rank test with continuity
correction

data:  tfpmean
V = 4851, p-value < 2.2e-16
alternative hypothesis: true mu is not equal to 0
```

The Shapiro-Wilk's Test is the test for normality recommended in this course.

```
> shapiro.test(tfpmean)
```

Shapiro-Wilk normality test

```
data: TFP
W = 0.9777, p-value = 0.09475
```

There is no significant deviation from normality (Shapiro-Wilk $W = 0.9777$ $Pr < W = 0.09475$).

Should we want particular quantile values, we achieve this by telling the `quantile()` function exactly which probabilities we require (you may check the help page for this function `?quantile`). For example:

```
> quantile(tfpmean,
probs=c(0,0.01,0.10,0.5,0.9,0.99,1),na.rm=T)
      0%      1%      10%      50%      90%      99%      100%
0.048500 0.049955 0.059850 0.067250 0.076500 0.084515 0.085000
```

The mode is not very useful as the data are not grouped. Percentiles are useful for defining extreme events, for example if phosphorus is implicated in algal blooms, the water authorities might wish to be notified if the total filterable phosphorus exceeds the 90th percentile, in this case 0.0765 mg/l.

To calculate the interquartile range (the difference between the 25th and 75th percentile), we calculate the difference between the 25th and 75th percentile:

```
> quantile(tfpmean,0.75,na.rm=T)-quantile(tfpmean,
0.25,na.rm=T)
      75%
0.006875
```

To find extreme observations, say the 5 lowest and 5 highest, we can sort the variable then index the required observations:

```
> sort(tfpmean)[1:5]
[1] 0.0485 0.0500 0.0515 0.0540 0.0560

> sort(tfpmean)[93:98]
[1] 0.0800 0.0805 0.0815 0.0825 0.0845 0.0850
```

To find where these observations can be found in the data set, we use `sort.list()`.

```
> sort.list(tfpmean)[1:5]
[1] 1 3 6 29 39

> sort.list(tfpmean)[93:98]
[1] 16 13 18 17 15 12
```

To actually see all the data corresponding to these observations, we index the dataframe for the required rows. First the five smallest observations:

```
> kurt[sort.list(tfpmean)[1:5],]
      site  tfp1  tfp2 turbid tfpmean
1 SITE-01 0.045 0.052    43  0.0485
3 SITE-01 0.047 0.053    43  0.0500
6 SITE-01 0.041 0.062    43  0.0515
29 SITE-03 0.054 0.054    15  0.0540
39 SITE-04 0.053 0.059    NA  0.0560
```

Next the five biggest observations:

```
> kurt[sort.list(tfpmean, decreasing=TRUE)[1:5],]
      site  tfp1  tfp2 turbid tfpmean
12 SITE-02 0.082 0.088    28  0.0850
15 SITE-02 0.085 0.084    25  0.0845
17 SITE-02 0.085 0.080    26  0.0825
18 SITE-02 0.083 0.080    25  0.0815
13 SITE-02 0.079 0.082    28  0.0805
```

Quick plots for examining raw data include the stem and leaf plot, boxplot and histogram. First the "Stem and Leaf Plot", which is effectively a histogram on its side.

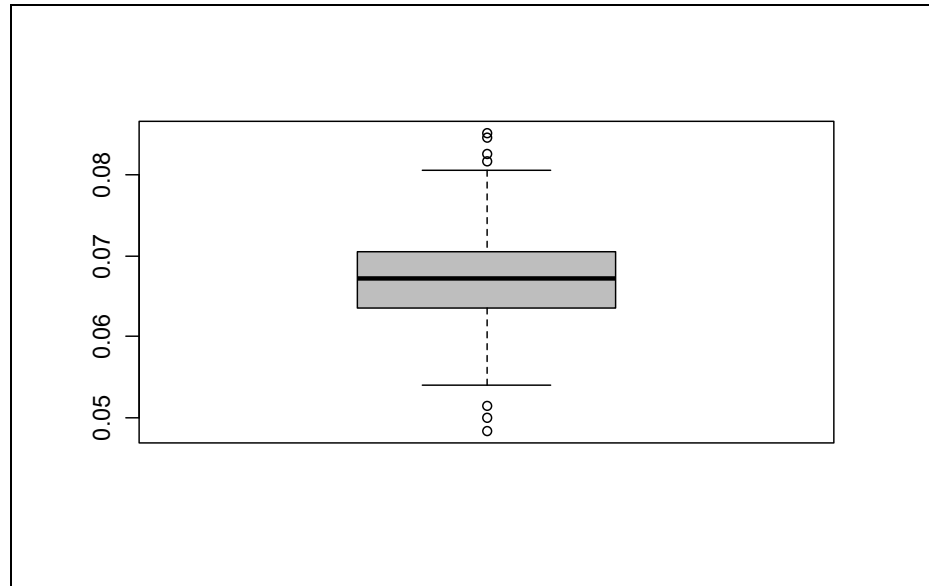
```
> stem(tfpmean)
The decimal point is 2 digit(s) to the left of the |
4 | 9
5 | 024
5 | 667899
6 | 0000222233333344444444
6 | 5555555566667777778888888888999999
7 | 0000001122444444
7 | 55566668
8 | 000123
8 | 55
```

The scale for the measurement variable needs to be multiplied by 10^{-2} (0.01) to be expressed in mg/l. The various digits used to make up the bars of the histogram indicate the value of an extra decimal place. For example, 2 of the 6 measurements in the column labeled "5" were 0.059 mg/l. Next we try a boxplot:

```
> boxplot(tfpmean, col="grey")
```

The resulting plot is shown in Figure 2-9. The bottom and top edges of the box show the 25th and 75th percentiles; the horizontal line shows the median. You can check the help for this function (`?boxplot`) to find out that the whiskers extend from the box are 1.5 times the inter-quartile range whichever is the lesser. It can be seen that the distribution is quite nicely symmetrical and probably normally distributed.

Figure 2–9.
Boxplot of variable
TFP obtained using
the `boxplot()`
function..

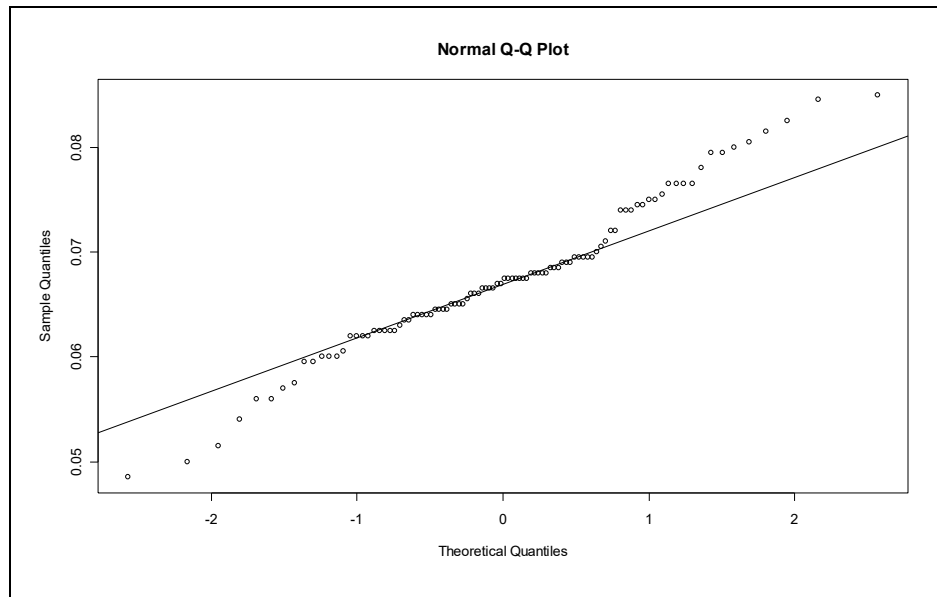


To check for normality (as suggested by the Shapiro-Wilk's Test) we can create a "Normal Probability Plot" using the `qqnorm()` function. This is similar in application to a plot of cumulative relative frequencies on probability paper. Normally distributed data appear as a straight line, whereas deviation from linearity indicates deviation from normality. We can plot this straight line using the `qqline()` function. For the total filterable phosphorous:

```
> qqnorm(tfpmean)
> qqline(tfpmean)
```

The results are shown in (Figure 2-10). There is only marginal evidence of deviation from a normal distribution in this plot. It can be seen that outer values left of the mean are a bit underrepresented and outer values right from the mean a bit overrepresented.

Figure 2-10.
Normal
probability plot
for total
filterable
phosphorus
from Lake
Burley Griffin,
Canberra.
Obtained using
the `qqnorm()`
function.



Graphical Presentation

We might choose at this point to draw upon the superior graphical capabilities of the function `hist()`, to produce a histogram of the `tfpmean` measurements for a report.

```
> hist(tfpmean, breaks=seq(0.04,0.09,0.005), xlab="tfpmean  
Midpoint", col="grey")
```



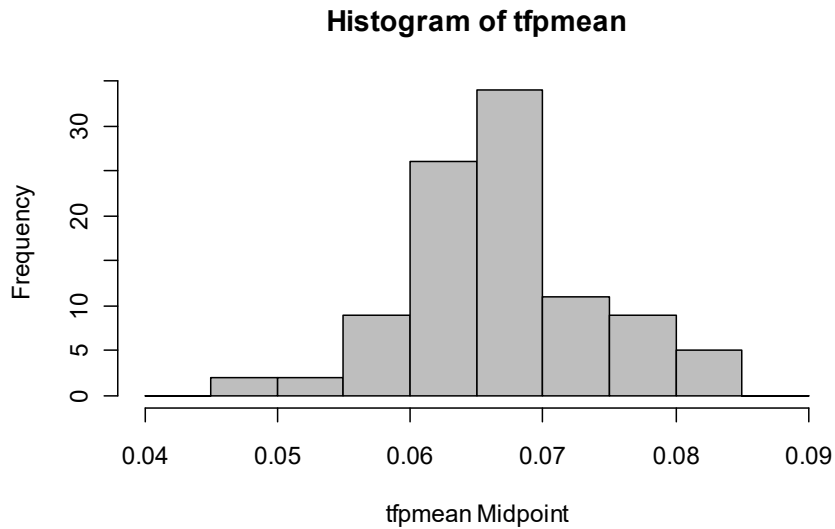
Note

To specify the breakpoints (R can do it automatically) we use the `seq()` function, so check what this function does.



Submit the above program for execution.

Figure 2–11.
Distribution of
measurements
of total filterable
phosphorus in
Lake Burley
Griffin. The data
are normally
distributed.



Report the results

The resulting histogram (Figure 2–11) is nicely bell-shaped, consistent with normal distribution. But what does this all mean for total filterable phosphorus? There is no evidence for suspecting that the data are not from a normally distributed population. The format of the following summary is appropriate for describing data that is normally distributed, and should be carefully followed in reports.

"Total Filterable Phosphorus in Lake Burley Griffin ranged from 48.5 $\mu\text{g/l}$ to 85.0 $\mu\text{g/l}$ (Mean 67.4 ± 0.72 , $n=98$) during the period of study. The variable was normally distributed (Shapiro Wilk Statistic=0.98, $p=0.09$, ns; visual examination of probability plot, Figure 2-11). The mean plus 3 standard deviations, useful as a definition of an extreme event, was 88.8 $\mu\text{g/l}$."

The mean is presented with its standard error for reasons that will become clear when the topic of statistical inference is covered in Workbook 3. Inclusion of the histogram shown in Figure 2–11 is optional, depending upon the emphasis you wish to place on the frequency distribution of TFP measurements in your report.

Note that no mention is made of the mode, median, quartiles or inter-quartile range. These statistics add no additional information for normally distributed data.

For normally distributed data, an equally useful alternative definition of extreme events could be made in terms of standard deviations from the mean. Less than 1% of values would be expected to lie outside three standard deviations from the mean.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Non-Normal data

Consider the output from a similar analysis for turbidity (turbid) taken from the previous example.

```
> summary(turbid)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
   9.00  15.00   18.00   21.73  26.00   43.00    3.00

> quantile(turbid,probs=c(0.05,0.25,0.5,0.75,0.95),na.rm=T)
   5%  25%  50%  75%  95%
12.8 15.0 18.0 26.0 42.0
```



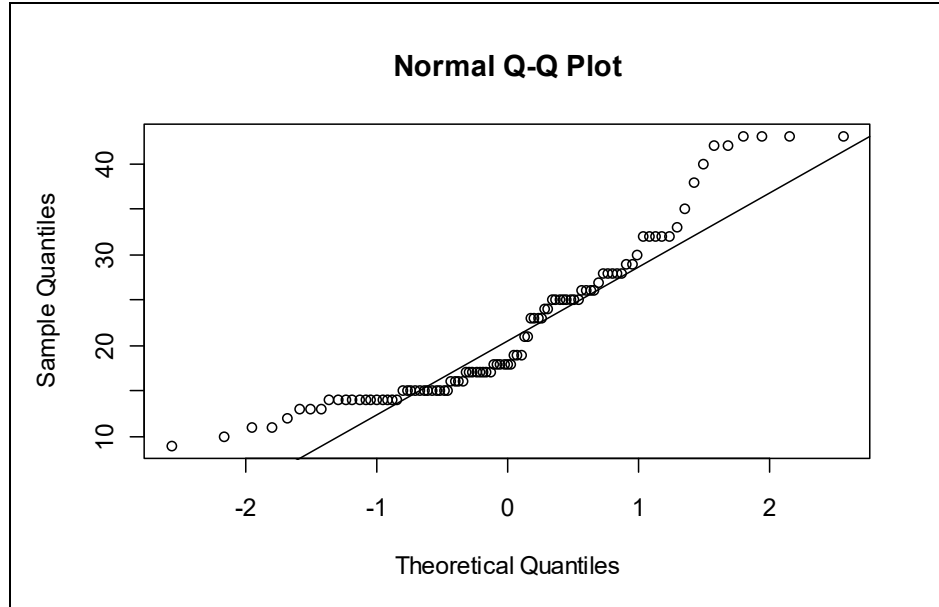
Submit the above programs for execution.

The main qualitative difference between the results for turbidity and those for total filterable phosphorus is that all indications suggest that turbidity is not normally distributed. We examine this as before:

```
> qqnorm(turbid)
> qqline(turbid)
```

The probability plot (Figure 2-12) reveals obvious deviations from linearity.

Figure 2-12.
Tests of
normality for
turbidity in Lake
Burley Griffin,
Canberra.



The Shapiro-Wilk's test was significant ($W = 0.89$, $p < 0.0001$).

```
> shapiro.test(turbid)
```

```
Shapiro-Wilk normality test
```

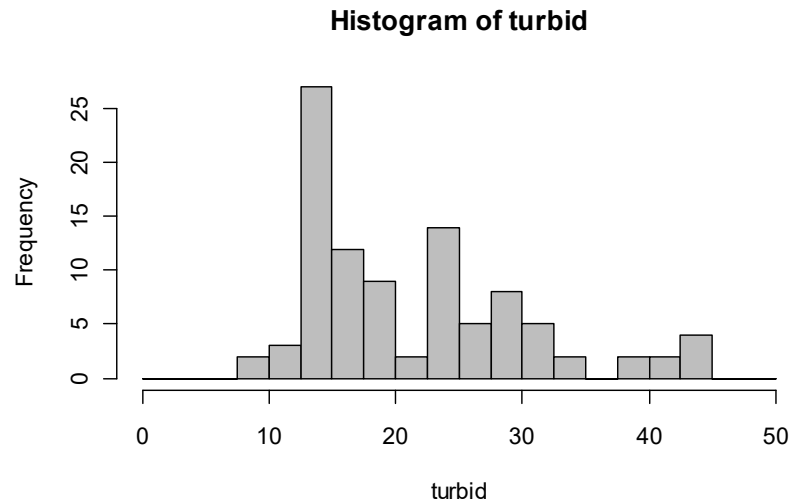
```
data:  TURBID  
W = 0.8926, p-value = 8.9e-07
```

We draw a histogram as before:

```
> hist(turbid, breaks=seq(0,50,2.5), col="grey")
```

Perusal of the histogram (Figure 2–13) reveals a bi-modal distribution (a primary mode and a secondary mode) with clear deviations from normality.

Figure 2–13.
Distribution of
measurements
of turbidity in
Lake Burley
Griffin. The
data are not
normally
distributed.



Report the results

A verbal summary of these non-normal data is more complicated than for the normally distributed TFP:

"Turbidity in Lake Burley Griffin ranged from 9 to 43 ntu (Mean 21.7 ± 0.88 , $n = 97$) during the period of the study. The variable was not normally distributed (Shapiro-Wilk Statistic = 0.89, $p < 0.0001$; visual examination of probability plot, Figure 2–11). The frequency distribution was bi-modal (highest mode at 14 ntu and skewed to the right, with a median of 18 and an inter-quartile range of 11. The 95th percentile, useful as a definition of an extreme event, was 42 ntu."

Again, you should be careful to follow this format for the description of non-normal data in your reports.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Example 2-4: Blue crab sizes

This is a sample analysis of ratio-level data, both normal and non-normal.

The Virginia Institute of Marine Science (VIMS) has conducted a routine trawl survey of Chesapeake Bay for over 40 years. The project began sampling in 1955 and continues in similar fashion today. The primary objective of the trawl survey is to monitor trends in seasonal distribution and abundance of juvenile fish of about twenty important finfish and invertebrates.

The data below were taken from 934 crabs, and includes a measurement of carapace width in mm and a variable containing sex and maturity status as follows:

JF: Juvenile Female

JM: Juvenile Male

FF: Mature Female

MM: Mature Male

The data are held in the disk file CRABLEN.DAT, with two columns of data, the first being MATURITY and the second being CARAPACE WIDTH.

We are interested to summarise these measurements for the blue Crab. We also want to learn something of the distribution of crab sizes as this may influence decisions made later in analyses. Many analysis options require that the data are normally distributed.

Start a R session



Double click on the Tinn-R icon and launch R from within Tinn-R (Click in the Menu on R->Initiate/Close Rgui->Initiate preferred Rgui)

Prepare the data



Note

Make sure you set your working directory using the `setwd("here is my//data")` function.

The first step is to read data from the disk. The name of the data file is "CRABLEN.DAT". The appropriate step would look like this:

```
> crab <- read.table("CRABLEN.DAT", header=FALSE,
na.strings=".")

> head(crab)
  V1 V2
1 JF 13
2 JF 14
3 JF 16
4 JF 17
5 JF 18
6 JF 18
```

This time there are no headings on top of the data columns, so R has generated `v1` and `v2` for us, which is not really informative. Therefore we need to change the headings using the `names()` function.

```
> names(crab) <- c("maturity","carawidth")
```

The resulting `data.frame` `crab` should now contain two variables—maturity and carawidth. Please check the data at this point to see if it has been read as intended (e.g. `head()`, `tail()`, `summary()`, `dim()`).

To see if there is any missing data, this time we use the `is.na()` function. Obviously we could also use the `complete.cases()` function as in the previous example.

```
> sum(is.na(crab))
[1] 0
```

Then we attach the `crab` `data.frame` to avoid retyping `crab` all the time.

```
> attach(crab)
```

Summary Statistics

Now that the data are read in, we can use `summary()` to compute some basic descriptive statistics. Means, standard deviations, standard errors, minimums, maximums and sample sizes are of interest, and the appropriate step is as follows:

```
> summary(carawidth)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
 11.00  56.00   73.00   82.26 114.00  165.00
```

To calculate the standard deviation we need to use the `sd()` function.

```
> sd(carawidth)
```

```
[1] 35.51195
```

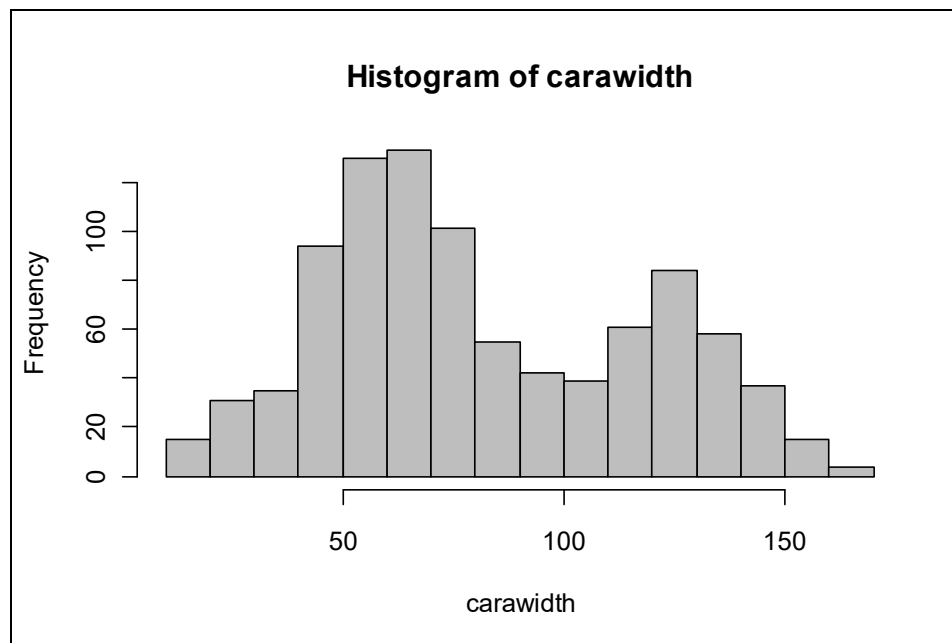


Submit the above commands for execution.

Now consider a more detailed analysis on just the variable CW. From the output above we see that the mean and median are quite different, which is a first indicator that the data are not normally distributed. To check this we first visualize our data by making a histogram (Figure 2-14).

```
> hist(carawidth, col="grey")
```

Figure 2-14. Distribution of measurements of turbidity in Lake Burley Griffin. The data are not normally distributed.



The histogram is not bell-shaped, as it would be if it were a normal distribution, but rather, is bimodal. Note that the median and mean shown on the box plot are not coincident, and that the interquartile range represented by the box is not symmetric about the mean. This indicates a skew to the right in this case.

As mentioned before the "Stem Leaf Plot" is effectively a histogram on its side. The scale for the measurement variable is expressed in mm, and the raw frequencies are given in the column headed "I" (Figure 2-15).

Figure 2-15.
Stem leaf plot
for carapace
widths of Blue
Crabs from
Chesapeake
Bay, USA.

```

The decimal point is 1 digit(s) to the right of the |
1 | 134667778899
2 | 00111111112222333333555677899999
3 | 0122233444555666777788888899
4 | 00000011111112222222223333333333334444445555666666667777778888+6
5 | 0000000000000011111222222222223333333333334444444455555555556666+48
6 | 0000000000000000111111111111222222222233333333333333333344444444+62
7 | 00000001111111111122222222233333333333334444444444445555555556+19
8 | 000000000111222222333334444444455556666677778888899999
9 | 000000111123333444455555666677778889999999
10 | 00011112233333334444555677778889999
11 | 00000122222222233333444444455555666677779999999
12 | 0000000001112222222233333333344444444445555556666677777+8
13 | 0000001111112222333444445555555666677777788888999999
14 | 000012222233334444455566667778889999999
15 | 012223456666799
16 | 1135

```

We clearly see that the distribution is bi-modal. A test on normality should reveal a significant deviation from normality.

```

> shapiro.test(carawidth)

Shapiro-Wilk normality test

data:  carawidth
W = 0.9549, p-value = 2.406e-16

```

The Shapiro-Wilks Test, (which is the one recommended in this course) indicate a strong evidence of deviation from normality (Shapiro-Wilk $W = 0.95$ [Pr < W] < 0.0001).

We further test this by creating a probability plot (Figure 2-16).

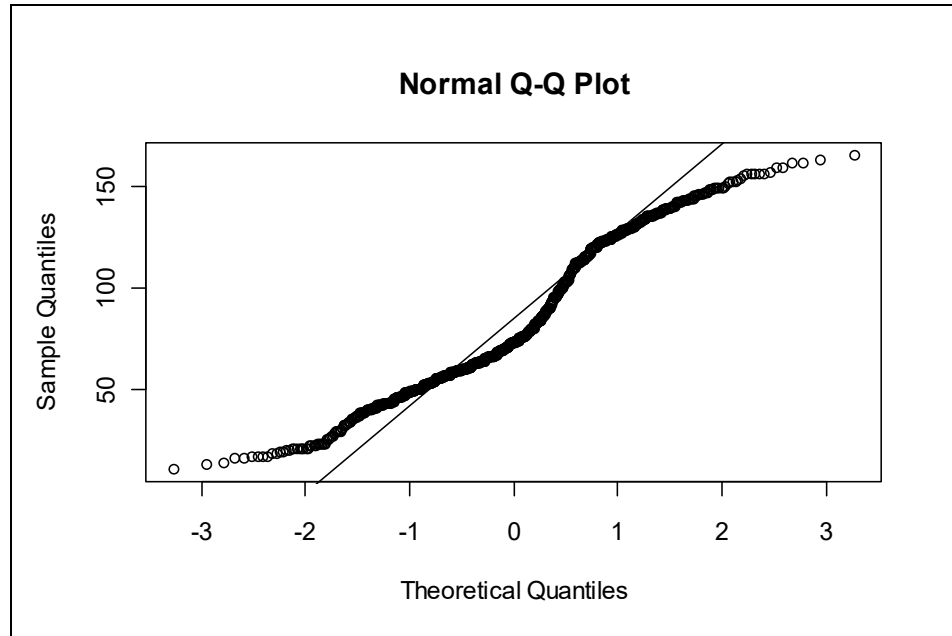
```

> qqnorm(carawidth)
> qqline(carawidth)

```

The "Normal Probability Plot" (Box 2–2G) is similar in application to a plot of cumulative relative frequencies on probability paper. Normally distributed data appear as a straight line, whereas deviation from linearity indicates deviation from normality. The data values are represented by *, whereas the + symbols define a reference straight line for comparison. The asterisks (*) depart from the linear trend (+), which is evidence of deviation from a normal distribution.

Figure 2–16.
Tests of normality based on a probability plot of the distribution of carapace widths for blue crabs in Chesapeake Bay.



 Submit the above program for execution.

Again we see quite a deviation from normality. So we need other non parametric descriptions such as quantiles and extreme values to describe this distribution.

```
> quantile(carawidth, probs =
c(0,0.01,0.05,0.1,0.25,0.5,0.75,0.9,0.95,0.99,1))
 0%   1%   5%   10%  25%  50%  75%  90%  95%
11.00 18.00 31.65 42.00 56.00 73.00 114.00 135.00 143.00

 99%  100%
156.00 165.00
```

The output contains the maximum (165), minimum (11) range, median, quartiles and various percentiles. The mode would not be useful as the data are not grouped. Percentiles are useful for defining extreme events or extreme individual sizes. A crab might be regarded as of exceptional size, for example, if its carapace width exceeds the 95th percentile, in this case 143 mm.

To identify outliers we look for the 5 lowest and highest values.

```
> sort(carawidth)[1:5] # 5 lowest values
[1] 11 13 14 16 16

> sort(carawidth, decreasing=TRUE)[1:5] # 5 highest values
[1] 165 163 161 161 159
```

This information can be very useful in identifying aberrant outliers.

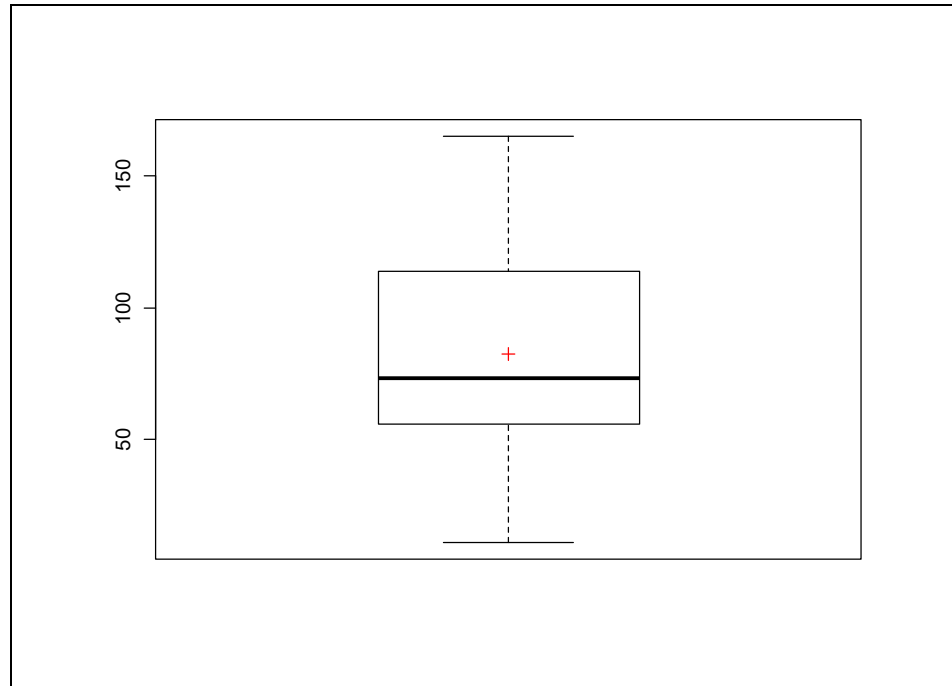
Finally we create a boxplot (Figure 2-17). This boxplot is another way to identify outliers. These would be plotted as open circles above the whiskers. Here we do not see outliers, but be careful as we are not dealing with unimodal data, the search for outliers is difficult anyway.

```
> boxplot(carawidth)
```

To make this plot a bit more informative we include the mean as a red cross.

```
> points(1,mean(carawidth), pch=3,col="red")
```

Figure 2-17.
Tests of
normality based
on a probability
plot of the
distribution of
carapace widths
for blue crabs in
Chesapeake
Bay.



The box plot is drawn on the basis of the scale on the histogram. The bottom and top edges of the box show the 25th and 75th percentiles; the horizontal line shows the median. The vertical line extends from the box as far as the range of the data or 1.5 times the inter-quartile range whichever is the lesser. Values more extreme than 1.5 inter-quartile ranges from the box are shown as 0 if they are less extreme than 3 inter-quartile ranges and as * otherwise.

Thus we have several lines of evidence to suggest that the size distribution of blue crabs is not normal. The Shapiro-Wilkes test demonstrated significant deviation from normality. This was evident in the stem-leaf plot, which showed the distribution to be bimodal and slightly skewed to the right. The mean and median did not coincide in the box diagram, and the probability plot showed considerable systematic deviation from linearity.



Submit the above program for execution.

Report the results

But what does this all mean for carapace widths of blue crabs? There is strong evidence for suspecting that the data are not from a normally distributed population. The format of the following summary is appropriate for describing data that is not normally distributed, and should be carefully followed in reports.

"Blue crabs in Chesapeake Bay ranged from 11 to 165 mm (Mean 82.3 ± 1.16 , $n = 934$) during the period of study. The distribution of crab sizes was not normally distributed, (Shapiro-Wilk Statistic = 0.95, $p < 0.0001$; visual examination of probability plot, Figure 2–9). The frequency distribution was bi-modal (highest mode at 65 mm) and skewed to the right, with a median of 73 mm and an inter-quartile range of 58 mm. A definition of a particularly large crab could be based on the 95th percentile of 143 mm."

The mean is presented with its standard error for reasons that will become clear when the topic of statistical inference is covered in Module 3. Inclusion of the histogram shown in Figure 2–14 is optional, depending upon the emphasis you wish to place on the frequency distribution of carapace width measurements in your report.

Note that the mean and standard deviation are insufficient to describe the data when it is drawn from a non-normal distribution. For this reason, we need to include the mode and median (which might differ from the mean), and selected percentiles. These statistics add additional information for non-normally distributed data beyond what is conveyed by the mean and standard deviation.

For non-normally distributed data, a useful definition of extreme events or exceptionally large individuals should be made in terms of an appropriate percentile, not in terms of standard deviations from the mean.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Normal data

We might now ask why the size distribution of crabs is not “Normal”. Recall that normality arises when a myriad of small influences come to bear upon the value taken by a variable. In this case however,

there are two dominating influential factors – sex and age. Why would we expect normality when the two sexes might differ in size, leading to bimodality, or if age has a dominant influence on size and cohorts differ in their contribution to the dataset.

We might expect crab sizes to be normal for a given sex and age, so let's have a look at the size distribution of mature female crabs only.

There are different ways of selecting only mature females. One is to define a new subset of the data that contains only mature female crabs using the `subset()` function.

```
> crab.ff <- subset(crab, maturity=="FF")
```

Now we need to calculate the same statistics as above. For convenience we include now `crab.ff` to the search path. Note before we do this we have to detach the `crab` data.frame because otherwise our new data will be masked by this data.frame, because their headings have the same name.

```
> detach(crab)
> attach(crab.ff)
> summary(crab.ff)

maturity   carawidth
FF:129    Min.      :107.0
JF:  0     1st Qu.:125.0
JM:  0     Median :135.0
MM:  0     Mean    :134.9
                3rd Qu.:144.0
                Max.    :163.0
```

As you see there are no other cases than FF in the maturity variable left. The mean and mode are similar, which is a first indication for normally distributed data. We calculate the mode:

```
>which.max(table(carawidth)) #mode
135
27
```

This should be read: The mode is 135, which is again similar to the mean and median. This value appears 27 times in the dataset. Check this by typing `table(carawidth)` into the R console.

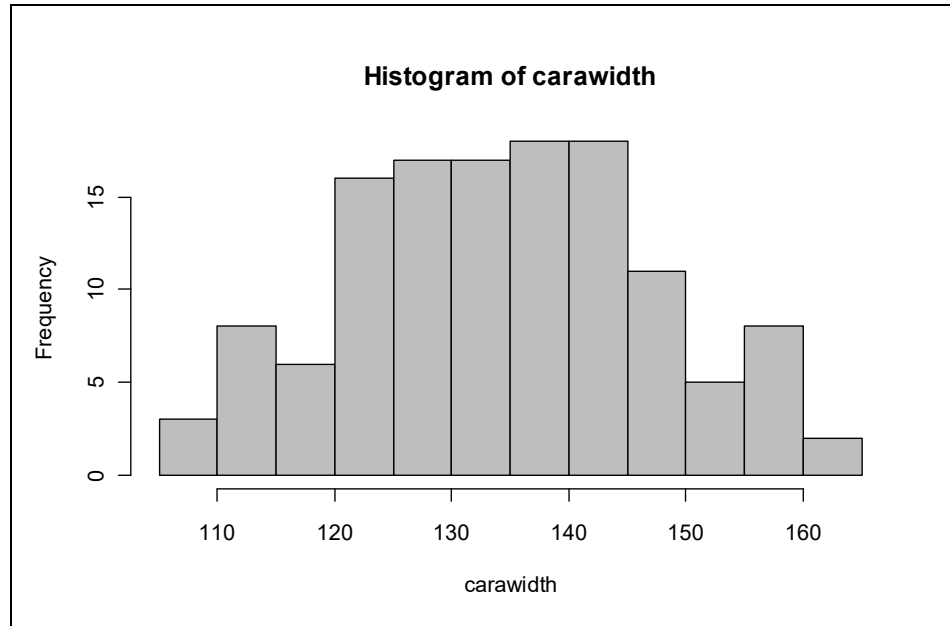
Now we plot the histogram of the new data (Figure 2-18) and also test for normality using the Shapiro-Wilk's Test.

```
> hist(carawidth, col="grey")
> shapiro.test(carawidth)

Shapiro-Wilk normality test
```

```
data: carawidth
W = 0.9889, p-value = 0.3865
```

Figure 2-18.
Tests of normality based on a probability plot of the distribution of carapace widths for blue crabs in Chesapeake Bay.



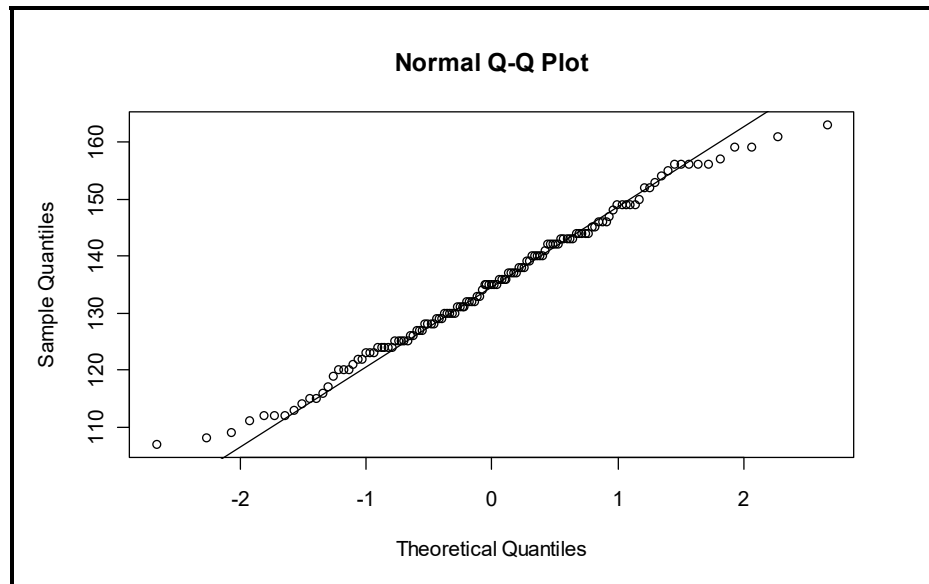
There is no indication from the Shapiro-Wilks Test of significant deviation from normality (Shapiro-Wilk $W = 0.99$ [$\text{Pr} < W$] = 0.39) (Box 2-3A).

The histogram is bell-shaped, consistent with expectation for a normal distribution. Note that the median and mean shown on the box plot are coincident. This output is consistent with normally distributed data.

Finally the "Normal Probability Plot" (Figure 2-19) does not depart from the linear trend, which would be evidence of deviation from a normal distribution.

```
> qqnorm(carawidth)
> qqline(carawidth)
```

Figure 2- 19.
Tests of normality
based on a
probability plot of
the distribution
of carapace
widths for mature
female blue crabs
in Chesapeake
Bay.



Thus we have several lines of evidence to suggest that the size distribution of mature female blue crabs is normal. The Shapiro-Wilkes test did not demonstrate significant deviation from normality. A bell-shaped curve was evident in the histogram plot. The mean, median and mode were coincident, and the probability plot showed no systematic deviation from linearity.

The format of the following summary is appropriate for describing data that is normally distributed, and should be carefully followed in reports.

"Mature female blue crabs in Chesapeake Bay ranged from 107 to 163 mm (Mean 134.8 ± 1.13 , $n = 129$) during the period of study. The distribution of mature female crab sizes was normally distributed, (Shapiro-Wilk Statistic = 0.99, $p < 0.39$; visual examination of probability plot, Figure 2–13). A definition of a particularly large female crab could be based on the 95th percentile of 156 mm."

The mean is presented with its standard error for reasons that will become clear when the topic of statistical inference is covered in Module 3. Inclusion of the histogram shown in Figure 2–13 is optional, depending upon the emphasis you wish to place on the frequency distribution of CW measurements in your report.

Note that no mention is made of the mode, median, quartiles or inter-quartile range. These statistics add no additional information for normally distributed data.

For normally distributed data, an equally useful alternative definition of extreme events could be made in terms of standard deviations from the mean. Less than 1% of values would be expected to lie outside three standard deviations from the mean.



Tidy up the program by ensuring there are no elements remaining of the program that did not work. Save the program to disk within Tinn-R for future reference. You may also want to save your workspace and Rhistory, if you want to return to this example at some stage.

Exit from R and Tinn-R by choosing File->Exit from the Menu Bar.

Source

The length frequency data on blue crabs were sourced from the Virginia Institute of Marine Science, Juvenile Fish and Blue Crab Trawl Survey. The web-based data retrieval system appears online [<http://www.fisheries.vims.edu/vimstrawldata/>]

Where have we come?

With some sound theory behind us from Lessons 1 and 2, it was time to get our hands dirty with some analyses. Here you were required to dust off what you learned about the R programming language in Module 1, and undertake some analyses by following the blow-by-blow sample analyses.

Skills imparted in Lesson 3 include

- How to analyse categorical data by constructing frequency tabulations, histograms and barcharts using `table()`. We introduced a few extra bells and whistles by using indexing to subgroup data, which we visualized using `barplot`.
- How to compute summary statistics with `summary()`, `quantile()`, `mean()`, `var()`, `sd()`, the mode and the coefficient of variation. Some statistics had to be calculated using several R functions such as the mode (e.g. `which.max(table(variable))`). The analyses included how to make an assessment of whether the data were drawn from a normal distribution using histograms, stem-leaf plots, probability plots and a comparison of mean, median and mode.
- How to report the results of your analysis for normal and non-normal data respectively. There is a very strong distinction in the statistics you report in each case, and you must be aware of this.

And of course, working through these examples should have reinforced a number of skills required to use R for statistical analyses, including the use of the `read.table()` function to read data in, assigning labels to values of a variable and how to index data from a vector in R, etc.

Lesson 4: Some Challenging Exercises

Exercise 2-1: Trawl Catch Statistics

A proforma in Word can be downloaded from the course website to assist in preparing your answer to this question.

Many fisheries agencies keep detailed statistics on fish stocks, sampling specifically for that purpose from research vessels. We have at hand trawl catch statistics for a coastal estuary for the years 1999 and 2000. The data are in the form:

SMALLMOUTH_FLOUNDER	1999 JUN	84
SPOT	1999 JUL	180
BLUE_CRAB	1999 MAY	27
BAY_ANCHOVY	1999 AUG	43
ATLANTIC_CROAKER	1999 FEB	253

where the first column is the fish species, the second column is the year, the third column is the month and the fourth column is fish length in mm. There are data for 53,856 fish in the dataset. In this exercise, you are asked to interrogate the dataset to answer some questions of specific interest.

Input the data to a data.frame called trawl.

Add the following headings to the data: “species”, “year”, “month”, “length.mm”. Transform the length measurements from mm to cm and add a new column named “length.cm” to the data.frame.

Confirm that that the data have been correctly input.

Generate summary statistics for each species, including only sample size, minimum, maximum and mean fish size.

Generate a barchart showing the relative abundance of the different species in the trawl dataset. Your analysis should yield a high quality barchart. It would be nice if you add a title to your graph.

Generate a histogram showing the size distribution for the most abundant fish species in the dataset. Use a subset statement to select only data for that fish species. Your analysis should yield a high quality histogram. Again a title would be nice.

Calculate a full set of summary statistics for length of the above species. Prepare a complete statistical summary for fish length (mm) of the above species. Make sure that your summary conforms to the

standard outlined in the worked examples.

Perform an appropriate analysis to yield a histogram showing the size distribution for Spotted Hake. Your analysis should yield a high quality histogram. Be sure to add a title to your graph.

Clearly fish length for Spotted Hake is not normally distributed, but it is unimodal. Repeat the analysis on this variable following a standard square root transformation and a log transformation.

Calculate a full set of summary statistics for length of the above species after applying the transformation that was most effective in normalizing the data.

Succinctly summarise what you conclude about the Normality of fish length for the above species **following transformation**. Include reference to supporting evidence in the form of graphs and/or tables.

Prepare a barplot to compare the size distributions of the two most common species in the dataset.

Prepare a barplot to compare the size distributions of the most common species in 1999 and 2000.

Exercise 2-2: Water Chemistry of Lake Carcoar

A proforma in Word can be downloaded from the course website to assist in preparing your answer to this question.

Data on chemical composition of the water of Lake Carcoar, near Cowra, were entered as part of a project on water quality management conducted at the University of Canberra.

Lake Carcoar is a relatively small storage in an agricultural district, so its water quality is of particular concern to the New South Wales water authorities.

The data are held in disk file CARCOAR.DAT and the measurements have been selected because of their known relationship to algal production, particularly production by diatoms. These algae are single-celled and secrete elaborate silica skeletons. Blooms of these microscopic organisms can cause severe deterioration of water quality.

Variable	Columns	Units
STATION NUMBER	1- 7	
DATE	8-13	ddmmyy
NITRATE	16-22	mg/l
SILICA	25-30	mg/l
SOLUBLE PHOS	33-37	mg/l
TOTAL PHOSPHORUS	40-44	mg/l
AMMONIA	47-51	mg/l
CHLOROPHYLL-A	54-56	UNESCO units
CONDUCTIVITY	59-61	microsiemens/cm
TURBIDITY		NTU

The Water Authorities would like to design a monitoring programme for this lake, based on knowledge of the typical concentrations of each of these key measurements. They would also like forewarning of algal blooms, and information that can be used to define upper acceptable limits for each of these variables would be most welcome.

Perform the appropriate analyses for one of NITRATE, TOTAL PHOSPHORUS or SILICA, and provide a brief reports on each for the Water Authorities, using the proforma supplied.

Undertake the appropriate analyses to determine whether the concentration is normally distributed. Present the outcomes of the analysis below. Be sure to include a histogram.

Compute a comprehensive set of summary statistics for the variable. Present the full set of statistics below in tabular form.

What do you conclude regarding the normality of the variable? Be sure to include supporting statistics or cross-references to diagrams and tables produced during the analysis.

Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Include in your summary, a description of the distribution of values, only those descriptive statistics appropriate to the data, and a working definition of an extreme value.

With regard to normality, are your results consistent with expectation for the variable? Why?

What advice would you give to anyone planning further statistical analyses on the variable?

What recommendations would you like to make to the NSW Water Authorities?

Append a full R program listing, cleaned up and free from error or redundant code.

Exercise 2-3: Inflows to Burrinjuck Dam

Data on water flow for Burrinjuck Dam were collected as part of a project on water quality management conducted by the University of Canberra for the Land and Water Resources Research and Development Corporation.

The data were provided by the New South Wales Department of Water Resources and comprise the following variables stored in the file BJUCK.DAT. DATE represents the date at which the water collections were taken. Depth, volume in megalitres (ML) and area were measured or estimated for that date, and inflow and outflow were measured using appropriate gauging stations.

The format of the data in BJUCK.DAT is shown in the table below. For example, the measurement for depth occupies position 9 to 13 on each line in the data file.

Variable	Columns	Units
DEPTH	9-13	m
VOLUME	16-21	ml
AREA	24-27	ha
INFLOW	30-34	ml/d
OUTFLOW	36-40	ml/d
RAINFALL	43-46	mm/d
EVAPORATION	49-52	mm/d

The New South Wales Department of Water Resources requires a detailed summary of the flows, rainfall and evaporation for Burrinjuck Dam.

Analysis of inflows

Perform the appropriate analyses for INFLOW only, and provide a brief report for the NSW Department of Water Resources, using the proforma supplied.

It is sound practice when analysing data that is not your own, to examine it before analysis. You should read the raw data into the Editor for perusal before beginning the analysis. Once you are satisfied, undertake the appropriate analyses, graphical and otherwise, to determine whether the inflows are normally distributed.

What do you conclude regarding the normality of the variable INFLOW? Be sure to include supporting statistics or cross-references to diagrams and tables produced during the analysis.

Compute a comprehensive set of summary statistics for the variable INFLOW. Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Include in your summary, a description of the distribution of INFLOW values, only those descriptive statistics appropriate to the data, and a working definition of an extreme inflow.

With regard to normality, are your results consistent with expectation for a variable such as INFLOW? Why?

What advice would you give to anyone planning further statistical analyses on INFLOW?

Append a full R program listing, cleaned up and free from error or redundant code.

Analysis following transformation

If the analysis of the Burrinjuck inflows shows that the variable INFLOW is not normally distributed, repeat the analysis on this variable following a standard log transformation and a square root transformation.

Undertake the appropriate analyses to determine whether the logged inflows are normally distributed. Repeat for the square root flows. Select the transformation that is the most successful in normalising the inflows. Present the outcomes of the analysis using the best transformation below. Be sure to include a histogram.

Compute a comprehensive set of summary statistics for the transformed inflows. Present the full set of statistics below in tabular form.

What do you conclude regarding the normality of the transformed inflows? Be sure to include supporting statistics or cross-references to diagrams and tables produced during the analysis.

Provide a concise summary of the results, such as might appear in the results section of a manuscript or report. Include in your summary, a description of the distribution of the transformed inflows, only those descriptive statistics appropriate to the data, and a working definition of an extreme inflow.

What advice would you give to anyone planning further statistical analyses on inflows?

What recommendations would you like to make to Department of Water Resources?

Append a full R program listing, cleaned up and free from error or redundant code.

Where have we come?

Lesson 4 is where the real learning occurs. In earlier lessons, you have read and understood written material and been led through worked examples. It is a bit like watching television. In Lesson 4 you were required to recall and integrate the information to complete some challenging real-world exercises. Recall in the context of problem solving is one of the best ways of achieving lasting learning. It is hard yakka.

In completing this module successfully, you will have achieved a number of core competencies, namely,

- Knowledge of the options available to you for summarizing data in tabular form, in graphical form and in the form of summary statistics.
- Understanding the distinction between the various statistical options available for summarizing univariate data, and when and when not to use them.
- A working knowledge of the R interface, the function of each window, and how to navigate among them in order to perform the statistical analyses.
- The ability and confidence to interpret the results of the analyses in a biological context based on demonstrated understanding of the analyses.
- The ability to present findings in a style appropriate to the scientific literature.
- Appropriate attitudes and efficient strategies for extending your abilities to conduct analyses and solve problems beyond the scope of this module, by using resource materials such as statistical texts, software manuals, and your colleagues.

References

Begg, R, Walsh, B, Woerle, F & King, S. (1983). Ecology of *Melomys burtoni*, the grassland *Melomys* (Rodentia, Muridae) at Cobourg Peninsula, N.T. *Australian Wildlife Research* 10:259-267.

Sokal & Rohlf (1994). *Biometry. The Principles and Practice of Statistics in Biological Research*, 3rd Ed, W.H. Freeman and Company, San Francisco, USA.

Virginia Institute of Marine Science (2003). Juvenile Fish and Blue Crab Trawl Survey. <http://www.fisheries.vims.edu/vimstrawldata/>