Chapter 1

Summarising Data

In statistical data analysis, the number of experimental or observational units (and the number of variables) is often large. For presentation purposes, it is impractical to present the whole data. Furthermore, the data are often not particularly informative when presented as a complete list of observations. A better way of presenting data is to pick out the important features using **summary measures** or **graphical displays**.

1.1 Summary Measures

The data in the file `silt2.dat` were collected as part of an investigation into soil variability. Soil samples were obtained in each of 4 sites in the province of Murcia, Spain, and the percentage of clay was determined. At each site, 11 observations were made (at random points in a 10m×10m area). The eleven observations for each of the first four sites are presented in the dotplot below.

```
   ....   .  . .. .      .           .
-------+---------+---------+---------+---------+---------Site 1


        :     . ...: :  .
-------+---------+---------+---------+---------+---------Site 2

                .
        . :   . :. .  . .
-------+---------+---------+---------+---------+---------Site 3


                      . .. ..  .    .:.             .
-------+---------+---------+---------+---------+---------Site 4
    28.0      35.0      42.0      49.0      56.0      63.0
```

Clearly there are some differences in the distributions of the observations at each of the sites. These differences can be described in terms of the **location** and **spread** of the data.

1.1.1 The Mean

Any summary measure which indicates the centre of a set of observations is a **measure of location** or a **measure of central tendency**. Perhaps the most often used measure of location is the **mean** of the observations.

Suppose that we have n observations of a variable X, and the values of the observations are denoted by $x_1, x_2, \ldots, x_n$, then we denote the mean by $\overline{x}$, and

$$\overline{x} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \quad \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

♡ **Example** 1.1.

For the data in the file `silt2.dat`, the mean percentage clay for the first site is given by

$$\begin{aligned} \overline{x} \quad &= \quad \tfrac{30.3+27.6+40.9+32.2+33.7+26.6+26.1+34.2+25.4+35.4+48.7}{11} \\ &= \quad \tfrac{361.1}{11} \quad = \quad 32.83 \end{aligned}$$

Similarly, the mean percentages of clay for sites 2, 3 and 4 are 34.80, 34.05 and 45.77 respectively. Clearly, presenting the mean conveys the information that the distributions of observations for sites 1,2 and 3 have similar locations while the observations for site 4 are generally larger.

In MINITAB

```
MTB > mean c1
```

Calc→Column Statistics

Stat→Basic Statistics→Display Descriptive Statistics

1.1.2 The Median

An alternative to the mean as a measure of location is the **median** of the observations. The median is the 'middle' value.

For example, the eleven observations of the clay percentage for the first site are, when placed in order

25.4 26.1 26.6 27.6 30.3 32.2 33.7 34.2 35.4 40.9 48.7

```
   ....   .  . .. .      .           .
-------+---------+---------+---------+---------+---------Site 1
    28.0      35.0      42.0      49.0      56.0      63.0
```

Similarly, the median percentages of clay for sites 2, 3 and 4 are 35.9, 34.5 and 44.5 respectively. Again, the median conveys the information that the distributions of observations for sites 1,2 and 3 have similar locations while the observations for site 4 are generally larger. If there are an **even** number of observations, then there isn't a single 'middle observation' and the median is defined to be half way between the 'middle two' observations.

In general:

if we have an odd number of observations, then the median is the value of the $\frac{n+1}{2}$th largest.

if we have an even number of observations, then the median is the mean of (half way between) the $\frac{n}{2}$th largest and the $(\frac{n}{2}+1)$th largest.

In MINITAB

```
MTB > median c1
```

Calc→Column Statistics

Stat→Basic Statistics→Display Descriptive Statistics

Why use the median rather than the mean?

The mean is the summary of location which is most often calculated and quoted. However, there are situations where the median provides a better summary of location.

The median is much less sensitive (more robust) in situations where there are a small number of extreme observations. It is a better measure of a 'typical observation'. (Indeed, it often is the value of an actual observation). However, the mean has many nice 'statistical properties' which we shall discuss later.

1.1.3 Measures of Spread

Any summary measure which indicates the amount of dispersion of a set of observations is a **measure of spread**.

The easiest measure of spread to calculate is the **range** of the data, the difference between the smallest and largest observations. For example, consider the eleven observations of the clay percentage for the first site.

```
   ....   .  . .. .      .           .
-------+---------+---------+---------+---------+---------Site 1
    28.0      35.0      42.0      49.0      56.0      63.0
```

The range for the percentages of clay for sites 2, 3 and 4 are 11.4, 11.3 and 21.4 respectively. This conveys the information that the observations for sites 2 and 3 have a very similar spread, which is somewhat smaller to that for sites 1 or 4.

However, the range is not a very useful measure of spread, as it is extremely sensitive to the values of the two extreme observations. Furthermore, it gives little information about the distribution of the observations between the two extremes.

A more robust measure of spread is the **interquartile range** (or quartile range). This is the difference between the **lower quartile** and **upper quartile**.

The lower and upper quartiles, together with the median, divide the observations up into four sets of equal size.

For example, for the eleven observations of the clay percentage for the first site

```
   ....   .  . .. .      .           .
-------+---------+---------+---------+---------+---------Site 1
    28.0      35.0      42.0      49.0      56.0      63.0
```

In general:

the upper quartile is the value of the $\frac{3}{4}(n+1)$th largest.

the lower quartile is the value of the $\frac{1}{4}(n+1)$th largest

If $n+1$ is not divisible by 4 then some interpolation is required. However, MINITAB does this for us.

The interquartile range may be interpreted as the range in which the 'middle half' of the observations lie.

For the sets of observations of clay percentages for the four sites, the interquartile ranges are 8.8, 4.9, 6.5 and 8.7, which again illustrates the difference in spread between the observations for sites 1 and 4, and those for sites 2 and 3.

Although the range and the interquartile range are easy to calculate and interpret, they do not have nice statistical properties. For future use, we shall define a further measure of spread called the **standard deviation**.

Recall that we denote the n observations by $x_1, x_2, \ldots, x_n$ and the mean of the sample by $\overline{x}$. Then for each observation x_i, $i = 1, 2, \ldots, n$, $x_i - \overline{x}$ is the difference between that observation and the mean.

```
   ....   .  . .. .      .           .
-------+---------+---------+---------+---------+---------Site 1
    28.0      35.0      42.0      49.0      56.0      63.0
```

Some values of $x_i - \overline{x}$ are positive and some are negative.

However, all values of $(x_i - \overline{x})^2$ are positive, and the larger values of $(x_i - \overline{x})^2$ correspond to values which are further away from the mean.

We define the **variance** of the observations to be the sum of the values of $(x_i - \overline{x})^2$ for all observations, divided by $n - 1$. (If we divide by n here, we would have the mean value of $(x_i - \overline{x})^2$, but this does not have such nice statistical properties). Hence the variance, denoted by s^2 is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

The **standard deviation** of the observations, which we denote by s, is the square root of the variance.

If the observations are more highly spread out, then in general they will be a greater distance from the mean (which indicates the 'centre' of the observations) and therefore the standard deviation will be greater.

Therefore, the standard deviation is a measure of spread.

For the sets of observations of clay percentages for the four sites, the standard deviations are 7.07, 3.66, 3.55 and 6.17, which again illustrates the difference in spread between the observations for sites 1 and 4, and those for sites 2 and 3.

Measures of spread in MINITAB

Calc→Column Statistics

Stat→Basic Statistics→Display Descriptive Statistics

1.1.4 Accuracy

Summary statistics such as means and standard deviations may often be produced with a large number of decimal places.

There is no 'golden rule' as to how many decimal places should be reported, but a number of points should be taken into consideration.

1. Consider the accuracy to which the data have been measured.

   If summaries are presented containing many more decimal places, then this provides 'spurious' accuracy which is not justified by the data collection process.

If summaries are presented containing many fewer decimal places, then important information may be lost.

2. For continuous data, consider the variability of the data.

   For example, if all the observations are the same up to and including the first decimal place, with variability occuring in the second decimal place and beyond, then clearly at least two, and probably more decimal places, are required.

3. For discrete data, there is no need for summaries to be reported on the same scale as the data.

   For example, it is perfectly reasonable that the mean of a set of counts may not be a whole number.

4. Do not truncate trailing zeros.

   Once you have decided on a certain number of decimal places to report, then report them all, even if the last one is a zero. Otherwise you are throwing away information.

1.2 Graphical Displays of Data

Often, a simple graphical display provides a more easily interpretable summary of the distribution of the observations than a collection of summary statistics.

One graphical display, which is easy to construct, and incorporates many of the features of the summary measures introduced in §1.1 is the **box-and-whisker plot** (or simply **boxplot**).

1.2.1 The Boxplot

We will illustrate this using data in the file `quake.dat` which represent the time in days between successive serious earthquakes worldwide, between 16th December 1902 and 4th March 1977.

Constructing a boxplot involves the following steps:

1. Draw a vertical (or horizontal) axis representing the interval scale on which the observations are made.

2. Calculate the median, and upper and lower quartiles (Q_1, Q_3) as described in §1.1. Calculate the interquartile range (or 'midspread') $H = Q_3 - Q_1$.

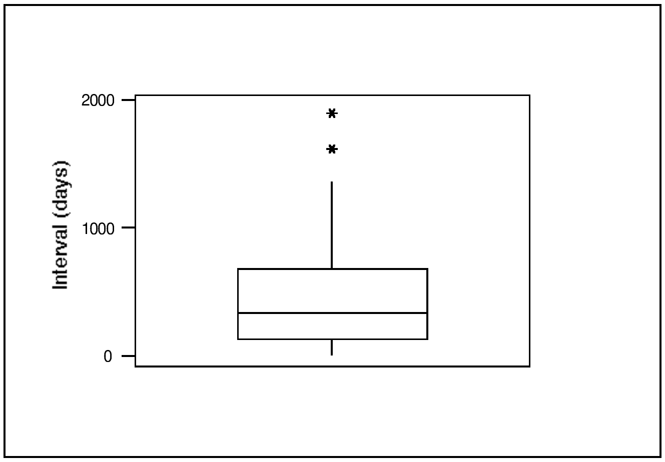
3. Draw a rectangular box alongside the axis, the ends of which are positioned at Q_1 and Q_3. (The box covers the 'middle half' of the observations). Q_1 and Q_3 are referred to as the '**hinges**'.

4. Divide the box into two by drawing a line across it at the median.

5. The **whiskers** are lines which extend from the hinges as far as the most extreme observation which lies within a distance $1.5 \times H$, of the hinges.

6. Any observations beyond the ends of the whiskers (further than $1.5\times H$ from the hinges) are **outliers** and are each marked on the plot as individual points at the appropriate values. (Sometimes a different style of marking is used for any outliers which are at a distance greater than H from the end of the whiskers).

From a boxplot, you can immediately gain information concerning the centre, spread, and extremes of the distribution of the observations.



In MINITAB

Graph→Boxplot

1.2.2 The Time Series Plot

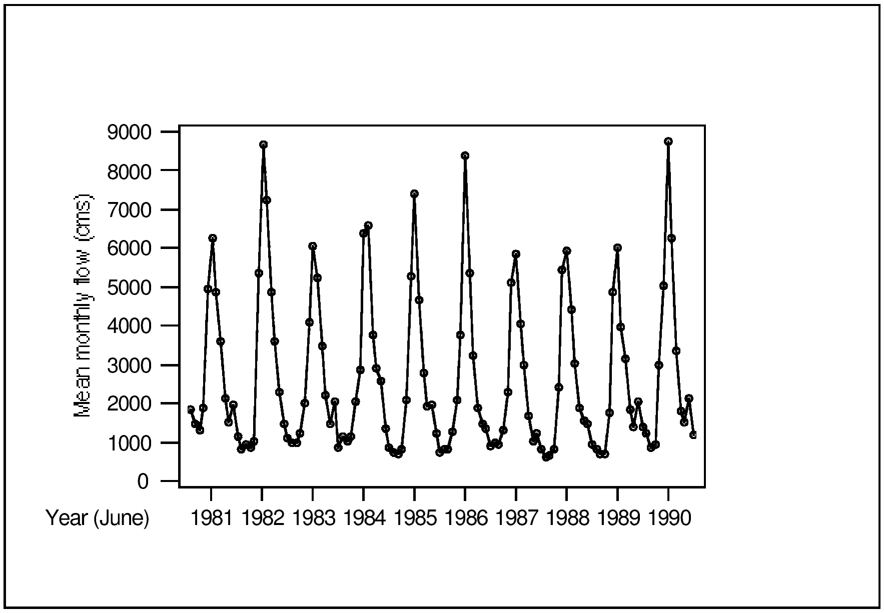
Often, the data collected are observations of the same quantity at different points in time (the units are time points). For example, weekly mean precipitation, monthly maximum sea level . . .

Where the time points at which the data have been collected are evenly spaced (or approximately so) then a **time series plot** may be used to illustrate the variation in the observations.

A time series plot is simply a plot of each observation x_i, $i = 1, 2, \ldots, n$ on the y-axis against its **index** i on the x-axis, in other words a plot of the points (i, x_i), $i = 1, 2, \ldots, n$.

Consecutive points are joined together to illustrate the way in which the observations vary over time.

For example, the data in the file `flow.dat` represent the mean monthly flow (in cms) of the Fraser River at Hope, B.C., Canada between January 1981 and December 1990.



In MINITAB

Graph→Time Series Plot

Time series plots may be used to detect **trend** or **seasonal** behaviour (or both).

Note that in many practical examples, there is no natural time ordering of the observations (for example, observations where the units are individuals). In such examples, time series plots are meaningless.

1.2.3 The Histogram

Histograms have the following properties.

1. The horizontal axis represents the scale on which the observations are measured, and the bars of the histogram adjoin each other with the boundaries between bars representing the boundaries between the categories.

2. If bars are not of equal width, then care must be taken when determining the height of each bar (particularly with MINITAB) to ensure that the **area of each bar is proportional to the number of observations in each category**.

3. The best choice of boundaries between bars is the one which best illustrates the distribution of the observations. This usually requires some experimentation (trial and error).

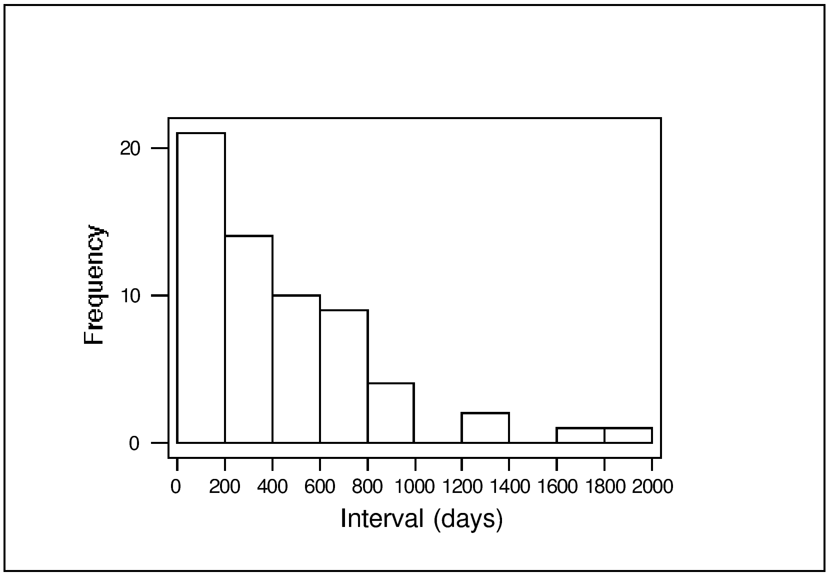


Figure 1.1: A histogram of the earthquake data (`quake.dat`) introduced in §1.2.1.

In MINITAB

Graph→histogram

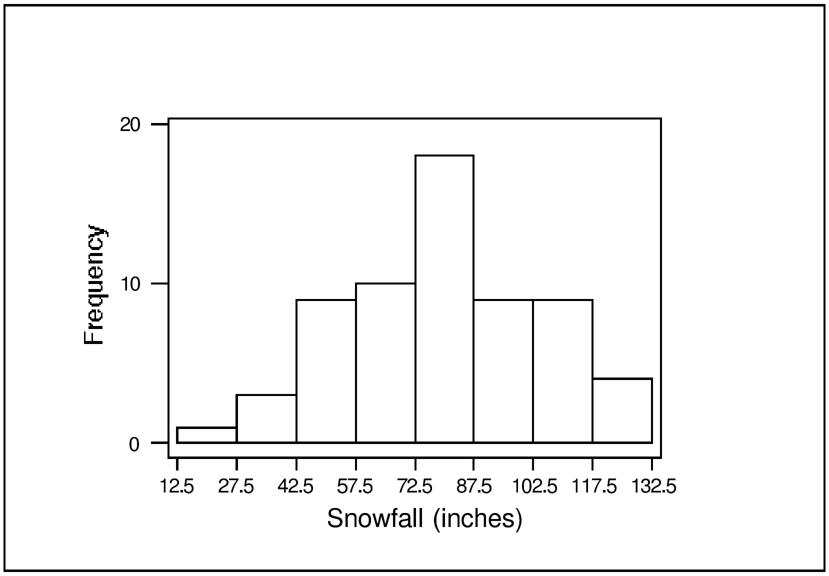
There are a number of features of the distribution of a set of observations which are not summarised by the summary measures described in §1.1. but which are illustrated by a histogram.

For example, we can determine if the distribution of the data is **symmetric** or **skew**.

The data in the file `snow.dat` represent the annual snowfall (in inches) in Buffalo, NY, for the years 1910 to 1972.

A histogram can also be used to determine if the distribution of the observations is **unimodal** (a single 'largest' category with categories generally becoming 'less common', above or below this category) or **multimodal**.

The data in the file `acidity.dat` are the measurements of an acidity index for each of 155 lakes in the Northeastern USA.



1.3 Summarising the Joint Distribution of a Pair of Variables

Many interesting problems in statistical data analysis concern the **relationship** or **association** between a pair of variables. When observations are made of two or more variables, on the same set of units, we can examine such relationships by investigating the **joint distribution** of pairs of observations.

The simplest way of summarising the joint distribution of a pair of variables is by a **scatterplot**. Suppose that we have observed n units and we denote the measurements of one variable by $x_1, x_2, \ldots, x_n$ and the measurements of the other variable by $y_1, y_2, \ldots, y_n$. Then a scatterplot is a plot of the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

We consider two examples here, and in each case the question of interest is what, if any, is the relationship between the two variables?.

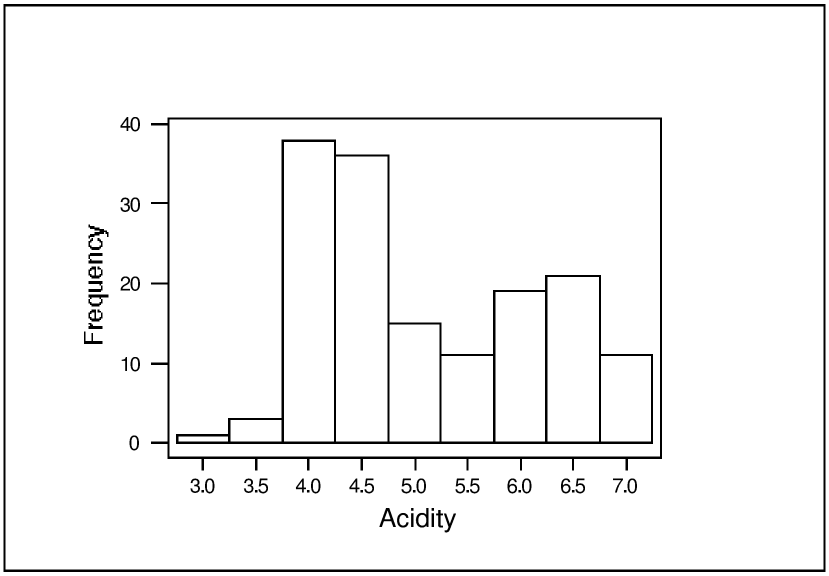
The data in the file `level.dat` record the level of Lake Victoria Nyanza for the years 1902–1921 (relative to a fixed standard) and the number of sunspots in the same years.

The data in the file `paving.dat` are the compression strength (Nmm^{-2}) and percentage dry weight of 24 paving slabs. In each case the question of interest is what, if any, is the relationship between the two variables?

In MINITAB

Graph→Plot

The strength of the association between the variables may be summarised by a single summary measure called the **correlation coefficient**.



To calculate the correlation coefficient, we first need to calculate the mean and standard deviation of the observations $x_1, x_2, \ldots, x_n$ of the first variable (call these $\overline{x}$ and s_x), and the mean and standard deviation of the observations $y_1, y_2, \ldots, y_n$ of the second variable (call these $\overline{y}$ and s_y). The correlation coefficient (denoted by r) is given by

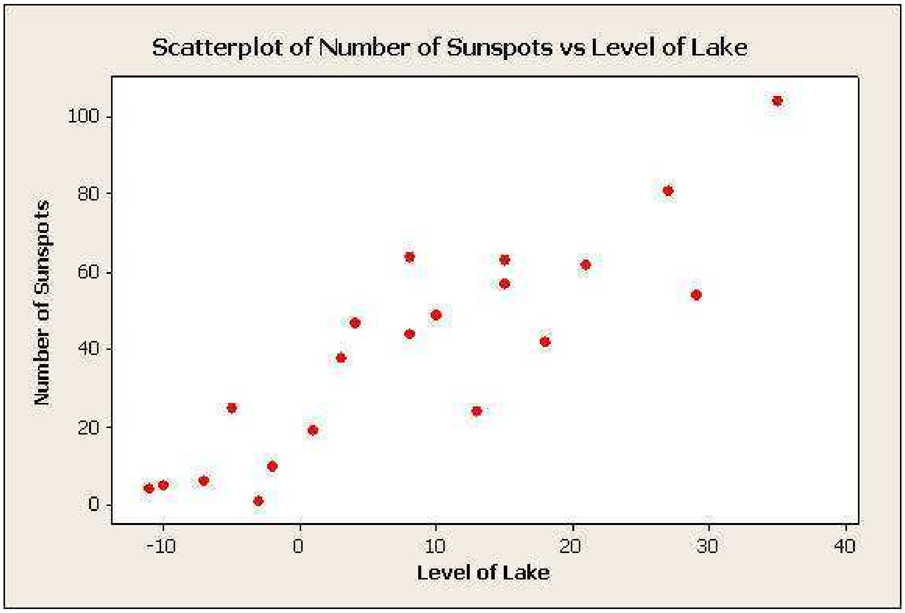
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y}.$$

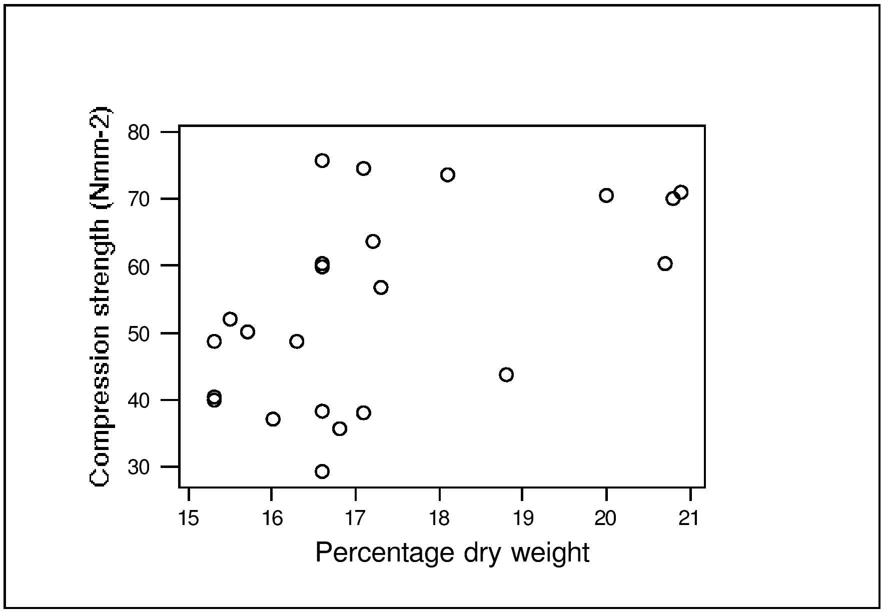
The correlation coefficient, which must lie between -1 and 1, measures the strength of the **linear** (straight line) relationship between the variables. It determines to what extent values of one variable increase as values of the other variable increase, and how close this relationship is to being a perfect straight line.

Hence, the correlation coefficient provides a measure of the extent of linear association. For example, the correlation coefficients for the two examples illustrated by scatterplots on the previous page are 0.526 between ‘strength’ and ‘dry weight’ and 0.879 between ‘lake level’ and ‘number of sunspots’. Therefore, both data sets show positive linear association, stronger between lake level and number of sunspots.

In MINITAB

Stat→Basic Statistics→Correlation


Scatterplot of Number of Sunspots vs Level of Lake
Number of Sunspots
100
80
60
40
20
0
-10
0
10
20
30
40
Level of Lake


Compression strength (Nmm-2)
80
70
60
50
40
30
15
16
17
18
19
20
21
Percentage dry weight