One Numerical Variable

Data may have a rich structure, but we first concentrate on the simplest data structure, a single numerical measurement from each 'individual' in a group. We call data of this form a batch of numbers, or a univariate numerical data set.

Examples of the types of 'individuals' from which we often record data are ...

- People
- Plants
- Samples of output from a manufacturing process
- Admissions of patients to a hospital

In this chapter, we describe several common ways to display the variability in a batch of numerical data. We show how such data can provide useful information.

Graphical Display of Values

Dot plots and stem and leaf plots show each value in a data set graphically.

Analysing Variation

Meaningful information can be obtained from variation in the values of a variable.

**Information from the variation in data **

Variation in data is not simply an annoyance — the variation itself can hold important information.

Simply sorting a data set into order can highlight features that are not obvious in the raw data, such as the lack of values between 3.4 and 4.9 in the data below.

6.1	5.2	7.9	2.3	3.4
1.4	5.3	7.1	3.2	2.8
5.1	6.9	6.1	3.4	5.2
5.5	2.0	1.3	4.9	6.4

Basic Dot Plot

A dot plot displays each value as a cross along a numerical axis.

Dot plots

Some ranges of values are more common than others — they have higher density.

The simplest graphical display of data that shows where there is high and low density is a dot plot. This shows each value as a cross (or dot) against a numerical axis.

The gap between 3.4 and 4.9 is more obvious on the right than in a textual list of values, whether ordered or not.

[TODO: IMAGE PLACEHOLDER]

Jittered Dot Plot

Jittering is a modification to the basic dot plot that avoids some problems associated with overlapping crosses

Jittering the crosses

In all but the smallest data sets, the crosses on a basic dot plot overlap, making it difficult to identify regions of high density.

Randomly moving crosses away from the axis reduces this problem by separating the crosses:

Note that the vertical jittering is random and therefore tells you nothing about the data.

[TODO: IMAGE PLACEHOLDER]

Stacked dot plots

Stacking of the crosses is an alternative to jittering that highlights ranges of high or low density.

Stacked dot plots

Stacking the crosses into columns is usually better than jittering them.

[TODO: IMAGE PLACEHOLDER]

Stacking highlights regions of high density well (tall stacks).

Stems and Leaf Plots

Stem and leaf plots are similar to stacked dot plots, but a digit is used instead of a cross to retain extra information.

Digits instead of crosses

Stem and leaf plots are closely related to stacked dot plots. The crosses are replaced by digits that provide a little more detail about the values that they represent.

Stem and Leaf

In a stem and leaf plot, the axis is replaced by a column of 'stems' — the most significant digits of the values in the data. The digits that replace the crosses are called 'leaves' and give a further significant digit of each value on a stem.

[TODO: IMAGE PLACEHOLDER]

A final refinement is to sort the leaves into increasing order on each stem.

Splitting the Stems

To increase the flexibility of the display, each stem may be repeated either 2 or 5 times, increasing the number of classes in the basic stem and leaf plot by a factor of 2 or 5.

Need for more flexibility

Sometimes basic stem and leaf plots are not flexible enough — there would be either too many or too few rows of leaves to show the varying density well.

Repeating each stem 2 times (with leaves 0-4 on the lower copy and leaves 5-9 on the upper one) or 5 times (with leaves 0-1, 2-3, 4-5, 6-7 and 8-9 on the different copies) gives intermediate numbers of stems.

[TODO: IMAGE PLACEHOLDER]

Drawing Stem and Leaf Plots

For data analysis, stem and leaf plots are rarely more informative than stacked dot plots, but they are easy to draw by hand.

**Smoothness **

When drawing a stem and leaf plot, the aim is for a smooth shape to the stem and leaf plot, and this is usually achieved by between 10 and 20 rows of leaves.

Drawing by hand

When data are analysed on a computer, a stacked dot plot usually describes a distribution of values more clearly than a stem and leaf plot.

However stem and leaf plots are easy to draw by hand:

- Decide on the stems to use — usually between 10 and 20 of them.
- Scan down the values identifying the leaf digits and writing them against the value's stem.
- Finally, sort the leaves on each stem into order.

To simplify drawing, values are truncated to give their stems and leaf digits, not rounded. For example, 7.98 and 7.90 would both be displayed as leaf '9' on the stem '7'.

2. Understanding Distributions

Outliers

Does the data contain any outliers -- values that are atypically large or small? The extreme values in a skew distribution are often mistaken for outliers.

**Outliers **

Values that are considerably larger or smaller than the bulk of the data are called outliers.

An outlier may have been incorrectly recorded, or there may have been other anomalous circumstances associated with it. Outliers must be carefully checked if possible. If anything atypical can be found, outliers should be deleted from the data set and their deletion noted in any reports about the data.

Outliers and skew distributions

Deciding whether a value is an outlier or not is affected by the shape of the distribution of values for the rest of the data.

Symmetric distribution

[TODO: IMAGE PLACEHOLDER]

Skew distribution

A distribution with a long tail to one side is called a skew distribution — positively skew if the long tail is to the right and negatively skew if the long tail is to the left. It is not unusual for the extreme value in a very skew distribution to be a fair distance from the other values and may not be an outlier.

[TODO: IMAGE PLACEHOLDER]

Clusters

Does the data split into separate clusters -- ranges of values with high density separated by ranges with low density? Clusters may correspond to different groups of individuals.

**Clusters **

If a dot plot, stem and leaf plot or histogram separates into two or more groups of values (clusters), this suggests that there may be more fundamental differences between the 'individuals' in the groups.

[TODO: IMAGE PLACEHOLDER]

Further investigation should be made of the individuals in the clusters to find whether they also differ in other ways.

If the clusters were less distinct, especially in small data sets, you would need external supporting evidence before concluding that the individuals separated into meaningful groups.

Distribution of Values

The distribution gives information about a typical value round which the data are spread (the distribution's location or centre) and the variability of the values (the spread of the distribution).

**Distribution of values **

Even when a data set has no outliers or clusters, the distribution of values also contains useful information. Important features are:

- The centre or location of the distribution — a 'typical value'
- The spread or variability of the distribution
- Whether a distribution is symmetric or skew — do the tails appear similar at both sides?
- Other aspects of the shape of the distribution

The concepts of centre and spread are particularly important.

Names of Individuals

Additional information about the items from which measurements have been made can help us understand the distribution of values in the data.

Extra information

When only a single value is known from each individual (or plant, item, etc), all that can be discovered is the shape of the distribution of these values.

Additional information about each individuals may give insight into why some values are bigger or smaller than others. Different types of information may be available. The simplest is a unique name for the individuals — a textual label. These names may help us to understand why values are outliers or group into clusters in a dot plot or stem and leaf plot.

Distinguishing Known Groups

If we know that the values come from 2 or more groups of individuals, dot plots can be modified to show this extra information.

Multiple groups of individuals

Sometimes we know that the individuals belong to two or more groups before the data are collected or, equivalently, that they have different values of an extra categorical variable.

Information about groups is best displayed by plotting the separate groups against a common axis.

[TODO: IMAGE PLACEHOLDER]

Back-to-back stem and leaf plots

Stem and leaf plots can be used to compare two groups of individuals, if drawn on different sides of a common column of stems. (They are less useful if there are three or more groups.)

[TODO: IMAGE PLACEHOLDER]

Dangers of overinterpretation

There is a risk of over-interpreting patterns in small data sets.

Warning: Features in the distribution of a small data set may not be meaningful

Be careful not to overinterpret patterns in small data sets. Clusters, outliers or skewness may appear by chance even if there is no meaningful basis to these features.

Pronounced outliers or clusters may be taken as indicative of something meaningful in the underlying process. However less pronounced outliers or clusters must be supported by outside evidence before these features can be interpreted as meaningful.

3. Histograms and Density

Density of Values

The heights of the stacks of crosses in a dot plot describe the density of values.

Density

In a stacked dot plot (or stem and leaf plot), the highest stacks contain the most values. These stacks have the highest density of values.

[TODO: IMAGE PLACEHOLDER]

Histograms more directly show how density varies along the axis.

Histogram with Equal Class Widths

In a simple histogram, the height of the rectangle above each class on the axis equals the number of values in the class -- the class frequency.

Histograms

In a simple histogram, the axis is split into sub-intervals of equal width called classes. A rectangle is drawn above each class with height equal to the number of values in the class — the frequency of the class.

[TODO: IMAGE PLACEHOLDER]

Choice of classes

Class width and start-point should be chosen to make the histogram as smooth as possible -- neither too blocky nor too jagged.

Aim of a 'smooth' histogram

There is considerable freedom in the choice of histogram classes. The exact shape depends on:

- Class width
- Start value for first class

We usually choose classes with the aim of smoothness in the outline of the histogram rectangles.

[TODO: IMAGE PLACEHOLDER]

The choice of 'best' classes is subjective, but...

WARNING: If your conclusions about what a histogram tells you about the data depend on the choice of histogram classes, you are over-interpreting its shape.

Histograms of small data sets

The shape of a histogram can be very dependent on the choice of classes if the data set is small; beware over-interpreting its shape. Stacked dot plots are a better display of small data sets.

**Warning for small data sets **

For small data sets, changing the class width and the starting position for the first class can give a surprising amount of variability in histogram shape, so be extremely wary of over-interpreting features such as clusters or skewness.

Indeed, it is probably better to avoid using histograms unless there is a reasonable number of values — stacked dot plots are far less likely to mislead you over minor features.

Relative frequency and area

In a histogram, the proportion of the total area that is above any class equals the relative frequency of the class.

Relative frequency

When all histogram classes are of equal width, histograms are often drawn with a vertical axis giving the frequencies (counts) for each class. The vertical axis can alternatively be labelled with the relative frequencies (proportions) for the classes.

(There is no harm in including both axes.)

[TODO: IMAGE PLACEHOLDER]

Area equals relative frequency

An important property of histograms is that the proportion of values in one or more classes equals the proportion of the histogram area above these classes.

[TODO: IMAGE PLACEHOLDER]

** Therefore: Relative frequency = proportion of the total area **

Comparing groups

The vertical axis should be relative frequency, not frequency, when comparing two groups with histograms. Population pyramids are often used to compare age distributions.

**Relative frequencies to compare two groups **

Histograms may be superimposed to compare two groups. However if the groups differ in size, it is usually more meaningful to compare relative frequencies (proportions) than the counts in the classes.

Use relative frequency histograms to compare groups.

[TODO: IMAGE PLACEHOLDER]

Histograms with varying class widths

If a histogram has varying class widths, the vertical axis must be 'density'. The histogram shape would be misleading if frequency or relative frequency was used for the vertical axis.

Mixed Class Widths

For some data sets, wider classes give a smoother histogram in some ranges of values (e.g. in the tail of a distribution) and narrower classes are better in other parts of the distribution (usually where there is greater density of values).

In a correctly drawn histogram, each value contributes the same area.

Histograms can be drawn with mixed class widths, but it would be badly misleading to make the rectangle heights equal to either the class frequency or relative frequency.

[TODO: IMAGE PLACEHOLDER]

Understanding histograms

The proportion of values in any classes always equals the proportion of the total histogram area that is above the classes.

**Area and proportion of values **

The details of drawing histograms by hand with varying class widths are unimportant — a computer should be used. To interpret their shape remember that:

** The proportion of the total area above any classes equals the proportion of values in them **

For example,

[TODO: IMAGE PLACEHOLDER]

Frequency Polygons

Frequency polygons are closely related to histograms but give a less 'blocky' display of density. Different groups can be compared more easily with them.

Frequency polygons

A frequency polygon is closely related to a histogram with equal class widths. It joins the midpoints of the tops of the class rectangles and tends to give a smoother outline than the corresponding histogram.

[TODO: IMAGE PLACEHOLDER]

It is easier to distinguish and compare superimposed frequency polygons for two groups than the corresponding histograms.

[TODO: IMAGE PLACEHOLDER]

Kernel Density Estimates (optional)

Kernel density estimates show density in a still smoother display.

Kernel density estimates

A kernel density estimate is an alternative to a histogram that often results in a smoother display of the density of values. Each data value on the axis is replaced by a 'blob' of ink (kernel) and these kernels are stacked.

[TODO: IMAGE PLACEHOLDER]

The widths of the kernels can be adjusted — if they are too narrow, the display becomes jagged, but if they are too wide, the display becomes too spread out and detail is lost.

Drawing Histograms by hand (optional)

Histograms are based on frequency tables. Class boundaries should avoid possible data values.

Frequency table

A computer is normally used to draw histograms. Hand-drawn histograms are based on a frequency table that lists the histogram classes and their frequencies.

To avoid ambiguity in the histogram, the class boundaries should be chosen to ensure that no data values are on boundaries. For example,

Data values	Class	Frequency
1.0 - 1.9	$0.95 \le x < 1.95$	2

|2.0 - 2.9|1.95 ≤ x < 2.95 | 3 |

|3.0 - 3.9| 2.95 ≤ x < 3.95 | 3 |

|4.0 - 4.9|3.95 ≤ x < 4.95 | 1|

|5.0 - 5.9|4.95 ≤ x < 5.95 | 5|

|6.0 - 6.9| 5.95 ≤ x < 6.95 | 4 |

|7.0 - 7.9| 6.95 ≤ x < 7.95| 2|

| | |20|

**Height of a histogram rectangle **

To draw a histogram by hand with equal class widths, each class rectangle can be drawn with height equal to its class frequency. If class widths vary, we need to calculate the density for each class with the formula:

Density = Relative frequency of class / Class width

and use this for the rectangle heights.

Median, quartiles, and boxplots

Box plots highly summarise the distribution of values in a data set. They are useful for comparing different batches of values.

The need to summarise

Histograms are based on frequency tables. Class boundaries should avoid possible data values.

Unhelpful detail when comparing groups

Dot plots, stem and leaf plots and histograms contain a lot of detail about the distribution of values in a data set. This level of detail is useful when examining a single data set, but when several groups of values are being compared, the detail distracts from the main differences between the groups.

For example, the jittered dot plots below do not concisely summarise the differences between the five groups.

[TODO: IMAGE PLACEHOLDER]

Median, quartiles, and boxplot

The median and quartiles split a batch of values into four equal-sized sets of values. A box plot is a graphical display of the median, quartiles and extremes.

Five-number summary

Five values are enough to capture a lot of information about the distribution of values in a data set.

- The two extremes (i.e. the minimum and maximum values).
- The lower quartile, median and upper quartile.

These values split the data set into four groups with approximately equal numbers of values.

Box plot

A box plot displays the five-number summary graphically.

[TODO: IMAGE PLACEHOLDER]

Details

The median, m, is the middle value if there is an odd number of values in the data set. If there is an even number of values, the median is the average of the middle two.

Different authors give slightly different definitions for the upper and lower quartiles. One definition of the lower quartile is the median of the lowest half of the data — i.e. of the values lower than m. (The upper quartile would then be defined as the median of the top half of the values.)

Important: Provided you are consistent, different definitions of the quartiles should lead you to the same conclusions.

Interpreting a box plot's shape

A box plot clearly shows the centre, spread and skewness of a data set. It splits the corresponding histogram into 4 approximately equal areas.

Box plots and histograms

Since the median and quartiles split the data set into quartiles, they also split a histogram of the data into four approximately equal areas.

[TODO: IMAGE PLACEHOLDER]

What does a box plot tell you about the distribution?

Centre

The median gives an indication of the centre of the distribution.

[TODO: IMAGE PLACEHOLDER]

Spread

The width of the box (the interquartile range) and the difference between the maximum and minimum (range) both give an indication of the spread of values.

[TODO: IMAGE PLACEHOLDER]

Skewness

> The distances of the minimum and lower quartile to the median, in relation to the corresponding distances of the maximum and upper quartile give information about the skewness of the distribution. If the maximum and upper quartile are further from the median, the distribution is skew with a long tail of higher values.

[TODO: IMAGE PLACEHOLDER]

Displaying outliers

The basic box plot is often modified to display outliers as separate crosses.

Outliers and skew distributions

Basic box plots cannot show whether the minimum and maximum in a distribution are outliers or simply the end of skew distributions.

[TODO: IMAGE PLACEHOLDER]

Clusters

Box plots cannot show clusters, so must never be used for data with clusters.

Box plots and clusters

Box plots cannot show clusters in data.

[TODO: IMAGE PLACEHOLDER]

Before using a box plot, always look at the data with a dot plot or histogram to make sure that there are no clusters.

Comparison of groups

Box plots are particularly effective for displaying differences between several groups of values.

Box plots to compare groups

To display the distribution of values in a single set of data, a dot plot or histogram is more useful than a box plot. However for comparison of two or more groups of values box plots are particularly effective — they highlight differences between the centres, spreads of values and skewness of the groups.

[TODO: IMAGE PLACEHOLDER]

Dangers of overinterpretation

Box plots are relatively stable, and contain less 'noise' than other displays. They can concisely describe differences between even small groups.

Stability of the shape of box plots

When used for small data sets, features in dot plots, stem and leaf plots and histograms are relatively unstable. Although more stable, the shapes of box plots also vary if different data are collected from the same process.

Important: Care must be taken not to over-interpret the shape of box plots for small data sets.

As with other displays, the larger the data set, the more stable the box plots become.

5. Describing Centre and Spread

Centre and Spread

The centre of a distribution is a 'typical value'. The spread describes how far the values are from the centre.

Summarising centre and spread

Two important aspects of a distribution of values are particularly important.

Centre

The centre is a 'typical' value around which the data are located.

Spread

The spread describes the distance of the individual values from the centre.

[TODO: IMAGE PLACEHOLDER]

We will describe centre and spread with numerical values called summary statistics. They provide particularly concise and meaningful comparisons of different groups.

Median, range, and IQR

The median is a summary of the centre of a distribution. The range and inter-quartile range both describe spread.

Simple summaries of centre and spread

Centre

The median is the simplest measure of centre. Half the data values are more than it, and half less.

Spread

The range (maximum - minimum) and interquartile range (upper quartile - lower quartile) are two different summary statistics that describe the spread of values in a data set.

Information from median and interquartile range

Given the median and interquartile range, it is possible to sketch a bell-shaped histogram that matches these values. Such a 'guess' is often close to the actual distribution of values.

[TODO: IMAGE PLACEHOLDER]

Summaries of centre

The median and mean are alternative measures of the centre of a distribution.

Median

Half of the data values are below the median and half are above it:

[TODO: IMAGE PLACEHOLDER]

**Mean **

The mean is:

\bar{x} = \sum x / n

If each value in a dot plot was a solid object resting on a beam with negligible mass, the mean is the value at which the beam will balance.

[TODO: IMAGE PLACEHOLDER]

Because of the leverage exerted by points far from the centre, the mean is further into the tail of a skew distribution than you might expect.

Properties of median and mean

When a data set is not symmetric, the mean and median may differ substantially.

Although both describe aspects of the 'centre' of a distribution, the median and mean are not the same and can occasionally have very different values.

Social vs economic indicator

For some data sets, the median can be considered to be a social indicator, whereas the mean can be interpreted as an economic indicator. In a company,

- the median salary indicates what the 'average employee' earns (half of the employees earn more and half earn less)
- the mean salary reflects the total amount paid as salaries in the company (it is total / n)

Outliers

An outlier has little effect on the median, but affects the mean more strongly. The median is said to be more robust.

Skew distributions

When a distribution is fairly symmetrical, the mean and median are similar, but if the distribution is skew, then the mean is usually further into the tail of the distribution than the median.

[TODO: IMAGE PLACEHOLDER]

Standard Deviation

The standard deviation is the most commonly used numerical summary of the spread of values in a data set.

Simple measures of spread

Range
Difference between maximum and minimum values
Inter-quartile range
The middle half of the values are within an interval of this length

These are (relatively) easy to understand and explain to others, but neither are commonly used.

Standard deviation

The standard deviation is a 'typical' distance of values from the sample mean.

[TODO: IMAGE PLACEHOLDER]

The standard deviation is denoted by the letter s and is defined by:

s = \sqrt{\frac{\sum(x - \bar(x))^2}{n-1}}

The numerator, $\sum\left(x-\bar{x}\right)^2$, depends on the distances of the values to the mean, so it will be small if the values are all close to the mean and big if they are far from the mean.

Variance

The square of the standard deviation, s2, is called the sample variance. Variances are sometimes reported and used but standard deviations are easier to interpret since they have the same units as the original data (e.g. kilograms or dollars).

Rule of thumb for standard deviation

The 70-95-100 rule-of-thumb is useful for understanding the numerical value of the standard deviation.

**'Quarter-range' rule of thumb **

For many data sets, the standard deviation is just under a quarter of the range.

[TODO: IMAGE PLACEHOLDER]

This is a simple rule, but is only very approximate. The standard deviation can be more than a quarter the range in distributions with short tails or much less if there are long tails or outliers.

The 70-95-100 rule of thumb

The 70-95-100 rule is more accurate. In many distributions,

- Approximately 70% of the values are within 1 standard deviation of the mean.
- Approximately 95% of the values are within 2 standard deviations of the mean.
- Nearly all of the values are within 3 standard deviations of the mean.

The 70-95-100 rule holds approximately for most reasonably symmetric data sets, but for skew data or distributions with long tails, outliers or clusters, it is often less accurate.

Understanding Means and Standard Deviation

It is possible to roughly guess the mean and standard deviation from a histogram and roughly sketch a symmetric histogram matching any given mean and standard deviation.

Understanding the definition of the standard deviation is much less important than knowing its properties and having a feel for what its numerical value tells you about the data.

Guessing s from histogram

About 95% of the values should be within 2s of the mean, so after dropping the top 2.5% and bottom 2.5% of the values (histogram area), the remainder should span approximately 4s. Dividing this range by 4 should approximate the standard deviation.

[TODO: IMAGE PLACEHOLDER]

**Sketching a histogram from the mean and s **

Similarly, you should be able to draw a rough sketch of a symmetric histogram with any mean and standard deviation that you are given. (It would be centred on the mean and 95% of the area would be within 2s of this.)

Warnings about mean and standard deviation

The mean and standard deviation cannot give any indication of the existance of outliers, skewness or clusters. A dot plot or histogram should be examined before reporting these numerical summaries.

Important: The mean and standard deviation hold no information about the shape of a distribution, other than its centre and spread.

Many different distributions have the same mean and standard deviation.

[TODO: IMAGE PLACEHOLDER]

Clusters, outliers and skewness are important features of a data set and should influence the analysis that you perform and the conclusions that you reach. In particular, if you ignore outliers or clusters, you could easily reach the wrong conclusions.

It is therefore essential that you look at a graphical display of a distribution before summarising with a mean and standard deviation.

6. More about Variation (optional)

Effect of Outliers

If a data set contains an outlier, the mean and especially the standard deviation can be badly affected. The values may be obviously wrong when the 70-95-100 rule is applied in the context of the data but examining a dot plot or box plot is best.

Outliers and the standard deviation

The mean and standard deviation are inadequate descriptions of distributions that have clusters, outliers or skewness.

An outlier has a strong influence on the mean of the data and an even bigger effect on the standard deviation. In the data below, one measurement was missing and coded as '999'. If this value (999) is included, the mean is a feasible value, but the standard deviation should be recognised as being unreasonable.

[TODO: IMAGE PLACEHOLDER]

A graphical display such as a dot plot is the best way to detect an outlier and you should always look at the data before summarising with a mean and standard deviation.

An outlier should be carefully examined. Was the value incorrectly recorded? Was there something unusual about the individual from which the measurement was obtained? If we are convinced that there was something wrong about the value, it should be removed from the data set before further analysis.

Standard Deviation of Grouped Data

The standard deviation within groups is usually lower than the overall standard deviation.

Within-group and overall standard deviation

In some data sets, the 'individuals' can be split into groups.

[TODO: IMAGE PLACEHOLDER]

When the three groups above (A, B and C) are combined, all information about the differences between the groups is lost. The overall variability is also considerably larger than variability within the groups.

Note: The standard deviation of the combined data set is often considerably higher than that of the separate groups.

It is therefore better to separately describe the distributions within the groups than to describe the overall distribution with a single mean and standard deviation.

Explained and Unexplained Variation

Splitting a data set into groups of 'similar' values results in more accurate predictions of future values if the group membership is known. The grouping is said to explain some of the overall variation.

Types of variation

The table below summarises monthly rainfall data in a city over several years:

	Rainfall	
Month	Mean	Standard deviation
January	32.13	2.11
February	31.44	2.17
March	31.24	2.08
April	30.46	1.73
May	28.53	1.69
June	26.10	1.37
July	26.43	1.32
August	30.04	1.28
September	33.44	1.24
October	34.93	1.01
November	34.34	1.49
December	32.62	1.75
Overall	30.99	3.17

We can distinguish between three types of variation in the rainfalls:

Overall variation

Ignoring the months, the overall standard deviation is 3.17.

Unexplained variation

Variation within months is unexplained — it is unpredictable from available information. The standard deviations within months are between 1.01 and 2.17.

Explained variation

This is the difference between the overall and unexplained variation. (We do not give a numerical definition here.) Knowing the month would help to predict rainfall, so the month explains part of the variation in rainfalls.

Variance and Degrees of Freedom (advanced)

The square of the standard deviation is called the variance; its value is harder to understand but it is the basis of important advanced statistical methods. The degrees of freedom are the number of pieces of information contributing to the standard deviation (or variance).

**Variance **

variance = (standard deviation)2 = $\frac{\sum\left(x - \overline{x}\right)^2}{n - 1}$

The units of the variance are the square of the units of the original values. For example, if the values are weights, the standard deviation might be 6 kg, but the variance would be 36 square kg. Since its units are easier to interpret, standard deviations are more easily understood measures of spread, but variances are important in advanced statistics. (An important collection of methods for analysing relationships between variables is called analysis of variance.)

**Degrees of freedom (optional) **

The divisor (n − 1) in the formula for the sample standard deviation is called its degrees of freedom. This is the number of 'independent pieces of information' that contribute to it.

Sample of size n = 1

With only a single value, there is no information about the spread of values, so there are 0 degrees of freedom.

Sample of size n = 2

With two values, x1 and x2, there is only a single piece of information about the spread — the difference between the values, x1 − x2 — and there is one degree of freedom.

Sample of size n

In general, there is one less 'piece of information about the spread' in the sample than the number of data points because the sample mean, $\bar{x}$, is one piece of information that does not give any information about the spread of the data. There are therefore (n − 1) degrees of freedom.

Root Mean Squared Error (advanced)

The root mean squared error summarises how close the values in a data set are to a target, k.

Distance of values from a target, k

The distance of a single random value from a target is called its error.

[TODO: IMAGE PLACEHOLDER]

Root mean squared error

One solution to the problem of negative errors is to square them before averaging, mean squared error $= \dfrac{\sum\left(x-k\right)^2}{n}$

To express this in the original units of the data (instead of units such as squared kg), we can take its square root:

root mean squared error $= \sqrt{\dfrac{\sum\left(x-k\right)^2}{n}}$

The root mean squared error is a 'typical' error.

Distances from the mean (advanced)

The standard deviation is similar to the root mean squared error, but summarises distances to the mean of the data. Its value can be interpreted in terms of the average area of squares on a graph.

Distances from the centre of the distribution

The population standard deviation is similar to the root mean square error but summarises the distances of the values from the centre of their distribution. It summarises the spread of values in the data.

population standard deviation $= \sqrt{\dfrac{\sum\left(x-\overline{x}\right)^2}{n}}$

This can be illustrated graphically — the squared standard deviation is the average of the squared distances of values to their mean:

[TODO: IMAGE PLACEHOLDER]

Standard deviations in reports are likely to be sample standard deviation.

7. Proportions and Percentiles

Illustrative Data Set

A data set containing annual rainfalls in Samaru, Nigeria, will be used for illustrative purposes.

Annual rainfall in Samaru, Nigeria

In most of Africa, the most important climatic variable is rainfall. Rainfall is usually highly seasonal and failure of crops is normally associated with late arrival of rain or low rainfall. A better understanding of the distribution of rainfall can affect the crops that are grown and when they are planted.

The annual rainfall (in mm) in Samaru, Northern Nigeria between 1928 and 1983 will be used as an example in this section.

[TODO: IMAGE PLACEHOLDER]

Cumulative Proportions

Half the data are lower than the median. A quarter and three quarters are lower than the lower and upper quartiles. At any other value, x, the proportion of data values that are x or lower is called its cumulative proportion.

Cumulative proportions

The proportion of values in the data set that are less than or equal to any value, x, is called its cumulative proportion.

For the median and quartiles, the cumulative proportions are:

Value	Proportion below
Lower quartile	0.25
Median	0.5
Upper quartile	0.75

The proportion of values greater than x is one minus its cumulative proportion,

Pr(values > x) = 1 - Pr(values <= x)

Equality

For continuous data, we do not need to distinguish between the proportion of values less than x and the proportion that are less than or equal to x. Provided the values are recorded accurately enough,

- these two proportions are usually equal for most x of interest, and
- since the same value rarely appears more than once, the difference is unlikely to be more than 1/n.

However for discrete data (counts) it is important to distinguish the terms 'less than' and 'less than or equal to'.

Graph of Cumulative Proportions

A graph of the cumulative proportion below x against x is a step function that increases from zero (at small x) to one (at high x).

Cumulative distribution function

The cumulative proportion of values less than or equal to x can be found for any x. They can be shown together in a single graph of the cumulative proportion against x. This is called the cumulative distribution function of the variable.

[TODO: IMAGE PLACEHOLDER]

The cumulative distribution function for a data set with n values is a step function that rises from 0.0 below the minimum x-value to 1.0 at the maximum x in the data. It increases by 1/n at each value in the data set.

Percentiles

Given any target proportion, p, it is possible to find a corresponding value, x, for which approximately this proportion of values is x or lower. For example, the percentile for p = 50% is the median.

[TODO: IMAGE PLACEHOLDER]

Finding percentiles

Given any proportion, p, between 0 and 1, we can find a value x such that approximately this proportion, p, of values is x or lower in our data set. This is called the p'th quantile in the data set. When p is given as a percentage, the same value is called the p'th percentile.

Important: The p'th percentile is the value x such that p percent of the data set are x or lower.

Percentiles can be read from a graph of the cumulative distribution function.

[TODO: IMAGE PLACEHOLDER]

Details (optional)

It may not be possible to find a value, x, such that exactly p percent of the data are lower, expecially if the sample size is not a multiple of 100. If n = 56, the cumulative distribution function is a step function that rises by 1/56 at each data value, so it is impossible to find an x-value for which exactly say 43% of values are lower.

There is no universally accepted general definition of percentiles and different statistical programs give slightly different values. The differences are minor and should not affect your interpretation of the data.

Displaying Percentiles

The 0, 25, 50, 75 and 100'th percentiles are displayed as a box plot. Other percentiles can be displayed in a similar shaded rectangle.

25, 50 and 75% percentiles

The 50th percentile is the median and the 25th and 75th percentiles are the lower and upper quartiles. A box plot therefore shows these percentiles.

[TODO: IMAGE PLACEHOLDER]

Displaying other percentiles

For some data sets, other percentiles are more important than the 25th and 75th ones. A similar 'box' can be used to graphically display any other percentiles. It is best to alter the way the box is drawn to avoid confusion with the standard box plot.

[TODO: IMAGE PLACEHOLDER]

Comparing Groups

Box plots are useful for comparing groups. If the groups are in order (e.g. the months of a year), the median, quartiles and extremes can be joined and shaded as bands. This effectively describes how the distribution of values varies.

Joined-up quartiles

Box plots are an effective way to compare the distributions of different groups of values. When the groups are ordered, an alternative to the conventional display of the box plots is to join up the medians, quartiles and extremes of the groups in shaded bands.

[TODO: IMAGE PLACEHOLDER]

Comparing Groups with Other Percentiles

In some applications, different percentiles are important. They can also be joined and shaded as bands to compare ordered groups.

Joined-up percentiles

A similar display can be used with other percentiles.

[TODO: IMAGE PLACEHOLDER]

Better Definition of Percentiles

The graph of cumulative probabilities is a step function. Most software reports percentiles that are equivalent to reading values off a smoothed version of this step function.

Different definitions of percentiles

It was mentioned earlier that there are several competing definitions of the upper and lower quartile. All such definitions split the data approximately into quarters but there is not a unique way to do this.

There is even less agreement about the precise definition of other percentiles, and different computer software finds them in different ways. The definitions are usually based on a smoothed version of the cumulative distribution function.

[TODO: IMAGE PLACEHOLDER]

The differences between the different definitions are small if the data set is large.

Important: If your conclusion about the data would change with a different definition of the percentiles, you are over-interpreting the data.

8. Transformations

Linear Transformations

Linear transformations of data affect the scale on the axis of graphical displays, but do not otherwise change the shape of the distribution of values.

**Linear transformations **

When the values are replaced by other using an equation of the form

new value = a + b × old value

we say that there has been a linear transformation of the original values. The original and transformed data can be displayed together with dual axes.

[TODO: IMAGE PLACEHOLDER]

**Centre and spread **

The centre and spread of the data are different, but the shape of the distribution otherwise remains unchanged. The mean and standard deviation are related:

new mean = a + b × old mean

new sd = |b| × old sd

Note that if the scale factor, b, is negative, we must change its sign since the standard deviation must always be positive.

Most other measures of centre (e.g. the median) and spread (e.g. the interquartile range) are similarly related.

Log Transformations

Nonlinear transformations change the shape of the distribution of values more profoundly. A logarithmic transformation can help detect patterns in very skew data sets.

Nonlinear transformations

Nonlinear transformations arise when the values are replaced by a nonlinear function of the original measurements, such as their logarithm or inverse. They have a more fundamental effect on the shape of a distribution than linear transformations.

The most commonly used nonlinear transformation is:

new value   = log10 (old value)

Natural logarithms (base e) have a similar effect on the distribution of values but base-10 logarithms are easier to interpret so we use them here.

Properties of logarithms

- Multiplying any value by 10 increases its logarithm by 1.
- Doubling any value increases its logarithm by log10(2) = 0.3010.

Consider four values 1, 10, 100 and 1000. The first two values are much closer to each other than the last two values. However their logarithms are 0, 1, 2 and 3, so their logarithms are evenly spaced out.

Effect on the shape of a distribution

A logarithmic transformation selectively spreads out low values in a distribution and compresses high values. It is therefore useful before analysing skew data with a long tail towards the high values. It will spread out a dense cluster of low values and may detect clustering or outliers that would not be visible in graphical displays of the original data.

[TODO: IMAGE PLACEHOLDER]

When to use a Log Transformation?

Logarithmic transformations are most useful for 'quantity' data that cover several orders of magnitude.

'Quantities'

Logarithmic transformation can only be used for data sets consisting of positive values — logarithms are undefined for negative or zero values. They are therefore particularly useful for quantities — i.e. amounts of something. Indeed, many researchers routinely apply logarithmic transformation to quantity data before analysis.

When are they effective?

A log transformation affects the shape of the distribution most when the ratio of the largest to the smallest value in the data is large. When this ratio is less than 10 (one order of magnitude) then the transformation has much less influence on the shape of the distribution, as in the data set below.

[TODO: IMAGE PLACEHOLDER]

Power Transformations (advanced)

Power transformations are a more flexible family of nonlinear transformations that are useful in data exploration.

Power transformations

A more general family of transformations that is flexible enough to reduce or eliminate the skewness in a wide range of data sets is:

[TODO: IMAGE PLACEHOLDER]

This family of power transformations includes many common ones:

[TODO: IMAGE PLACEHOLDER]

Power Transforms and Skewness

The effect of power transformations on the skewness of data is evident in a wide range of graphical displays.

Effect of power transformations

Power transformations affect the skewness of data.

If a power transformation with $p > 1$ is applied to data with a symmetric distribution, it will make the data skew with a long right tail. If the power transformation has $p < 1$, the distribution will become one with a long left tail.

In practice, power transformations are used to do the opposite. They can change many skewness distributions into fairly symmetric ones.

9. Discrete Data (counts)

Discrete and Continuous Data

Discrete data sets contain counts whereas continuous data sets could potentially contain any values within an interval. Stacked dot plots are good displays of small discrete data sets containing small counts.

It is important to distinguish two types of numerical data.

Discrete data

When the values in the batch are whole numbers (counts), the data set is called discrete.

Continuous data

When the data are not constrained to be whole numbers, the data set is called continuous.

Dot plots for counts

Dot plots can be used to display count data. However since discrete values are often repeated several times in a data set, the crosses need to be jittered or, preferably, stacked.

[TODO: IMAGE PLACEHOLDER]

If there is a stack for each integer value, the stacked dot plot is a complete representation of the data.

Histograms for Counts

When the range of possible counts is moderate or large, a histogram is an effective display of the distribution. Class width should be a whole number and class boundaries should end in '.5'.

Displaying moderate or large counts

For discrete data sets whose values are large counts, a histogram can be used to give a 'smooth' summary of the shape of the distribution of values.

If the counts are a bit smaller, the exact definition of the histogram classes becomes important. The class boundaries should end in '.5' to ensure that data values do not occur on the boundary of two classes.

[TODO: IMAGE PLACEHOLDER]

Bar Charts

When the range of possible counts is small, a bar chart is a better representation of the data than a histogram.

Displaying small counts

When the range of values in a discrete data set is small, a histogram can be drawn with class width 1 (and with class boundaries ending in '.5'). These classes are centred on 1, 2, 3, etc.

This can be improved by narrowing the histogram rectangles into bars to emphasise the discrete nature of the data. This is called a bar chart.

[TODO: IMAGE PLACEHOLDER]

Mean and Standard Deviation (advanced)

A frequency table is often used to summarise discrete data. The mean and standard deviation can be evaluated easily from the frequency table.

Calculating the mean from a frequency table

x	ƒx
1	140
2	180
3	60
4	100
5	60
6	40
7	20
total	600

The mean can be easily calculated from this table:

[TODO: IMAGE PLACEHOLDER]

More generally,

\bar{x} = \frac{\sum(x) \times f_x}{n}

where the summation is over the distinct values in the data set, rather than all individuals.

Calculating the standard deviation

A similar formula holds for the standard deviation, using the formula

[TODO: IMAGE PLACEHOLDER]