Glossary of statistical terms

Roger Stern, Ian Dale and Sandro Leidi

Index of entries

Absolute value (modulus)

The absolute value is the value of a number, disregarding its sign. It is denoted by a pair of "|" signs. For example the modulus of –2.5 is |-2.5| = 2.5. See also *mean deviation*.

Alternative hypothesis

The alternative *hypothesis*, H_1, is a statement of what the test is set up to establish. For example if comparing average annual rainfall in El Nino and ordinary years then we could have:

- H_0,that the two means are not equal, i.e. there is a difference between the two types of year..

The conclusion from the hypothesis test is either "Reject H_0 in favour of H_1." or "Do not reject H_0." If H_0 is rejected then the analysis usually continues by establishing, in this case, the extent of the difference between the two types of year.

Average

For a numeric variable the average is a loosely used term for a measure of location. It is usually taken to be the *mean*, but it can also denote the *median*, the mode, among other things.

Assumptions

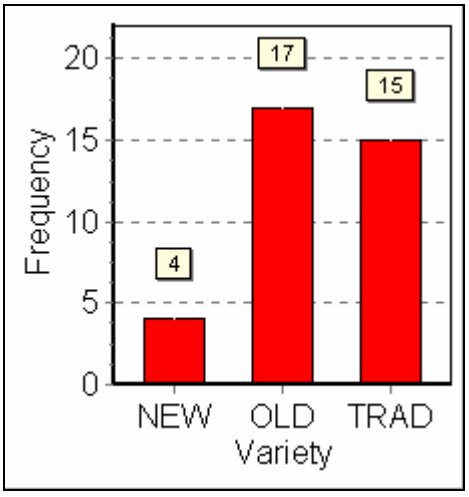
Statistical inference usually involves using a *sample* to *estimate* the *parameters* of a *model*. The conclusions, i.e. the validity of the estimates, only hold if certain assumptions are true.

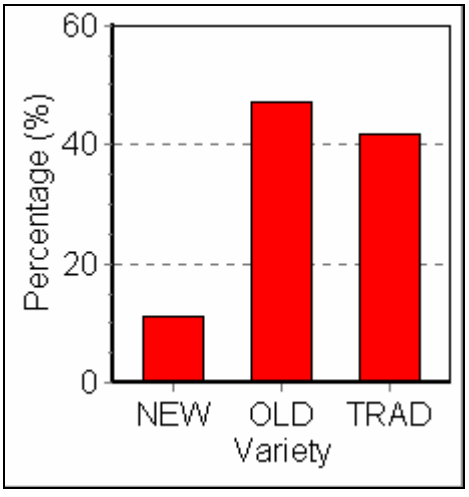
For example a sample of 50 years of rainfall data may be used to estimate the parameters of a *normal* model. The assumptions are then that:

- The 50 years behave like a random sample. (They are not random, they are 50 successive years)
- The data are from a single *population*, i.e. there is no climate change
- The population has a normal distribution.

Bar chart

A diagram for showing the *frequencies* of a *variable* that is *categorical* or *discrete*. The lengths of the bars are proportional to the frequencies or the *percentages*. The widths of the bars should be equal.





Bias

See the entry under *precision*.

Binomial distribution

The binomial distribution is used to model data from *categorical variables*, when there are just 2 categories, or *levels*. For example suppose we wish to *estimate* the chance of rain at a given site at the start of April. We have data for 10 days each year, and for 30 years. So there are n = 300 days, out of which, say, r = 90 were rainy and 210 were dry. Then the observed chance of rain, p = r/n = 90/300 = 0.3. So we observed 30% of the days had rain.

The binomial model has a single *parameter*, θ, the true chance of rain. Hence we can estimate θ by p = 0.3 or 30% and then use the properties of the binomial model to give the *standard error* and *confidence interval* for the estimate. The result, using a statistical package is that the standard error is 0.026 and the 95% confidence interval for the chance of rain is 0.25 to 0.36. So the true chance of rain at this time, is likely to be somewhere between 25% and 36%.

These confidence limits depend on the assumptions of the binomial model, and they are as follows:

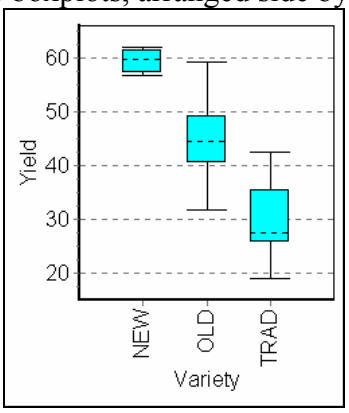
- Each day has an equal chance of rain
- This chance is independent of the previous day

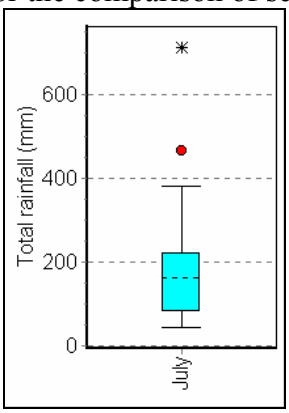
In this example, one problem might be the assumptions that the successive days in the same year are independent. In many countries rainy days follow each other.

If the assumptions are not true, then a different model must be used.

Boxplot (or "box and whisker" plot)

A graphical representation of numerical data, based on the *five-number summary* and introduced by *Tukey* in 1970. The diagram has a scale in one direction only. A rectangular box is drawn, extending from the lower *quartile* to the upper quartile, with the *median* shown dividing the box. 'Whiskers' are then drawn extending from the end of the box to the greatest and least values. Multiple boxplots, arranged side by side, can be used for the comparison of several samples.





In **refined boxplots** the whiskers have a length not exceeding 1.5 times the *interquartile range*. Any values beyond the ends of the whiskers are shown individually as *outliers*. Sometimes any values further than 3 times the interquartile range are indicated with a different symbol as extreme outliers

Categorical variable

A variable with values that range over categories, rather than being numerical. Examples include gender (male, female), paint colour (red, white, blue), type of animal (elephant, leopard, lion). Some categorical variables are *ordinal*.

Central Limit Theorem

This result explains why the *normal distribution* is so important in statistics.

Often we want to use the sample *mean* $\bar{x}$ to *estimate* the mean, μ, of the population. The central limit theorem says that, as long as the sample size is reasonably large, the distribution of $\bar{x}$ about μ will be roughly normal, whatever the distribution of the data. How amazing – but true!

Confidence interval

A confidence interval gives an estimated range of values that is likely to include an unknown *population* parameter.

For example suppose a study of planting dates for maize, and the interest is in estimating the upper *quartile*, i.e. the date by which a farmer will be able to plant in ¾ of the years. Suppose the estimate from the sample is day 332, i.e. 27th November and the 95% confidence interval is from day 325 to 339, i.e. 20th November to 4th December. Then the interpretation is that the true upper quartile is highly likely to be within this period.

The width of the confidence interval gives an idea of how uncertain we are about the unknown *parameter* (see *precision*). A very wide interval (in the example it is ± 7 days) may indicate that more data needs to be collected before an effective analysis can be undertaken.

Continuous variable

A numeric variable is continuous if the observations may take any value within an interval. Variables such as height, weight and temperature are continuous.

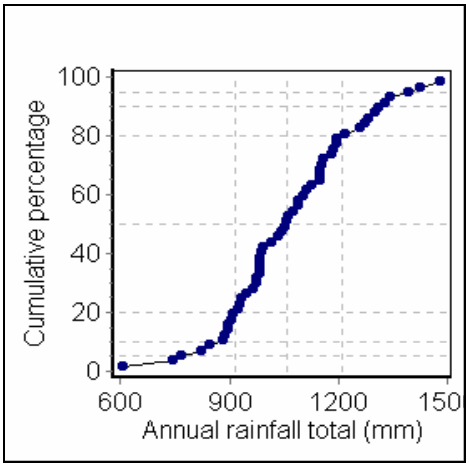
In *descriptive statistics* the distinction between *discrete* and continuous variables is not very important. The same summary measures, like mean, median and standard deviation can be used.

There is often a bigger difference once *inferential* methods are used in the analysis. The *model* that is assumed to generate a discrete variable is different to models that are appropriate for a continuous variable. Hence different *parameters* are *estimated* and used.

(See also *discrete variable*, *mixed variable*.)

Cumulative frequency graph

For a numerical variable, the cumulative frequency corresponding to a number x is the total number of observations that are less than or equal to x.



The y-axis of this graph can show the *frequency*, the *proportion* or the *percentage*.

With the percentage, this graph allows any *percentile* to be read from the graph. For example in the graph above, the 25% point (lower *quartile*) is about 940mm. The 90% point is about 1300mm.

Decile

Deciles are used to divide a numeric variable into 10ths, whereas the quartiles divide it into quarters, and percentiles into 100ths. An approximate value for the rth decile can be read from a

cumulative frequency graph as the value corresponding to a cumulative relative frequency of 10r%. So the 5th decile is the *median* and the 2nd decile is the 20th *percentile*. The term decile was introduced by *Galton* in 1882.

Descriptive statistics

If you have a large set of data, then descriptive statistics provides graphical (e.g. *boxplots*) and numerical (e.g. summary tables, means, *quartiles*) ways to make sense of the data. The branch of statistics devoted to the exploration, summary and presentation of data is called descriptive statistics.

If you need to do more than descriptive summaries and presentations it is to use the data to make inferences about some larger population. *Inferential statistics* is the branch of statistics devoted to making generalizations.

Discrete variable

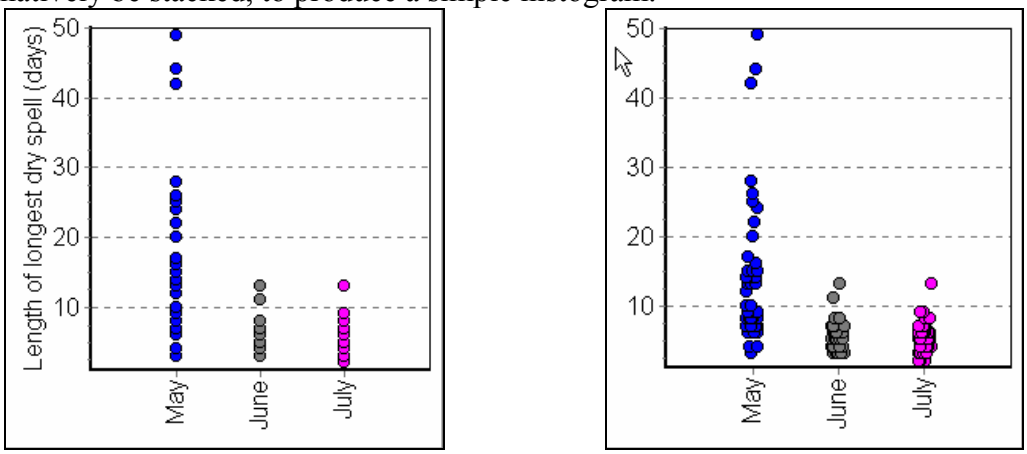
A set of data is discrete if the values belonging to it are distinct, i.e. they can be counted. Examples are the number of children in a family, the number of rain days in the month, the length (in days) of the longest dry spell in the growing season. (See also *continuous variable* for a more complete discussion.)

Dispersion

See *variability*

Dot plot

A dot-plot is an alternative to a *boxplot* where each value is recorded as a dot. It is used when there are only few data values. The dots may be *jittered*, so each value is made visible. They may alternatively be stacked, to produce a simple histogram.



Estimation

Estimation is the process by which *sample data* are used to indicate the value of an unknown quantity in a *population*.

The results of estimation can be expressed as a single value, known as a point estimate. It is usual to also give a measure of precision of the estimate. This is called the *standard error* of the estimate.

A range of values, known as a *confidence interval* can also be given.

Estimator

An estimator is a quantity calculated from the *sample data*, which is used to give information about an unknown quantity (usually a *parameter*) in the *population*. For example, the sample *mean* is an estimator of the population mean.

Estimators of population parameters are sometimes distinguished from the true (but unknown) population value, by using the symbol "hat". For example

$\hat{\mu} = \overline{x}$ is used to estimate the population mean, μ.

$\hat{p}_{10} = (\overline{x} - 1.28s)$ is used to estimate the 10% point, or the 10 year return period

If $\overline{x} = 210$ mm and s = 45mm then the population mean is estimated to be 210mm and the 10% point is estimated to be 152mm.

Five-number summary

For a numeric variable, the least value (*minimum*), the lower *quartile*, the *median*, the upper quartile, and the greatest value (*maximum*), in that order. These are shown graphically in a *boxplot*.

For the following data (when put into ascending order), the five numbers are shown.

Data	11	12	13	14	15	18	18
Summaries	Minimum	Lower quartile		Median		Upper quartile	Maximum

Factor

Another word for *categorical*.

Fisher Sir Ronald Aylmer (1890-1962).

Fisher was an English statistician; arguably the most influential statistician of the twentieth century. Fisher was educated at Harrow and at Cambridge University, where he studied mathematics. His initial interest in statistics developed because of his interest in genetics, and he pursued both subjects for the rest of his life. Fisher's first paper introduced the method of maximum likelihood, his second the mathematical derivation of the t-distribution, his third the distribution of the correlation coefficient.

Following the First World War (which saw Fisher employed as a teacher because his dreadful eyesight prevented him from fighting), he joined the agricultural research station at Rothamsted. During his time there he virtually invented the subjects of experimental design and ANOVA, which motivated his derivation of the F-distribution. In 1925 the first edition of his *Statistical Methods for Research Workers* appeared. The extent of the influence of this work can be gauged from the fact that there was a new edition about every three years until 1958, and a posthumous fourteenth edition in 1970.

At a conference in India he said 'To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.'

Frequencies

The frequency is the number of times that particular values are obtained in a variable. For example, with the data:

dry, rain, dry, dry, dry, rain, rain, dry,

the frequency of "rain" is 3 and the frequency of "dry" is 5. The sum of the frequencies is the number of observations, n, in the variable. In this example, n = 8. The percentage is (100*frequency/n), = 100* 3/8 = 37.5% for rain, in this example. With as few as 8 values *percentages* are not usually used, instead state 3 out of the 8 values were rain.

Galton (Sir Francis 1822 – 1911)

An English doctor, explorer, meteorologist and statistician. Galton was also first cousin of Charles Darwin, the author of The Origin of Species. Galton studied medicine at Cambridge University. On coming into money he abandoned this career and spent the period 1850-1852 exploring Africa. In the 1860's he turned to meteorology and devised an early form of the weather maps used by modern meteorologists. He coined the term "anticyclone". Subsequently, perhaps inspired by Darwin's work, Galton turned to inheritance and in his 1869 book *Hereditary Genius* he used the term correlation in its statistical sense. His best-known work, published in 1889, was entitled Natural Inheritance and included great use of the *normal distribution*. He is quoted as saying "whenever you can, count."

Greek letters

In statistics, Greek letters are used for the parameters of the population and for a few other things. The table below gives the pronunciation of those that are most used.

Greek letter	Pronounced	Uses
α	Alpha	Significance tests, Intercept in regression line $y=\alpha + \beta x$
β	Beta	Significance tests, Slope in regression line $y=\alpha + \beta x$
χ	Kai	χ^2 test is a much (over)used significance test
μ	Mew	Mean of the population
π	Pai	The number 3.14
θ	Theta	Also sometimes used for the population proportion Sometimes used for the population proportion
ρ	Rho	Correlation coefficient of the population
σ	Sigma	Standard deviation of the population σ^2 is the variance of the population
Σ	Sigma	Shorthand for "The sum of"

Hypothesis Test

Testing hypotheses is a common part of statistical inference. To formulate a test, the question of interest is simplified into two competing hypotheses, between which we have a choice. The first is the *null hypothesis*, denoted by H_0, against the *alternative hypothesis*, denoted by H_1 .

For example with 50 years of annual rainfall totals a hypothesis test could be whether the mean is different in El Nino and Ordinary years. Then usually

- The null hypothesis, H_0, is that the two means are equal, i.e. there is no difference.
- The alternative hypothesis, H_1, is that the two means are unequal, i.e. there is a difference.

If the 50 years were considered as being of three types, El Nino, Ordinary, La Nina then usually:

- The null hypothesis, H_0, is that all three means are equal.
- The alternative hypothesis, H_1, is that there is a difference somewhere between the means.

The hypotheses are often statements about population parameters. In the first example above it might be:

- H_0, is that $\mu_E = \mu_O$.
- H_1, is that $\mu_E \neq \mu_O$.

The outcome of a hypothesis test is either

- Reject H_0 in favour of H_1, or
- Do not reject H_0.

Inference

Inference is the process of deducing properties of the underlying distribution or *population*, by analysis of data. It is the process of making generalizations from the sample to a population.

Inter-quartile range

The interquartile range is the difference between the upper and lower *quartiles*. If the lower and upper quartiles are denoted by Q1 and Q3, respectively, the interquartile range is (Q3 - Q1). The phrase 'interquartile range' was first used by *Galton* in 1882.

Jittered dot-plot

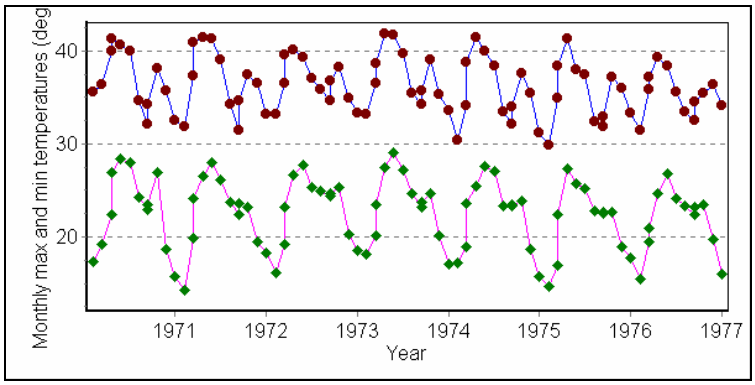
In a simple dot plot the dots may overlap. In some cases they may coincide completely, so obscuring some of the points.

A solution is to randomly move the dots perpendicularly from the axis, to separate them from one another. This is called jittering. It results in a jittered dot-plot. See *dot-plot* for an example

Line graph

A line graph is a *scatter plot* where individual points are connected by a line. The line represents a sequence in time, space, or some other quantity.

Where the graph also includes a *category* variable, a separate line may be drawn for each *level* of this variable.



Levels

The levels are the number of categories in a *categorical* (factor) variable. The categorical variable for gender (male, female) has 2 levels. There are 5 levels in the variable with the categories (Very bad, Bad, Middling, Good Very good).

Maximum

The maximum is the highest value in a *numerical variable*. In the variable with the values

12, 15, 11, 18, 13, 14, 18

then 18 is the maximum value.

Mean

The mean is a measure of the "middle", sometimes called the "average". It is often given the symbol $\bar{x}$.

To calculate the mean, sum all the data values, and divide by the number of them. For example with the 7 values 12, 15, 11, 18, 13, 14, 18, the mean is:

$$\bar{x} = (12+15+11+18+13+14+18)/7 = 101/7 = 14.4$$

As a formula, the mean is given by $\bar{x} = \sum x/n$,

where $\sum$ is short for "the sum of", "x" signifies that each value is taken in turn, and *n* is the number of observations.

Mean deviation

The mean deviation is a measure of *spread*. It is given by the formula:

$mean\ deviation = \sum |x-\bar{x}|/(n-1)$, where the | symbol is to take the *absolute value*, see the example below.

The calculation is shown in the Table below:

Data values	$(x-\bar{x})$	$\lvert x-\bar{x}\rvert$	
12	-2.4	2.4	
15	0.6	0.6	
11	-3.4	3.4	
18	3.6	3.6	
13	-1.4	1.4	
14	-0.4	0.4	
18	3.6	3.6	
	Sum = 15.4		mean deviation = 15.4/6 = 2.57

The mean deviation is, as is shown by the formula, an average (mean) difference from the mean. The value it gives is similar, but slightly smaller than the *standard deviation*.

Although it is intuitively simpler than the standard deviation it is used less. The reason is largely because the standard deviation, s, is used in inference, because the *population* standard deviation, σ, is one of the *parameters* of the *normal distribution*.

Median

The median is the "middle value" of a list. If the list has an odd number of entries, the median is the middle entry after sorting the list into increasing order. If the list has an even number of entries, the median is halfway between the two middle numbers after sorting.

For example with the same 7 values shown for the mean and maximum above the sorted data are as follows:

11 12 13 14 15 18 18.

The median is therefore the 4th value in the sorted list, i.e. 14.

Minimum

The maximum is the lowest value in a *numerical variable*. In the variable with the values

12, 15, 11, 18, 13, 14, 18

then 11 is the minimum value.

Mixed variable

Some variables are between being *categorical* and *numerical*. For example daily rainfall is exactly zero on all dry days, but is a continuous variable on rainy days. Wind speed is similar, with zero being calm. Similarly crop yield of maize is zero for all the farmers who did not grow maize, or for whom the crop failed.

Often there is a single *"special value", here zero*, and otherwise the variable is continuous. This is not always the case. For example sunshine hours expressed as a fraction of the day length, is zero on cloudy days and 1 (or 100%) on days with no cloud (or haze).

In the analysis it is usual to treat the categorical and the numerical parts separately. For example consider 10 rainy days with the data as follows

0 0 21 9 0 0 5.3 0 2.7 0

The total rainfall is 38mm and this may be a useful summary. When considering the mean rain per day this is 38/10 = 3.8 mm per day. However, categorical variables are usually summarized using frequencies and percentages. Hence if the "mixed" nature of the data is included, then the analysis could be in two parts, as follows.

- "Four out of the ten days had rain." If more need be said, then "The frequency of rain = 4/10 = 0.4 (or 40%)", but take care with the misleading impression given by percentages with such a small data set.
- On those rainy days the mean rain per rain day was 38/4 = 9.5mm per rain day.

This is sometimes called a conditional analysis.

Model (statistical)

The word "model" is used in many ways and means different tings, depending on the discipline. For example a meteorologist might think of a global climate model, used for weather forecasting, while an agronomist might think of a crop simulation model, used to estimate crop growth and yields.
Statistical models form the bedrock of data analysis. A statistical model is a simple description of a process that may have given rise to observed data.
For example, suppose the data were the presence or absence of rain on 1st April each year for a given site, for 50 years:

Year	1	2	3	4	…	…	49	50	Total
Rain (no = 0, yes = 1)	1	0	0	1			0	1	18

A simple model might be to assume that the chance of rain each year, was independent of the previous year and it had the same value. These assumptions (same value each year, and independence) would imply that the frequency of rain on 1st April could be modeled using a *binomial distribution*.
There are many probability distributions that are key parts of models in statistics, including the *normal* distribution and the binomial distribution.

Normal distribution

The normal distribution is used to model some *continuous* variables. It is a symmetrical bell shaped curve that is completely determined by two *parameters*. They are the distribution (or population) mean, μ, and the standard deviation, σ.

Thus, once the mean and standard deviation are provided, it is possible to calculate any percentile (or risk) of the distribution. For example for a normal distribution

- The lower quartile, or 25% point is (μ - 0.68σ)

- The 10% point is (μ - 1.28σ)
- The 1% point is (μ - 2.33σ), so by symmetry the 99% point is (μ + 2.33σ)

Also

- Roughly 70% of the values are within 1 standard deviation of the mean, i.e. between (μ - σ) and (μ + σ)
- Roughly 95% of the values lie within 2 standard deviations of the mean, i.e. between (μ - 2σ) and (μ + 2σ)

This is the origin of the " 70%, 95%, 100% rule of thumb" that is used to help interpretation of the *sample standard deviation*, s.

The real reason that the sample *mean*, $\bar{x}$ and sample *standard deviation*, s, are so important is because as well as being simple summaries of *average* and *spread*, they can also be used to estimate the parameters of the normal distribution.

In addition, the *central-limit theorem* justifies the use of the methods of *inference* developed for data from a normal model (and hence also the use of the sample mean and standard deviation), even when the raw data are not normally distributed.

Null hypothesis

The null hypothesis, H_0, represents a theory that has been put forward, usually as a basis for argument. For example if comparing average annual rainfall in El Nino and ordinary years then we could have:

- H_0, that the two means are equal, i.e. there is no difference between the two types of year..

The null hypothesis is usually simpler than the alternative hypothesis and is given special consideration. Thus the conclusion is given in terms of the null hypothesis. We either "Reject H_0 in favour of H_1." or "Do not reject H_0." If we reject H_0, then we declare the result to be "statistically significant", and this provides evidence that H_1 is true

If we conclude "Do not reject H_0.", then the result is declared to be "not statistically significant". This does not necessarily mean that H_0 is true, only that we do not have sufficient evidence to reject it.

Numerical variable

Refers to a variable whose possible values are numbers (as opposed to *categories*).

Ordinal variable

An ordinal variable is a *categorical variable* in which the categories have an obvious order, *e.g.* (strongly disagree, disagree, neutral, agree, strongly agree), or (dry, trace, light rain, heavy rain).

Outlier

An outlier is an observation that is very different to other observations in a set of data. Since the most common cause is recording error, it is sensible to search for outliers (by means of summary statistics and plots of the data) before conducting any detailed statistical modelling.
Various indicators are commonly used to identify outliers. One is that an observation has a value that is more than 2.5 *standard deviations* from the mean. Another indicator is an observation with a value more than 1.5 times the *interquartile* range beyond the upper or the lower *quartile* *(see boxplot*).
It is sometimes tempting to discard outliers, but this is imprudent unless the cause of the outlier can be identified, and the outlier is determined to be spurious. Otherwise, discarding outliers can cause one to underestimate the true variability of the data.

P – value

The probability value (p-value) of a *hypothesis test* is the probability of getting a value of the test statistic as extreme, or more extreme, than the one observed, if the null hypothesis is true.

Small p-values suggest the *null hypothesis* is unlikely to be true. The smaller it is, the more convincing is the evidence to reject the null hypothesis.

For example suppose the annual rainfall totals are being used to test the null hypothesis that the average is the same in El Nino and Ordinary years, and we find that the test statistic $(\bar{x}_E - \bar{x}_O)$ = 100mm resulting (once the spread of the data is given) in a p-value of 0.003.

This small probability value implies that if H_0 is true i.e. $(\mu_E - \mu_O) = 0$, then the data we have occurs vary rarely, i.e. only 3 times in 1000. This low p-value is used to reject H_0.

If, on the other hand, the p-value were found to be 0.2, (or 0.4 or 0.6) then it implies that the data and the null hypothesis are compatible. Thus the data do not provide evidence to reject the null hypothesis.

In the pre-computer era it was common to select a particular p-value, (often 0.05 or 5%) and reject H_0 if (and only if) the calculated probability was less than this *fixed value*. Now it is much more common to calculate the exact p-value and interpret the data accordingly.

Parameter

A parameter is a numerical value of a *population*, such as the population mean. The population values are often modelled from a distribution. Then the shape of the distribution depends on its *parameters*. For example the parameters of the *normal distribution* are the mean, μ and the standard deviation, σ. For the *binomial distribution*, the parameters are the number of trials, n, and the probability of success, θ.

Pattern

A good statistical analysis is one that takes account of all the "pattern" in the data. In *inference* this can be expressed or "modelled" as

data = pattern + residual

The idea is that the data has *variability*, i.e. the values differ from each other. Some of the variability can be understood, and it therefore is part of the "pattern" or "signal" in the data. What is left over is not understood and is called the residual, (or "noise" or "error"). A good analysis is one that explains as much as possible. Hence if you can still see any patterns in the residual part, then consider how it can be moved over into the pattern (or model).

Even with a *descriptive statistics*, it is important that the analysis reflects the possible patterns in the data. At least the obvious patterns in the data should be considered when doing the analysis.

As an example, consider 2 variables, that give the date of planting for two farmers, called "Bold" and "Cautious". There are 65 observations for each variable, giving their date from 1936 to 2000. First consider the column "Bold". Obvious sources of "pattern" are:

- **Climate change** – is there any indication of a trend in the data; perhaps recent years tend to have later planting dates?
- **El Nino** effect– Is there any evidence of different planting dates in El Nino years; perhaps La Nina years tend to have earlier planting date?
- **Outliers –** Are there any obvious *outliers*, that can be explained; perhaps very late planting in a particular year is found to be due to missing values that were mistakenly coded as zero.

In the analysis to compare the two variables, we assume they were the same years for both.

- Hence consider analysing the difference in planting dates. This is usually better than analysing the two variables separately, and then comparing the summary values.
- When processing the new variable (the differences), you may find that some years the difference is zero, i.e. the farmers plant on the same day. This is a *mixed variable* and analysing the years with zero difference separately from the other years is usually helpful.

Percentage

For a variable with n observations, of which the frequency of a particular characteristic is r, the percentage is 100*r/n. For example if the frequency of replanting was 11 times in 55 years, then the percentage was 100*11/55 = 20% of the years.
Percentages are widely used (and misused). Whenever percentages are used it must be made clear what is the 100%. In the example above it was the value 55.

Percentile

The *p*th percentile of a list is the number such that at least *p*% of the values in the list are no larger than it. So the lower quartile is the 25th percentile and the median is the 50th percentile. One definition used to give percentiles, is that the p'th percentile is the 100/p*(n+1)'th observation. For example, with 7 observations, the 25^{th} percentile is the 100/25*8 = 2^{nd} observation in the sorted list.

Similarly, the 20^{th} percentile = 100/20*8 = 1.6^{th} observation.

For example for the data:

11 12 13 14 15 18 18.

the 25^{th} percentile is the value 12, while the 20^{th} percentile is 11.6

An approximate value for the pth percentile can be read from a *cumulative frequency graph* as the value of the variable corresponding to a cumulative frequency of
r%. So the lower *quartile* is the 25th percentile and the *median* is the 50th percentile. The term 'percentile' was introduced by *Galton* in 1885.

Population

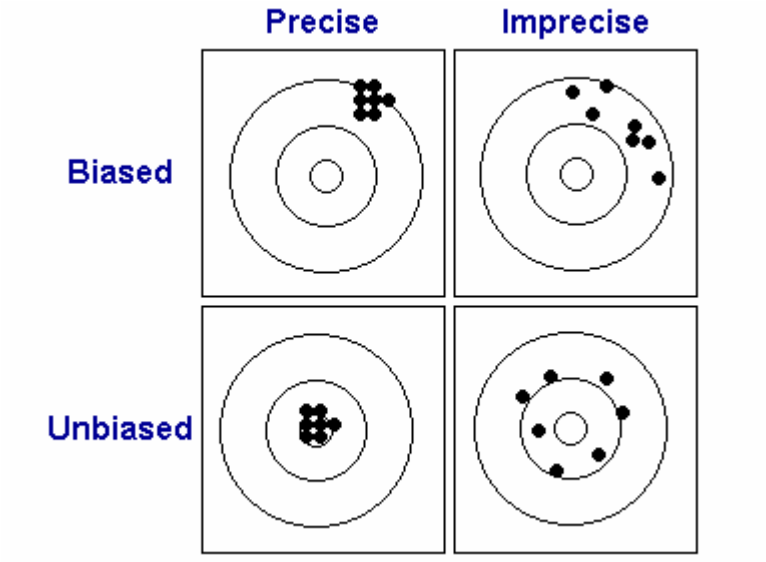
A population is a collection of units being studied. This might be the set of all people in a country. Units can be people, places, objects, years, drugs, or many other things. The term population is also used for the infinite population of all possible results of a sequence of statistical trials, for example, tossing a coin.
Much of statistics is concerned with estimating numerical properties (*parameters*) of an entire population from a *random sample* of units from the population.
Greek letters, e.g. μ, σ, θ are usually used for population parameters. This is to distinguish them from sample statistics, e.g. $\bar{x}$, s, p.

Precision

Precision is a measure of how close an *estimator* is expected to be to the true value of a *parameter*. Precision is usually expressed in terms of the standard error of the estimator. Less precision is reflected by a larger standard error.
For example, if the sample mean, $\bar{x}$, is used to estimate the population mean, μ, then the *standard error* s.e. (i.e. the measure of precision) is given by s.e. = s/√n, where s is the standard deviation of the data, and n is the sample size.

This formula shows that to improve precision, i.e. decrease the standard error, either the sample size, n, must be increased, or the spread of the data, s, must be decreased. One way to decrease the spread is to try and explain some of the variation in the data, see *pattern*.

The following diagram illustrates precision and *bias*, where the target value is the bullseye.



For example, the police decide to estimate the average speed of drivers using a 4-lane road. One suggested method is to use marked police cars to tail drivers and record their speed as the same as the police car. This is a precise method, because the speedometers are accurate, but is likely to give a biased result, because fast drivers may slow down when they see a police car behind them. An alternative method is to use a hidden speed gun. This should give an unbiased result, but the speed gun is less precise than a car's speedometer.

Proportion

For a variable with n observations, of which the frequency of a particular characteristic is r, the proportion is r/n. For example if the frequency of replanting was 11 times in 55 years, then the proportion was 11/55 = 0.2 of the years, or one fifth of the years. (See also *percentages*.)

Quantiles

Quantiles are a set of "cut points" that divide a numerical variable into groups containing (as far as possible) equal numbers of observations. Examples of quantiles include *quartiles*, *quintiles*, *deciles*, *percentiles*.

Quartiles

There are three quartiles. To find them, first sort the list into increasing order. The first or lower quartile of a list is a number (not necessarily in the list) such that 1/4 of the values in the sorted list are no larger than it, and at least 3/4 are no smaller than it.
With n numbers, one definition is that the lower quartile is the $(n+1)/4^{th}$ observation in the sorted list. For example, with 7 numbers the lower quartile is the 8/4 = 2nd value in the sorted list.
The second quartile is the median. The third or upper quartile is the "mirror image" of the lower quartile. So, with 7 numbers, it is the $3*8/4 = 6^{th}$ observation in the sorted list.
For example, with the sorted data:

11 12 13 14 15 18 18.

The lower quartile (2^{nd} value) is 12, the median (4^{th} value) is14, and the upper quartile (6^{th} value) is 18.

Quintile

Like a *quartile*, but dividing the data into five sets, rather than four. The lowest quintile is the 20% point, see *percentiles*.

For example, with the 7 values given above (see quartiles), the first quintile (20% point) is the (n+1)/5 = =8/5 = 1.6th value. This is).6 of the way between the two lowest values, and is therefore 11.6. The 40% point is the 3.2th value = 13.2.

Random sample

See *sample*

Range

The range is the difference between the *maximum* and the *minimum* values. It is a simple measure of the *spread* of the data.

For example with the data:

12 15 11 18 13 14 18

the range is (18 – 11) = 7.

Return period

The return period is the average time to one occurrence of an event. For example, if the probability of an event each year, say having to replant, is p = 0.2, or 20%, then the

return period = 1/p = 1/0.2 = 5 years.

So on average the event will "return" once in five years.

Risk

The risk of an event is the probability of that event occurring. If replanting is needed on 10 years out of 50, this is a probability, or risk, of 0.2 (or 20%).

Sample, random sample

A sample is a group of units, selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions (*inferences*) about the *population*.

A sample is usually used because the population is too large to study in its entirety. The sample should be representative of the population. This is best achieved by random sampling. The sample is then called a random sample.

Sampling distribution

A sampling distribution describes the probabilities associated with an *estimator*, when a *random sample* is drawn from a population. The random sample is considered as one of the many samples that might have been taken. Each would have given a different value for the estimator. The distribution of these different values is called the sampling distribution of the estimator.

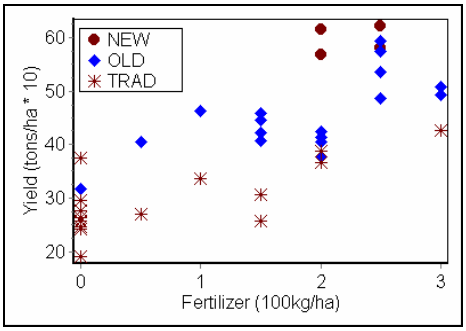
Deriving the sampling distribution is the first step in calculating a *confidence interval*, or in conducting a *hypothesis test*.

The standard deviation of the sampling distribution is a measure of the variability of the estimator, from sample to sample, and is called the *standard error* of the estimator. In many instances the sampling distribution of an estimator is approximately *normal*. This follows from the *central limit theorem*. Then approximate (95%) confidence intervals are found by simply taking the value of the estimate ±2* s.e.(estimate).

Scatter plot

A simple display when the data consists of pairs of values. The data are plotted as a series of points. If the data are ordered (for example, in time) then it may be sensible to join the successive points with a line. This is then called a *line graph*.

If there are other categorical variables, their values can be indicated using different plotting symbols or different colours.



Significance level, of a hypothesis test

The significance level of a statistical *hypothesis test* is a fixed probability of wrongly rejecting the null hypothesis H0, if it is in fact true.

It is the probability of a type I error and is set by the investigator in relation to the consequences of such an error. That is, we want to make the significance level as small as possible in order to protect the null hypothesis and to prevent the investigator from inadvertently making false claims.

Usually, the significance level is chosen to be 0.05 (or equivalently, 5%).

Skew, skewness

If the distribution (or "shape") of a variable is not *symmetrical* about the median or the mean it is said to be skew. The distribution has **positive skewness** if the tail of high values is longer than the tail of low values, and **negative skewness** if the reverse is true.

Spread, measures of

Most data sets exhibit variability -- all values are not the same! Two important aspects of the distribution of values are particularly important, they are the centre, and the spread.

The "centre" is a typical value around which the data are located. The mean and median are examples of typical values.

The spread describes the distance of the individual values from the centre. The *range* (maximum – minimum) and the *inter-quartile range* (upper quartile – lower quartile) are two summary measures of the spread of the data. The *standard deviation* is another summary measure of spread.

Standard deviation

The standard deviation (s.d.) is a commonly used summary measure of variation or spread of a set of data. It is a "typical" distance from the *mean*. Usually, about 70% of the observations are closer than 1 standard deviation from the mean and most (about 95%) are within 2 s.d. of the mean.

The standard deviation is a symmetrical measure of spread, and hence is less useful and more difficult to interpret for data sets that are *skew*. It is also sensitive to (i.e. its value can be greatly changed by) the presence of *outliers* in the data.

With the data values

12 15 11 18 13 14 18

the *variance*, s^2 was calculated as 7.62, so the standard deviation, $s = \sqrt{7.62} = 2.8$.

The mean, $\bar{x}$ was 14.4, so $(\bar{x} - s) = (14.4 - 2.8) = 11.6$

and $(\bar{x} + s) = 17.2$

So 4 of the 7 observations are within one standard deviation of the mean, while the other three are outside.

Standard error

The standard error (s.e.) is a measure of *precision*. It is a key component of statistical *inference*. The standard error of an *estimator* is a measure of how close it is likely to be, to the *parameter* it is estimating.

For example, if a survey is conducted to estimate the proportion of farmers who apply fertiliser in a particular season, the result from the sample may be given as an estimate of 32% with a standard error of 3%. The interpretation is that the true (population) proportion could easily be anything up to 3% less than, or more than the sample value, i.e. easily it could be as low as 29% or as high as 35%.

The standard error is a key building block in calculating *confidence intervals*. The 95% confidence interval is usually roughly the estimate ± 2*s.e.. So, in this example the 95% confidence interval is 32%-6% to 32%+6% or 26% to 38%. The interpretation is that the true population proportion is highly likely to be somewhere in the range 26% to 38%.

Symmetrical

A list of numbers is symmetrical if the data values are distributed in the same way, above and below the middle.

Symmetrical data sets:

- are easily interpreted;
- allow the presence of *outliers* to be detected similarly (i.e. using the same criteria), whether they are above the middle or below;
- allow the spread (variability) of similar data sets to be compared;

Some statistical techniques are appropriate only for data sets that are roughly symmetrical, (e.g. calculating and using the *standard deviation*). Hence *skew* data are sometimes *transformed*, so they become roughly symmetric. For example in economics, earnings or expenditure of a household may be transformed by taking the logarithm for this reason.

A *boxplot* is a useful graph to indicate the symmetry, or skewness of a set of data.

Table

When data are split into *categories*, tables provide a way of summary. A simple table gives the *frequency*, or the *percentage*, in each category. An example is below.

A one-way table of frequencies

Variety	Count
NEW	4
OLD	17
TRAD	15
All	36

A two-way table with percentages

	Variety			
Village	NEW	OLD	TRAD	All
SABEY	20	50	30	100
KESEN	0	43	57	100
NIKO	0	40	60	100
NANDA	14	50	36	100
All	11	47	42	100

The examples above show one-way tables, because they use one category column. There are as many cells in the table, as there are categories, plus the last cell, which is called the margin.

Tables can summarise data for 2 or more factors (category variables), and an example is shown below.

The contents of a table may, as above, be the frequencies (or percentages) at each combination of the factor levels. Alternatively they may be summary values of a numeric variable, for each category, as is shown below.

A two-way table showing median yields

	Variety			
Village	NEW	OLD	TRAD	All
SABEY	59.4	50.7	30.6	47.6
KESEN		40.7	26.4	27.6
NIKO		36.1	26.3	29.6
NANDA	59.7	45.8	37.6	42.5
All	2 values: 58.1, 61.4	44.6	27.6	40.5

Test statistic

A test statistic is a quantity calculated from the sample of data. It is used in *hypothesis testing*, where its value dictates whether the null hypothesis should be rejected or not.

The choice of a test statistic depends on the assumed *model* and the hypothesis being tested. For example, if comparing average annual rainfall in El Nino and Ordinary years then the *null hypothesis* could be that the two *population* means are equal, and the test statistic could be the difference in the *sample* means, i.e. $(\overline{x}_E - \overline{x}_O)$.

Time series

A series of measurements of a variable over time, usually at regular intervals.

Transforming variables

If there is evidence of marked *skewness* in a variable, then applying a transformation may make the resulting transformed variable more *symmetrical*. The example below shows what happens when square roots are taken:

Original variable	**Transformed variable**
1	1
4	2
4	2
1	1
36	6

Transforming skew data was very important 50 years ago, because the analysis was often simpler for variables that were symmetrical. This was partly because a *normal distribution* was then often an appropriate model, and much of the *statistical inference*/modeling depended on the data being from a normal distribution.

Recent advances in statistics have led to analyses being (almost as) simple for a wide range of statistical models, some of which are appropriate for modeling skew data. So now it is more important to consider the appropriate statistical model than to assume that data always need to be transformed if they lack symmetry.

Transforming data is not "cost free". For example, if the data above were the rainfall in mms, then analyzing the square root of the rainfall is not so easy to interpret as an analysis of the original data.

Beware of transforming when there are zeros in the data. A popular action used to be to add a small arbitrary value to the zeros and then to transform. Analysing the zeros separately is almost always to be preferred.

Tukey, John Wilder (1915-2000).

Tukey was an influential American statistician. Tukey was a chemistry graduate at Brown University, gaining his MA in 1937. He followed this with a PhD in topology at Princeton University in 1939. During the second world war he worked in the Fire Control Research Office alongside Wilks and Cochran. After the war he joined Wilks at Princeton University becoming a full professor at the age of 35. In 1946 he coined the word 'bit' as a shorthand for the 'binary digits' used by computers, and, in 1958, the word 'software' to describe the programs used by computers. In 1962 he introduced the trimmed mean as one of a number of robust summary statistics. In 1965, with John Cooley, he introduced the fast Fourier transform.

His 1970 book *Exploratory Data Analysis* introduced the *boxplot*.

Variable

The characteristic measured or observed when an observation is made. Variables may be non-numerical *(see categorical* or factor variable) or *numerical*.

The distinction between a categorical variable and a numerical variable is sometimes blurred. A categorical variable can always be coded numerically, for example, a gender (Male, Female) can be coded as 1 for Male or 2 for Female (or vice versa).

Similarly a numerical variable can be recoded into categories if needed. For example the variable "age" could be recoded into the three categories, of Young (<18yrs), Middle (18 to 60) and Old (>60).

An *average* (e.g. mean or median) and a measure of *spread*, (e.g. standard deviation or quartiles) are often used to summarize a numerical variable.

A *table* of the frequencies or percentages, at each level (or category) is often used to summarize a categorical variable.

Variability or variation or dispersion

The variability (or variation) in data is the extent to which successive values are different. Here are some examples where variability occurs (sometimes without us noticing it.)

- The time of arrival of a bus at a station.
- The score of Manchester United in a football match.
- The date of the year at which a crop can be planted.
- The length of the longest dry spell in the growing season.

The amount of variability in the data, and the different causes of the variability are often of importance in their own right. The variability (or "noise") in the data can also obscure the important information (or "signal").

Variance

The variance is a measure of variability, and is often denoted by s^2. In simple statistical methods the square root of the variance, s, which is called the *standard deviation*, is often used more. The standard deviation has the same units as the data themselves and is therefore easier to interpret. The variance becomes more useful in its own right when the contribution of different sources of variation are being assessed. This leads to a presentation called the "analysis of variance", often written as ANOVA.

The formula for the variance is $s^2 = \sum (x - \bar{x})^2 / (n-1)$

The calculation of the variance is shown in the Table below:

Data values	$(x - \bar{x})$	$(x - \bar{x})^2$	
12	-2.4	5.76	
15	0.6	0.36	
11	-3.4	11.56	
18	3.6	12.96	
13	-1.4	1.96	
14	-0.4	0.16	
18	3.6	12.96	
	Sum of squares = 45.72		$s^2 = 45.72/6 = 7.62$

Zero values

Variables may include zero values or other special values. For example the monthly rainfall total a dry site in March might be as follows:

Year	1	2	3	4	5	…	…	49	50
Monthly total (mm)	0	12.2	93.2	0	1.9			0	15.6

Zeros should be considered as an opportunity, rather than a problem. The data should usually be analysed (or modelled) in two parts. The first considers the chance of getting a zero value (as opposed to non-zero). Then the non-zero data are analysed further examine what happens in those years when there is some rain.

In the past, one strategy was to treat the zeros as representing something that had to be hidden, usually by adding a small value and then *transforming*. This is almost always counter-productive.

Acknowledgements

We acknowledge ideas and some entries from existing glossaries. In particular we have taken some of the biographical information, with permission, from "A Dictionary of Statistics" by Graham Upton and Ian Cook, published by the Oxford University Press. Some of the definitions were adapted from the STEPS project, http://www.stats.gla.ac.uk/steps/glossary/ by Valerie Easton and John McColl.