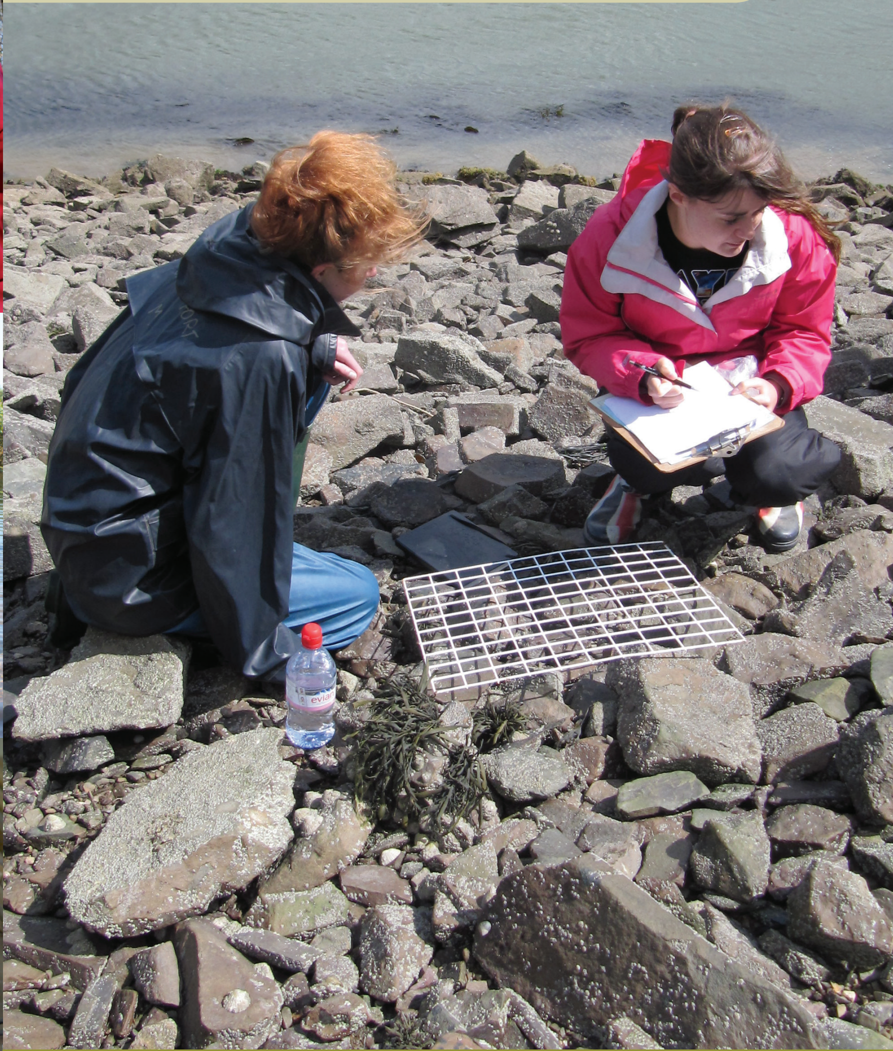
16-19 Biology

Maths skills for biologists



sampling strategies • mathematical and statistical skills • data presentation

BRITISH ECOLOGICAL SOCIETY

FSC

Planning field investigations

Asking ecological questions

Aim Statement of what you are trying to find out

Hypothesis Statement which can be scientifically tested to explain certain facts

Prediction Statement of what might happen in the future or in related situations

Designing a sampling strategy

A well designed sampling strategy should be:

- Unbiased No prejudice to a specific outcome
- Repeatable If using the same method, similar results are obtained
- Reproducible If repeated by another person, similar results are obtained
- Representative Samples are chosen to reflect relevant characteristics of the whole population
- Valid Experimental design is suitable to answer the question being asked

Sampling

Sample A data set selected from the population by a defined procedure. In other words, a small part of the population that is intended to show what the whole population is like

Sub-sample A sample of a sample

Sample area The whole area on the ground from which the sample is taken

Identifying units and measures of abundance (dependent variable)

Frequency How many times each species is present in samples within a given area. Can be expressed as a percentage

e.g. in a gridded quadrat of 100 squares, 28 squares are at least half-occupied by a particular plant species:

% frequency in this quadrat
= 28 ÷ 100 x 100 = 28%

Density The number of individuals in a given area (e.g. number of buttercup plants in a quadrat)

Cover An estimate of the area covered by a species (mostly used in plant investigations). Can be expressed as a percentage

Data definitions

Independent variable A variable which is changed or selected by the investigator

Dependent variable A variable which is measured for each change in the independent variable

Control variable A variable which has to be kept constant (or at least monitored)

Quantitative data Measurements (e.g. numbers, frequencies, rates, sizes)

Qualitative data Subjective assessments:

- Species lists The names of species present
- ACFOR scales An example might be plant percentage cover where: Abundant > 80%, Common = 50-80%, Frequent = 20-50%, Occasional = 5-20%, Rare < 5%

Matched (or paired) data A value from one data set corresponds with a value from another data set

Unmatched (or unpaired) data A value from one data set does not correspond with a particular value from another data set

Continuous data Numerical values given a magnitude by counting, ranking or measurement

- Interval data Data ordered, and the difference is equal and standardised (e.g. difference between 0°C and 10°C is the same as between 40°C and 50°C)
- Ordinal data Data ordered on an arbitrary scale (e.g. levels of aggression in apes)

Categorical (discontinuous) data: Values given labels (e.g. red, pink or blue flowers). No obvious ordering of categories (called nominal data)

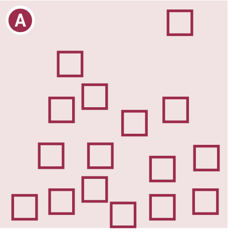
Planning field investigations

How to decide on sample points

Unbiased sub-samples can be taken from a sample area using a quadrat by:

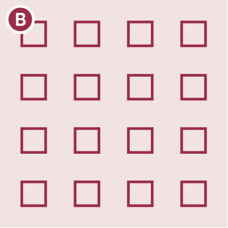
Random sampling Ⓐ

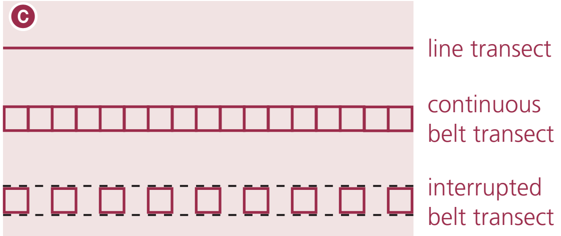
- Every point must have an equal chance of being chosen. Assumes conditions are the same across the sample area. Most often used when comparing two contrasting areas



Systematic (non-random) sampling

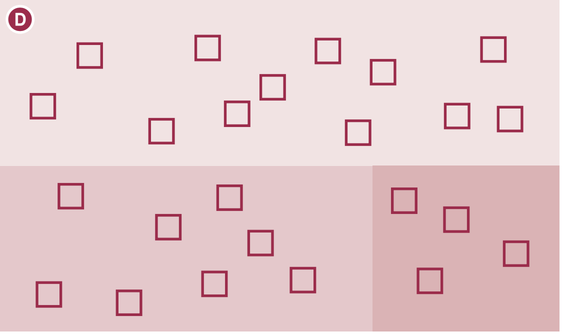
- Observations taken in a planned pattern Ⓑ
- Observations taken at regular intervals along a transect, especially where there is variation across the area. Often used when investigating relationships Ⓒ





Stratified sampling Ⓓ

- Observations taken from pre-selected parts of the larger sample area. The parts are actively selected to show a particular pattern



Sampling motile organisms

Many organisms are motile (they can move) so other sampling equipment may be required. Points should still be selected by random, systematic or stratified sampling

Pond nets Ⓔ for sampling organisms from still or running water, e.g. by stone washing and kick or sweep sampling

Sweep nets Ⓕ Fine-mesh nets for catching insects flying above or around vegetation

Beating trays Ⓖ Place a white sheet on the ground below vegetation. Beat and shake the branches so that invertebrates fall onto the sheet

Pooters Ⓗ Used to catch invertebrates directly from leaves

Mark-release-recapture

Technique for estimating population size of motile organisms

1. Take a sample from the population. Count and mark them (= M)
2. Release the sample back into the population
3. Allow time for marked individuals to mingle randomly within the population
4. Take a second sample in the same way
5. Count total number in the second sample (= S) and number recaptured, i.e. those marked in the first sample (= R)
6. Estimated population size (= P) is calculated using the Lincoln Index:

$$P = \frac{M \times S}{R}$$

M = Animals marked in first sample
S = Total animals in second sample
R = Recaptured in second sample

BRITISH ECOLOGICAL SOCIETY
FSC

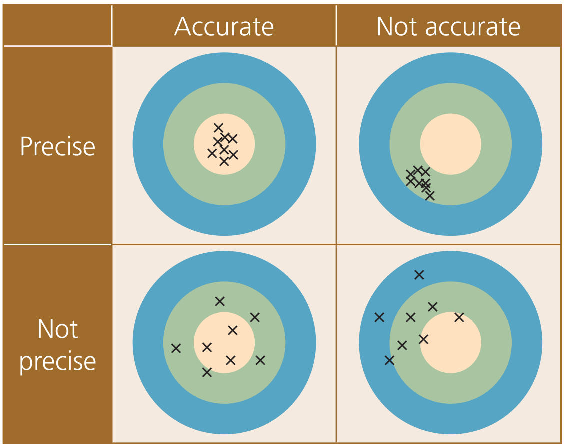
Mathematical skills

Accuracy, precision and errors

True value Value that would be obtained in an ideal measurement

Accuracy How close result is to the true value

Precision How much spread there is about the mean value



Resolution Smallest change in quantity that gives a perceptible change in the reading when using a measuring instrument

Uncertainty Interval within which the true value can be expected to lie, with a given level of confidence, e.g. 'temperature is 20 °C ± 2 °C, at a level of confidence of 95%'

Calibration marking a scale on a measuring instrument using reference values, e.g. placing a thermometer in melting ice to see if it reads 0°C, in order to check if it is calibrated correctly

Measurement error Difference between a measured value and the true value

Random error Causes readings to be spread about the true value due to unpredictable variation. Can be reduced by taking repeats

Systematic error Causes readings to differ from the true value by a consistent amount for each measurement. Cannot be dealt with by repeats

Anomaly A value in a set of results judged not to be part of the variation caused by random uncertainty

Averages

Mean Sum of all values divided by sample size (n). Used for normally distributed interval data:

- **True mean** (μ) Every individual in a population is measured
- **Sample mean** ($\bar{x}$) Only a sample of individuals in a population is measured

Median Middle value if data ordered from lowest to highest. Data do not have to be normally distributed. Used for interval or ordinal data

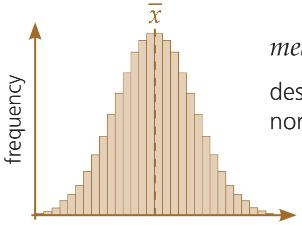
Mode Value in a data set which occurs most often. Can be used with continuous or categorical (nominal) data

Measures of dispersion

Dispersion Spread of data

Range Difference between maximum and minimum values of a particular data set

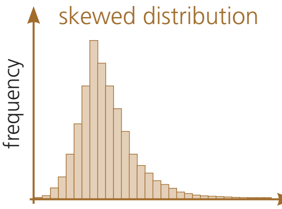
Normal distribution When measurements are plotted on a frequency histogram they form a symmetrical bell-shaped curve. Mean in the middle, with equal number of smaller and larger values on either side

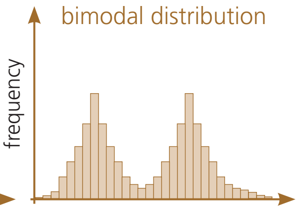


mean ± standard deviation

describes dispersion for a normal distribution

Non-normal distribution Measurements do not form a symmetrical bell-shaped curve





median ± interquartile range

describes dispersion for a non-normal distribution

BRITISH ECOLOGICAL SOCIETY

FSC

Mathematical skills

Standard deviation

A measure of dispersion of normally distributed data around a mean. There are two types:

- Population standard deviation (σ) Every individual in a population is measured
- Sample standard deviation (s) Only a sample is measured. Most commonly used with ecological data

The square of the standard deviation is called the variance (s^2)

$$s = \sqrt{\frac{\Sigma\,(x_i - \overline{x})^2}{n-1}}$$

Σ = sum of
x_i = individual values
$\overline{x}$ = sample mean
n = sample size

Worked example: standard deviation

Shell lengths (cm) of garden snails (*Helix aspersa*) were measured in three different habitats

Habitat	Shell lengths (cm)									
Woodland	0.1	0.9	0.4	0.5	0.5	0.5	0.5	0.6	0.5	0.5
Wall	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.6
Field edge	0.1	0.2	0.5	0.8	0.5	0.7	0.4	0.5	0.6	0.7

Sample means are the same (0.5 cm), but data are dispersed differently about the mean

Habitat	n	range	$\overline{x}$	s
Woodland	10	0.1 – 0.9	0.5	0.192
Wall	10	0.4 – 0.6	0.5	0.047
Field edge	10	0.1 – 0.8	0.5	0.221

Standard error

If more samples were taken, different sample means would be found. The standard error of the mean is the standard deviation of all these sample means around the true mean. It shows how close the sample mean is to the true mean

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

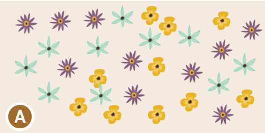
$s_{\overline{x}}$ = standard error
s = sample standard deviation
n = sample size

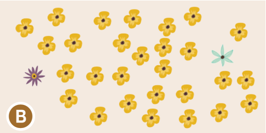
Measuring diversity

Species richness The number of different species

Species diversity Uniformity (or evenness) of the number of species and their relative abundance

Areas Ⓐ and Ⓑ are used in the worked examples of species diversity indices. Species richness is the same in both (3), but species diversity is different





Simpson's Diversity Index

$$D = \frac{N\,(N-1)}{\Sigma n\,(n-1)}$$

N = total organisms all species
n = organisms in each species
Σ = sum of

D ranges from 1 (only 1 species) to infinity (many species, all equally abundant). D has no units

Ⓐ	n	$n{-}1$	$n(n{-}1)$
✿	10	9	90
✳	10	9	90
✸	10	9	90

$\Sigma n\,(n-1) = 270$
$N\,(N-1) = 870$
$D = 3.22$

Ⓑ	n	$n{-}1$	$n(n{-}1)$
✿	28	27	756
✳	1	0	0
✸	1	0	0

$\Sigma n\,(n-1) = 756$
$N\,(N-1) = 870$
$D = 1.15$

Simpson-Yule Diversity Index

$$D = 1 - \sum\left(\frac{n}{N}\right)^2$$

N = total organisms all species
n = organisms in each species
Σ = sum of

D ranges from 0 (only 1 species) to 1 (many species, all equally abundant). D has no units

Ⓐ	n	n/N	$(n/N)^2$
✿	10	0.33	0.11
✳	10	0.33	0.11
✸	10	0.33	0.11

$\Sigma\,(n\,/\,N)^2 = 0.33$
$D = 0.67$

Ⓑ	n	n/N	$(n/N)^2$
✿	28	0.93	0.87
✳	1	0.03	0.00
✸	1	0.03	0.00

$\Sigma\,(n\,/\,N)^2 = 0.87$
$D = 0.13$

BRITISH
ECOLOGICAL
SOCIETY

FSC

Further information

The FSC is passionate about its cause and its long-standing history of helping people develop their knowledge of biology, ecology and taxonomy which spans many locations, involves numerous organisations and benefits people at various stages of their life.

We aim to encourage and develop passion for the natural world from a young age through the FSC Kids Fund, Young Darwin Scholarship and through family holidays.

We offer 350 wildlife, conservation and natural history courses each year, working in partnership with The Mammal Society, British Ecological Society and the British Trust for Ornithology to name a few which offer bursaries to help people attend our courses.

Each year over 135,000 publications are produced including fold-out charts to help people identify and learn about what they encounter outdoors as well as high quality, clearly written identification guides for non-specialists.

BRITISH ECOLOGICAL SOCIETY

The purpose of the British Ecological Society is to 'generate, communicate and promote ecological knowledge and solutions.' We are a thriving not for profit organisation with over 6,500 members across the world. Our activities include; scientific publishing, conferences, education, public engagement and grant giving to support the ecological science community in the UK and the developing world.

For more information about the BES visit: **www.britishecologicalsociety.org**

Want to find out more about wildlife?

Each year FSC runs a range of identifcation courses on insects, from a nationwide network of study Centres. Find out more at:
www.field-studies-council.org/naturalhistory

FSC also provides a wide range of wildlife guides to help you get to grips with identification. Find out more at:
www.field-studies-council.org/publications

This guide was developed by Mark Ward[1] and Simon Norman[1] with the assistance of Yoseph Araya[2], Dan Forman[2], Pen Holland[2], Louise Johnson[2], Sara Marsham[2], Rachel White[2] and Amy Padfield[2].

[1]Field Studies Council, [2]British Ecological Society.

© Text and concept FSC 2018. OP181. ISBN 978 1 908819 41 3.

MIX
From responsible sources
FSC® C022289
www.fsc.org

Statistical tests

Why do we use statistical tests?

Statistical tests are a tool for helping support sample data where sample sizes are small or any trends are unclear. They allow null hypotheses behind research questions to be accepted or rejected. Students must always return to the data to discuss possible scientific explanations

Statistical test Determines the level of confidence in sample data (= significance level), by calculating a statistical value which is used to accept or reject a null hypothesis. The statistical value is compared to a critical value

Null hypothesis (H_0) There is NO significant difference / correlation / association
i.e. there is no obvious pattern in the data

Alternative hypothesis (H_1) There IS a significant difference / correlation / association
i.e. there is a pattern in the data

Parametric Parametric statistical tests make the assumption that the distribution of data is normal. You can check this by drawing a size frequency histogram (see Mathematical skills)

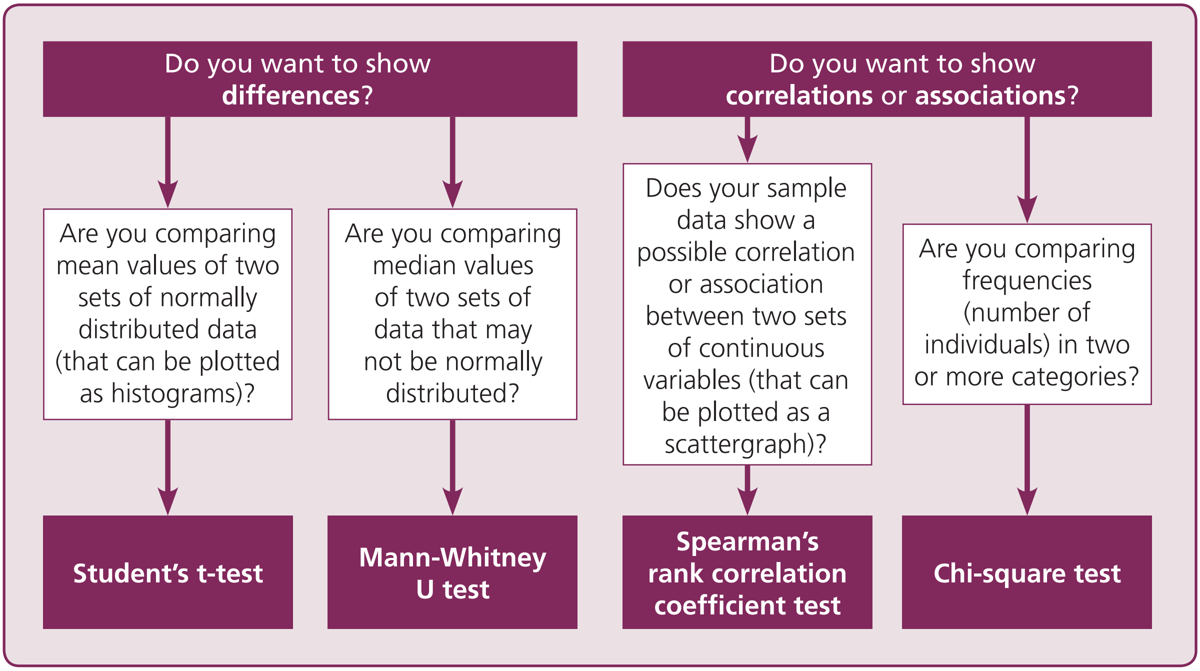
Non-parametric Non-parametric statistical tests make no assumptions about the distribution of the data

Critical values (from statistical tables)
These depend on:

- **Probability levels** (p) A value of 0.05 (5% significance or 95% confidence) means there is a 0.05 probability that any difference, correlation or association could still have occurred by chance. Null hypotheses are rejected (significant) at probabilities less than or equal to 0.05
- **Degrees of freedom** The bigger the sample the more freedom there will be for the data to be statistically significant. Degrees of freedom are calculated slightly differently for different tests

Association Any type of relationship between two variables where as one variable changes then there is a corresponding change in the second variable

Correlation An association where the relationship between the two variables is linear



Do you want to show **differences?**

Are you comparing mean values of two sets of normally distributed data (that can be plotted as histograms)?

Student's t-test

Are you comparing median values of two sets of data that may not be normally distributed?

Mann-Whitney U test

Do you want to show **correlations** or **associations?**

Does your sample data show a possible correlation or association between two sets of continuous variables (that can be plotted as a scattergraph)?

Spearman's rank correlation coefficient test

Are you comparing frequencies (number of individuals) in two or more categories?

Chi-square test

Student's t-test (unmatched)

This test is used to determine whether the sample means of two data sets are significantly different

- data must be be normally distributed (parametric)
- data must be unmatched (unpaired) and continuous, with interval level measurements
- designed for small sample sizes ($n < 30$)

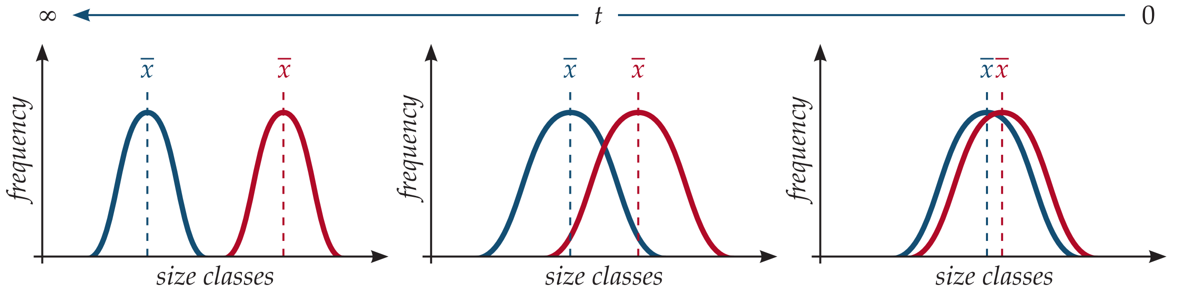
The null hypothesis is rejected at a stated probability value (e.g. p = 0.05) and for a given sample size (degrees of freedom) if calculated t is greater than or equal to the critical value

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}$ = mean value of x
s = standard deviation
n = sample size
(for data sets 1 and 2)

Critical values from a statistical table		
Degrees of freedom	p = 0.05	p = 0.01
18	2.10	2.88

$$\textit{degrees of freedom} = (n_1 + n_2) - 2$$



Worked example

An investigation into the difference in toothed topshell (*Osilinus lineatus*) length between an exposed and a sheltered shore.

Habitat	Shell lengths (mm)										n	$\bar{x}$	s
Exposed	12	23	21	17	24	31	29	25	26	24	10	23.2	5.53
Sheltered	30	27	26	31	28	28	35	29	21	25	10	28.0	3.74

Null hypothesis: 'There is no significant difference in the mean shell length of toothed topshells between the exposed and the sheltered shore.'

$$t = \frac{|23.2 - 28.0|}{\sqrt{\frac{(5.53)^2}{10} + \frac{(3.74)^2}{10}}} \qquad t = \frac{4.8}{\sqrt{\frac{30.62}{10} + \frac{14.0}{10}}} \qquad t = \frac{4.8}{\sqrt{4.46}} \qquad t = 2.27$$

The t value (2.27) is greater than critical value (2.10) at the 0.05 probability level with 18 degrees of freedom. The null hypothesis is rejected. Therefore there is a less than 0.05 probability that the difference in mean shell length of toothed topshells between the two shores is due to chance.

Related statistical tests

Student's t test (matched) Different version of t formula for when data are matched (paired).

Z test Similar to the t test (n-1 is replaced by n in the formula), but can be used with larger sample sizes ($n > 25$) and does not rely as much on the data being normally distributed.

ANOVA test Based on the t and z tests, but used to compare more than two means at one time. An F value is calculated that compares variance both within and between different samples.

Mann Whitney U test

This test is used to determine whether the medians of two data sets are significantly different

- data do not need to be normally distributed (non-parametric)
- data must be unmatched, and can be interval or ordinal
- both data sets need $n > 5$

The null hypothesis is rejected at a stated probability value (e.g. p = 0.05) and for the given sample sizes (n_1 and n_2) if calculated U is less than or equal to the critical value

$$U_1 = n_1 \times n_2 + 0.5\, n_2\, (n_2 + 1) - \Sigma R_2$$
$$U_2 = n_1 \times n_2 + 0.5\, n_1\, (n_1 + 1) - \Sigma R_1$$

n_1 = sample size of the first data set
n_2 = sample size of the second data set
ΣR_1 = sum of the ranks of the first data set
ΣR_2 = sum of the ranks of the first data set

Critical values from a statistical table (p = 0.05)

		n_1		
		8	9	10
n_2	8	13	15	17
	9		17	20
	10			23

Worked example

An investigation into % cover of dog's mercury (*Mercurialis perennis*) in two areas of woodland.

Quadrat	1	2	3	4	5	6	7	8
Site A	26	14	8	6	26	20	11	13
Site B	28	16	26	25	25	24	16	28

Null hypothesis: 'There is no significant difference in % cover of dog's mercury in the two areas.'

Re-organise both data sets in increasing order of size. Rank them from lowest to highest as if they were one set of data. When two or more values are the same an average rank is calculated.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Site A	6	8	11	13	14			20				26	26			
Rank (R_1)	1	2	3	4	5			8				13	13			
Site B						16	16		24	25	25			26	28	28
Rank (R_2)						6.5	6.5		9	10.5	10.5			13	15.5	15.5

Sum the ranks for each set of data

Site A: $\Sigma R_1 = 1 + 2 + 3 + 4 + 5 + 8 + 13 + 13 = 49$

Site B: $\Sigma R_2 = 6.5 + 6.5 + 9 + 10.5 + 10.5 + 13 + 15.5 + 15.5 = 87$

Calculate U_1 and U_2

$U_1 = 8 \times 8 + 0.5 \times 8\ (8{+}1) - 87 = 13$

$U_2 = 8 \times 8 + 0.5 \times 8\ (8{+}1) - 49 = 51$

Use the smaller U value as your statistic.

The U value (13) is equal to the critical value (13) at the 0.05 probability level with sample sizes of $n_1 = 8$ and $n_2 = 8$. The null hypothesis is rejected.

Therefore there is a less than 0.05 probability that the difference in % cover of dog's mercury in the two areas is due to chance.

Spearman's rank correlation coefficient test

This test is used to determine if there is a significant correlation between two variables

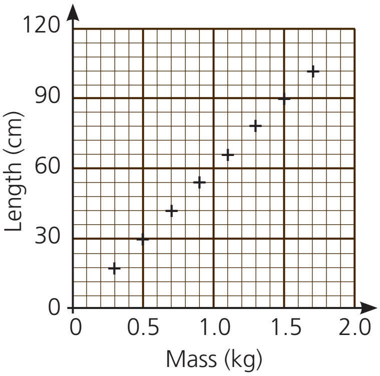
- data should be unmatched and continuous, and can be ordinal or interval
- number of pairs of data (n) should be greater than 8 and less than 30
- data can be plotted on a scattergraph

$$r_s = 1 - \left(\frac{6\Sigma D^2}{n(n^2-1)}\right)$$

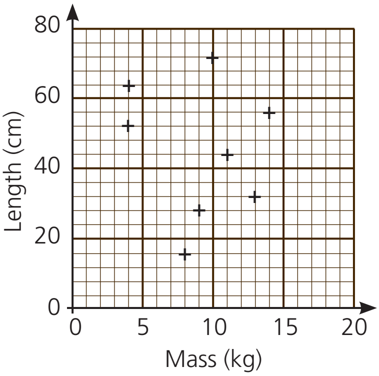
D = difference between ranks
n = number of pairs of data (sample size)

Critical values from a statistical table		
Number of pairs (n)	p = 0.05	p = 0.01
8	0.738	0.881

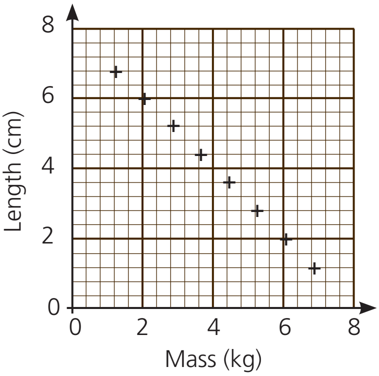
The null hypothesis is rejected at a stated probability value (e.g. p = 0.05) and for a given sample size (number of pairs) if calculated r_s is greater than or equal the critical value



Perfect positive correlation: r_s = +1.0



No correlation: r_s = 0



Perfect negative correlation: r_s = -1.0

Worked example

An investigation into the relationship between the length (cm) and mass (g) of grass snakes (*Natrix natrix*).

Null hypothesis: 'There is no significant correlation between the length and mass of grass snakes'.

Variable 1: length (cm)	Rank length (R_1)	Variable 2: mass (g)	Rank mass (R_2)	Difference (D)	D^2
10	1	150	1	0	0
15	2	230	2	0	0
23	3	360	3	0	0
46	5	820	5	0	0
29	4	510	4	0	0
70	6.5	1120	6	0.5	0.25
70	6.5	1210	7	-0.5	0.25
120	8	1820	8	0	0
				ΣD^2	0.50

Calculated $r_s = 0.994$

The r_s value (0.994) is greater than the critical value (0.881) at the 0.01 probability level where $n = 8$. The null hypothesis is rejected.

Therefore there is a less than 0.01 probability that the correlation between the length and mass of grass snakes is due to chance.

Related statistical tests

Pearson's Product-Moment of Correlation Test used to determine if there is a significant correlation between two variables. Data must be normally distributed. This is best used if you want to draw a line of best fit or if you want to predict values of the dependent variable from the independent variable

BRITISH ECOLOGICAL SOCIETY
FSC

Chi-square test

This is used to determine whether an observed distribution of a variable differs from a theoretical distribution or from a previous set of data

- data should be in the form of frequencies in a number of categories (i.e. nominal data)
- observations should be independent (i.e. one observation does not affect another)
- all expected frequencies should be larger in value than 5
- no more than 20 categories in total

The null hypothesis is rejected at a stated probability value (e.g. p = 0.05) and for a given size of data set (degrees of freedom) if calculated χ^2 is greater than or equal to the critical value

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = observed values
E = expected values
Σ = sum of

Observed values (O) Frequencies in each category that are actually counted or observed

Expected values (E) Frequencies predicted in each category if the null hypothesis were true (e.g. a null hypothesis might predict an even spread of results across a series of categories)

Critical values from a statistical table		
Degrees of freedom	p = 0.05	p = 0.01
4	9.49	13.28

degrees of freedom = (number of categories) – 1

Worked example

An investigation into the association between flat winkle (*Littorina obtusata*) frequencies and species of seaweed ($1\ m^2$ of each species).

Null hypothesis: ‘There is no significant difference between the observed and expected frequencies of flat winkles found on 5 species of seaweed’.

Species of seaweed	Observed frequencies (O)	Expected frequencies (E)	$(O - E)$	$(O - E)^2$	$(O - E)^2 \div E$
serrated wrack	45	20	25	625	31.3
bladder wrack	38	20	18	324	16.2
egg wrack	10	20	-10	100	5.0
spiral wrack	5	20	-15	225	11.3
oarweed	2	20	-18	324	16.2
TOTAL	100	100			79.9

Calculated $\chi^2 = 79.9$

The χ^2 value (79.9) is greater than critical value (13.28) at the 0.05 probability level with 4 degrees of freedom. The null hypothesis is rejected.

Therefore there is a less than 0.01 probability that the difference in frequencies of flat winkles between the 5 species of seaweed is due to chance.

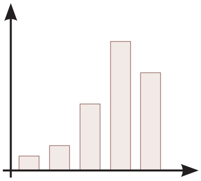
Related statistical tests

Chi-square test for association is another version used when you need to determine if there is a significant association between different data by comparing expected values (derived from a null hypothesis) with observed values.
In this case, *degrees of freedom = (number of rows – 1) × (number of columns – 1)*

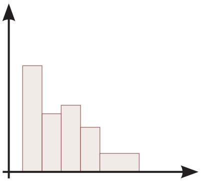
Data presentation

Results of an investigation can be displayed using a range of graphical techniques:

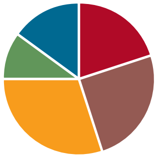
Bar chart Categorical (nominal) data. Categories on x-axis. Bar height represents frequency. Leave gaps between bars as data are discontinuous



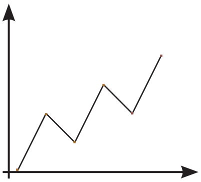
Histogram Interval data. Bar area shows frequency. Bars not necessarily equal width. No gaps between bars as data are continuous



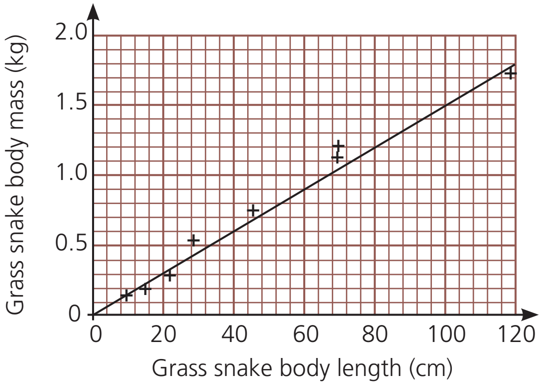
Pie chart Nominal or ordinal data. Area of circle segment represents proportion. Multiple pie charts can be used with the radius of the circle representing relative values



Line graph Ordinal or interval data. Both axes numerical. Independent variable plotted on x-axis. Dependent variable plotted on y-axis. Only join up points if data are continuous

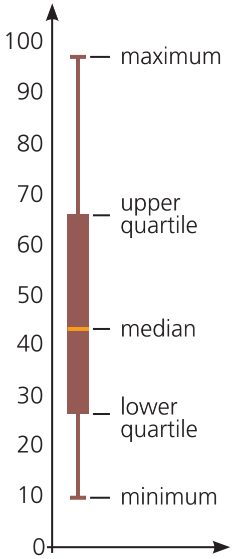


Scattergraph Interval data (continuous). Shows relationships between two variables. Independent variable plotted on x-axis, dependent variable on y-axis. Do not join up each point



Line of best fit Line drawn through a scattergraph to show relationship between two variables, drawn by eye or calculated using regression analysis

Box and whisker plots Display the median, upper quartile, lower quartile and range of a set of data



Quartile value half-way between either lowest or highest value and median. Hence there is an upper quartile (above median) and a lower quartile (below median)

Interquartile range Distance between upper and lower quartile values

Worked example

An investigation into girth (cm) of ash trees (*Fraxinus excelsior*) in a plantation (all trees planted at the same time)

10 20 25 26 27 32 33 34 43 52 53 56 60 70 75 80 96

LQ 26.5 | median 43 | UQ 65

interquartile range = 65 – 26.5 = 38.5

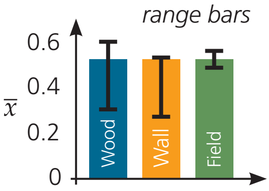
Range bars Added to a bar chart to display maximum and minimum values

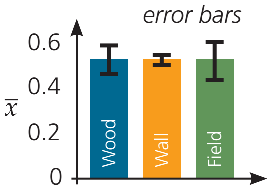
Error bars Added to a bar chart (or points on a line graph) *either* to display standard error *or* 1× standard deviation *or* 2× standard deviation

Worked example

Shell lengths (cm) of garden snails (*Helix aspersa*) were measured in 3 different habitats

Habitat	*max*	*min*	$\bar{x}$	*s.e.*
Wood	0.60	0.30	0.50	0.14
Wall	0.52	0.25	0.50	0.06
Field	0.55	0.45	0.50	0.18





BRITISH ECOLOGICAL SOCIETY

FSC