STANDARD LEVEL

EXTRA

PEARSON BACCALAUREATE

Environmental Systems and Societies

2nd Edition

ANDREW DAVIS • GARRETT NAGLE

Supporting every learner across the IB continuum

ALWAYS LEARNING

PEARSON

Appendix: Basic statistics and data analysis

Sampling

A sample is a representative body of data. A large number of items (the total population) can be represented by a small sub-section (the sample) when it is impractical or impossible to measure the total population. Sampling is therefore an efficient use of time and resources which makes it possible to make statements about the total population while using a representative section. Sampling makes fieldwork investigations manageable.

There are different types of sampling which have their own strengths and weaknesses. In general, there are two main types of sampling – spatial sampling and temporal sampling. Spatial sampling refers to samples that vary in where they are taken from. Temporal sampling refers to samples that are taken over different time periods. Both can be used – for example, monitoring water quality changes above and below a sewage outlet between summer and winter.

Both temporal and spatial sampling can be sub-divided into three main sub-types: random, systematic and stratified (Figure 1).

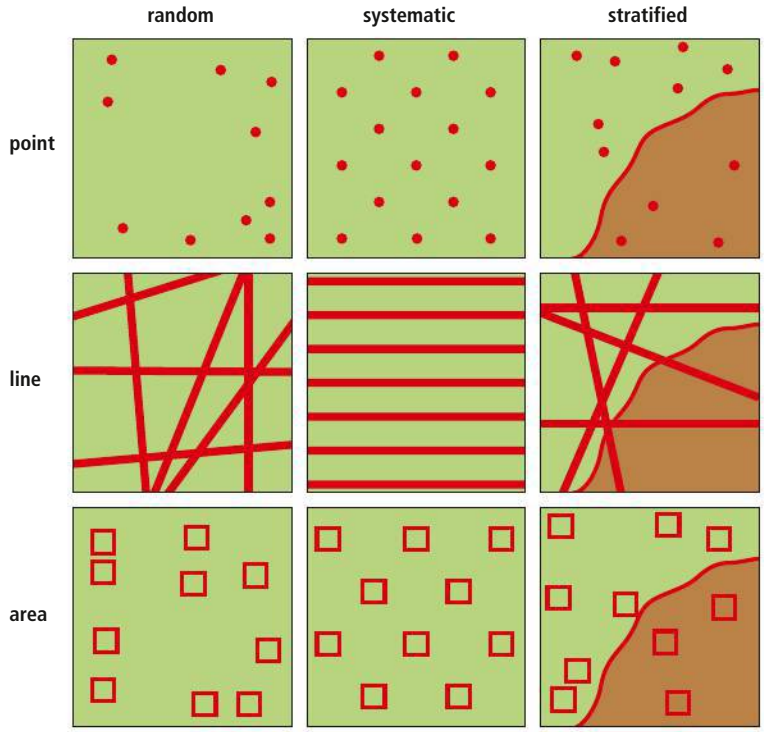


Figure 1 Random, systematic and stratified sampling

Before selecting one or more types of sampling, a number of questions should be considered.

- What is the population being studied and in what area/time?
- What is the minimum size of sampling needed to produce reliable information and results?
- What is the most appropriate form of sampling for the enquiry?

Random sampling

In a random sample, each item has an equal chance of being picked. Samples are often picked by using a random number table (Table 1). This is a table with no bias in the sequence of numbers. Once a number is chosen, it can be related to a map, a grid reference, an angular direction and distance. Although fair, the random sample may miss important parts of the survey area. It is also very time-consuming to do properly (page 135).

Table 1 Random number table

17	42	28	23	17	59	66	38	61	02	10	78
33	53	70	11	54	48	63	50	90	37	21	46
77	84	87	67	39	95	85	54	97	37	33	41
61	05	92	08	29	94	19	96	50	01	33	85
50	14	30	85	38	97	56	37	08	12	23	07
27	26	08	79	61	03	62	93	23	29	26	04
03	64	59	07	42	95	81	39	06	41	29	81
17	08	72	87	46	75	73	00	26	04	66	91
40	49	27	56	48	79	34	32	81	22	60	53

Systematic sampling

Systematic sampling is much quicker and easier than random sampling. Items are chosen at regular intervals (e.g. every 5 m, every tenth person, and so on). However, it is possible that a systematic sample will miss out important features. For example, in a survey of soil moisture and temperature in a ploughed field, if samples are taken on every ridge (or every furrow) and disregard other important microclimates, the results will be biased. The major problem with this type of sampling is that it can easily give a biased result because the sample is too small and, as a result, large areas are not included in the sample.

Stratified sampling

If it is known that there are important sub-groups in an area, for example different rock types which could influence soil types or farming types, it is possible to make a representative sample that takes into account all the sub-groups in the study area. It is also possible to weight the sample so that there is a proportionate number of samples related to the relative size of each sub-group.

Sample size

Determining the appropriate size of a sample is a critical matter. It depends on the nature and aims of the investigation but also on the time available (and other practical considerations such as access, land ownership, safety and so on). There are statistical formulae that can be used to determine sample size for a survey. Such statistical tests often depend on confidence limits (i.e. the statistical limits of probability that tell you how significant your results are likely to be). These are shown opposite in Table 2. For example, suppose in a survey of vegetation in an area with a sample of 100 points, 90% of the points were seen to be occupied by deciduous woodland, the true figure (at the

95% confidence level) is 90% ± 6% (i.e. between 84% and 96%). The larger the sample size, the narrower the limits of the true population.

Percentage calculated	Sample size 25	Sample size 50	Sample size 100
98% or 2%	5.6	4.0	2.8
97% or 3%	6.8	4.9	3.4
96% or 4%	7.8	5.6	3.9
95% or 5%	8.7	6.2	4.4
94% or 6%	9.5	6.8	4.8
92% or 8%	10.8	7.7	5.4
90% or 10%	12.0	8.5	6.0
85% or 15%	14.3	10.1	7.1
80% or 20%	16.0	11.4	8.0
75% or 25%	17.3	12.3	8.7
70% or 30%	18.3	13.0	9.2
65% or 35%	19.1	13.5	9.5
60% or 40%	19.6	13.9	9.8
55% or 45%	19.8	14.1	9.9
50%		20.0	14.2

Table 2 Range of error of estimates of population with one characteristic* at 95% confidence limit

* For example, the proportion of deciduous woodland in a survey of vegetation types in an area.

Confidence limits are based on normal probability (Appendix page 8). This assumes that 50% of the values are above the average (mean) and 50% are below. It also assumes that most of the values are within one standard deviation (Appendix page 8) of the mean. Probability states that in a normal distribution:

- 68% of samples lie within ± 1 standard deviation of the mean
- 95% of samples lie within ± 2 standard deviations of the mean
- 99.9% of samples lie within ± 3 standard deviations of the mean.

95% confidence limits are used in ecological investigations.

In other words, there is less than a 1 in 100 chance that the mean lies outside the sample mean ± 3 standard deviations, and less than a 1 in 20 chance that the true population mean lies outside the sample mean ± 2 standard deviations.

Descriptive statistics

There are many types of statistics, some of them extremely easy and some very complex. At the most basic, there are simple descriptive statistics. These include the mean or average, the maximum, minimum, range (maximum–minimum), the mode (most frequently occurring number, group or class) and the median (middle value when all the numbers are placed in ascending or descending rank order).

There are also three different types of data.

- Nominal data refer to objects which have names, such as rock types, land-uses, dates of floods, famines, and so on.
- Ordinal or ranked data are placed in ascending or descending order, for example settlement hierarchies are often expressed in terms of ranks. Spearman's rank correlation coefficient (Appendix pages 11–13) is used to compare two sets of ranked data such as infant mortality rate and purchasing power parity (Appendix pages 11–13).

- Interval or ratio data refer to real numbers – interval data have no true zero (as in the case of temperature which can be in °C or °F) whereas ratio data possess a true zero (as in the case of rainfall).

Summarizing data

The mean or average is found by totalling (Σ) the values (x) for all observations and then dividing by the total number of observations (n), thus Σx is divided by n. In Table 3, the average carbon dioxide emission per country is $\frac{21\,804.8}{20} = 1090.24$

Table 3 Carbon dioxide emissions for selected countries

Country	Million tonnes of carbon dioxide
USA	6044.0
China	5005.7
Russia	1523.6
India	1341.8
Japan	1256.8
Germany	808.0
Canada	638.8
UK	586.7
South Korea	465.2
Italy	449.5
Mexico	437.6
South Africa	436.6
Iran	433.2
Indonesia	377.9
France	373.4
Brazil	331.5
Spain	330.2
Ukraine	329.7
Australia	326.5
Saudi Arabia	308.1
Total ($\sum$)	21 804.8

The mode refers to the group or class which occurs most often. In Table 3, every value occurs once, so there is no mode. If, however, there were two values of 436 (for instance), the mode would be 436.

The median is the middle value when all the data are placed in ascending or descending order. In Table 3, there are two middle values (the 10th and 11th values), so we take the average of these two. In this case, the values are 449.5 and 437.6, so the median value is 443.55, which is not actually a value in the data set.

Summarizing groups of data

In some cases, the data we collect are in the form of groups (e.g. daily rainfall, slope angles or ages). Such data may be recorded as 0–4, 5–9, 10–14, 15–19, etc.

Table 4 shows daily rainfall in an area of rainforest. To make recording simpler, groups of 5 mm rainfall have been used. Finding an average or mean is slightly more difficult. We use the mid-point of the group, multiply it by the frequency and then proceed as before. So, from Table 4, $n = 100$ and $\Sigma x = 870$. The mean is $\frac{870}{100} = 8.7$.

Daily rainfall / mm	Mid-point	Frequency	Mid point × frequency
0–4	2	20	40
5–9	7	42	294
10–14	12	24	288
15–19	17	12	204
20–24	22	2	44
Total		100	870

Table 4 Daily rainfall for an area of tropical rainforest

The modal group is the one which occurs with the most frequency (i.e. 5–9 mm). The median or middle value is the average of the 50th and 51st values when ranked: these are both in the 5–9 mm group.

Measures of dispersion

The range is the difference between the maximum (largest) and the minimum (smallest) value. Going back to Table 3, the maximum is 6044.0 and the minimum is 308.1, hence the range is 6044.0 – 308.1 = 5735.9. An alternative measure is the inter-quartile range (IQR). This is similar to the range but gives only the range of the middle half of the results – by this the extremes are omitted. The IQR is found by removing the top and bottom quartiles (quarters) and stating the range that remains. The top quartile is found by taking the 25% highest values and then finding the mid-point between the last of the top 25% and the next point. The lower quartile is found by taking the 25% lowest values and finding the mid-point between the first of these and the next highest value. The first quartile is termed Q1, and the third quartile Q3.

Hence the IQR in the case of carbon dioxide emissions (Table 3) is from mid-way between the 5th and 6th values (i.e. half way between 1256.8 and 808 = 1032.4) to mid-way between the 15th and 16th values (i.e. half-way between 373.4 and 331.5 = 352.45). The result is 1032.4 – 352.45 = 679.95 – a much smaller variation than when all values (including extremes) are included.

Not every case is as easy! For example, there may be a number of observations not divisible by 4. In those situations, we have to make an informed guess at where the quartile would be.

If we add the figure for Poland (307.0 million tonnes) to Table 3, we get 21 observations. The quartiles are then at $5\frac{1}{4}$ and $15\frac{3}{4}$ (as each quarter is $5\frac{1}{4}$ in size).

The principle is the same as before. Find the values which represent 25% and 75% of the values. Then, find half the difference between the bottom of the top 25% and the next value below. Then find half the difference between the top of the lowest 25% and the next value above.

The 25% value is now found a quarter of the way between 1256.8 and 808.0, while the 75% value lies three-quarters of the way between 331.5 and 330.2. Thus, the first

quartile is found by subtracting one-quarter of the difference of 1256.8 and 808.0 from 1256.8.

$$1256.8 - \frac{(1256.8 - 808.0)}{4} = 1144.6$$

Q1 is mid-way between 1144.6 and 808.0: 976.3.

The 75% value is found by taking three-quarters of the difference of 331.5 and 330.2 from 331.5.

$$331.5 - 3\,\frac{(331.5 - 330.2)}{4} = 330.525$$

Q3 is located midway between 330.52 and 331.5: 331.0125

Thus, the IQR is 976.3 – 331.0125 = 645.2875.

Suppose we now add the figure for the 22nd largest producer of carbon dioxide, Thailand (267.8 million tonnes), to the table. There are now 22 observations.

The 25% and 75% values now are found at $5\frac{1}{2}$ and $16\frac{1}{2}$ (as each quarter is $5\frac{1}{2}$ in size, i.e. $\frac{22}{4}$). Thus the 25% value is found half-way between the 5th and 6th figures, 1256.8 and 808.0: 1032.4. The 75% value is found half-way between the 17th and 18th values, 330.2 and 329.7: 329.95. Hence Q1 is found half-way between the 25% value and the next value below, midway between 1032.4 and 808.0, namely 920.2. Q3 is found half-way between the 75% value and the next value above, the midpoint between 329.95 and 330.2, namely 330.075.

Thus the IQR in this case is 920.2 – 330.075 = 590.125.

Standard deviation

Another way of showing grouping around a central value is by using the standard deviation. This is one of the most important descriptive statistics because:

- it takes into account all the values in a distribution
- it is necessary for probability and for more complex statistics.

Standard deviation measures the dispersal of figures around the mean, and is calculated by first measuring the mean and then comparing the difference of each value from the mean.

Standard deviation is based on the ideas of probability. If a number of observations are made, then we would expect most to be quite close to the average, a few to be very much larger or smaller, and equal proportions above and below the mean.

The formula for the standard deviation (s) is:

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

where x refers to each observation, $\bar{x}$ to the mean, n is the number of points, and $(x - \bar{x})^2$ tells us to subtract the mean from each observation, and then to square the result.

Table 5 (opposite) shows the values worked out for $\bar{x}$, $(x - \bar{x})$ and $(x - \bar{x})^2$.

Standard deviation is found by putting the figures into the formula.

$$s = \sqrt{\frac{46\,724\,131.81}{20}} = \sqrt{2\,336\,206} = 1528 \text{ approx.}$$

Thus the average deviation of all values around the mean (1090.24) is 1528. This gives a much more accurate figure than the range or IQR, as it takes into account all values and is not as affected by extreme values. Given normal probability, we would expect that about 68% of the observations fall within 1 standard deviation of the mean, about 95% within 2 standard deviations of the mean, and about 99% within 3 standard deviations (Figure 2). Here we can see quite clearly that the rich countries are well above average (and some are over the mean plus two standard deviations, whereas the poorer countries are much more similar in income – they are all within one standard deviation of the mean).

Table 5 Working out the values to calculate standard deviation

Country	Millions of tonnes of carbon dioxide, x	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
USA	6044.0	1090.24	4953.76	24 539 738.14
China	5005.7	1090.24	3915.46	15 330 827.01
Russia	1523.6	1090.24	432.76	187 281.2176
India	1341.8	1090.24	251.56	63 282.4336
Japan	1256.8	1090.24	166.56	27 742.2336
Germany	808.0	1090.24	−282.24	79 659.4176
Canada	638.8	1090.24	−451.44	203 798.0736
UK	586.7	1090.24	−507.54	257 596.8516
South Korea	465.2	1090.24	−625.04	390 675.0016
Italy	449.5	1090.24	−640.74	410 547.7476
Mexico	437.6	1090.24	−652.64	425 938.9696
South Africa	436.6	1090.24	−653.64	427 245.2496
Iran	433.2	1090.24	−657.04	431 701.5616
Indonesia	377.9	1090.24	−712.34	507 428.2756
France	373.4	1090.24	−716.84	513 859.5856
Brazil	331.5	1090.24	−758.74	575 686.3876
Spain	330.2	1090.24	−760.04	577 660.8016
Ukraine	329.7	1090.24	−760.54	578 421.0916
Australia	326.5	1090.24	−763.74	583 298.7876
Saudi Arabia	308.1	1090.24	−782.14	611 742.9790
$\sum$				46 724 131.81

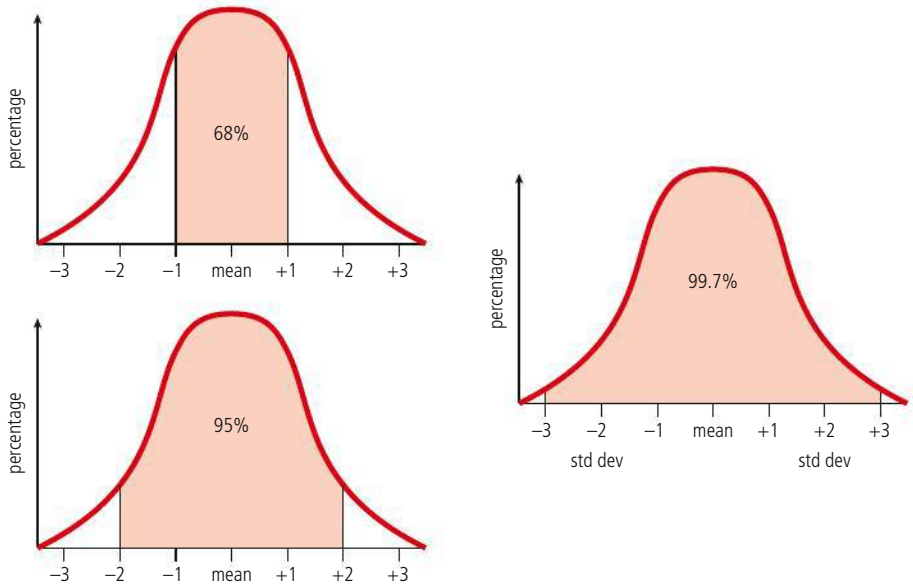


Figure 2 Standard deviations from the mean

Inferential statistics

Inferential statistics use results from surveys to make estimates or predictions (i.e. they make an inference about the total population or about some future situation). To understand inferential statistics, it is important to grasp three related concepts: probability, sampling and significance.

Probability

One of the main tasks of inferential statistics is to establish the likelihood of a particular event or value occurring – this is known as probability. Probability is measured on a scale from 0 to 1. The value 1 represents absolute certainty (e.g. everyone will eventually die), whereas the value 0 represents absolute impossibility (a non-American citizen will become President of the USA). In statistics, probability (p) is often expressed as a percentage:

- $p = 0.05$ (a 1-in-20 chance) is a 95% level of probability
- $p = 0.01$ (a 1-in-100 chance) is a 99% level of probability
- $p = 0.001$ (a 1-in-1000 chance) is a 99.9% level of probability.

Sampling

See Appendix pages 1–2 for a discussion of sampling methods. The key aspect here is to decide how reliable our sample size is and how accurately it allows us to predict (i.e. what is the probability that our sample is truly representative?).

Ecology uses the 95% significance level (a 1-in-20 chance of a result occuring by chance); this correlates with the 95% (p = 0.05) probability level occurring by chance.

Significance

Significance relates to the probability that a hypothesis is true. In statistics, it is the convention to use a null hypothesis (a negative statement that we aim to disprove). A null hypothesis might state, for example, that there is no difference in the water quality above and below a sewage outlet. The alternative hypothesis (aka the research hypothesis) would state that there is a difference between the water quality above

and below a sewage outlet. The probability at which it is decided to reject the null hypothesis is known as the significance level. The significance level indicates the number of times that the observed differences could be caused by chance. The practice is to refer to results as 'significant', 'highly significant' and 'very highly significant', respectively at the 95%, 99% and 99.9% levels of significance (Figure 3). This means there is a 1-in-20, 1-in-100 and 1-in-1000 chance (probability) of the result occurring by chance.

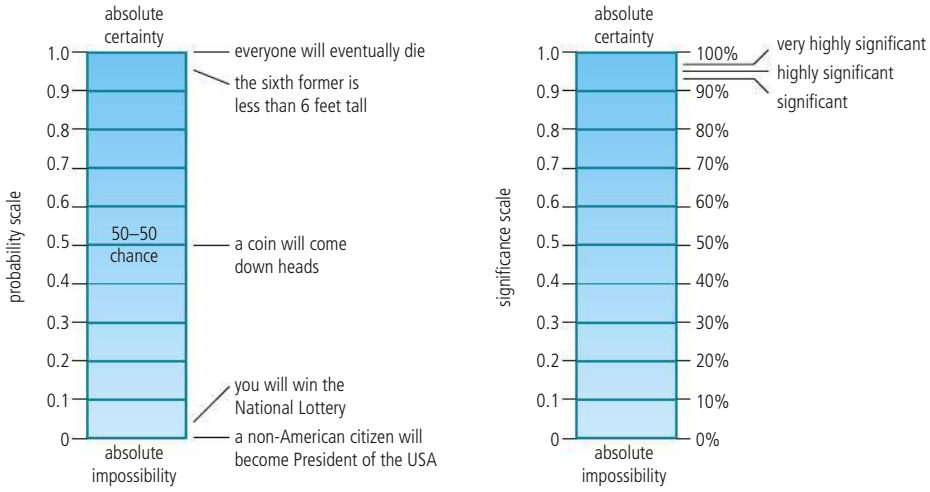
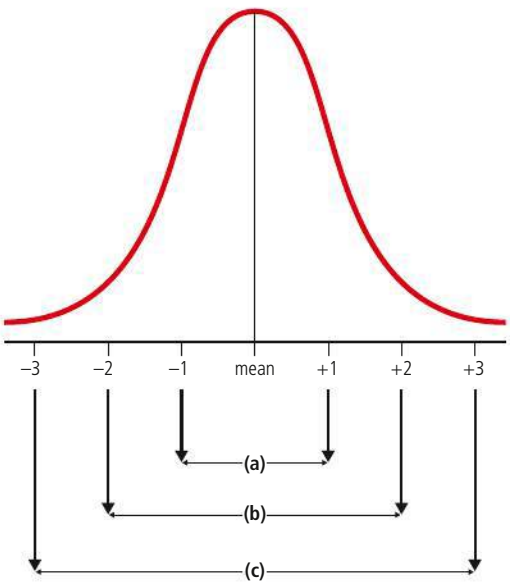


Figure 3 Probability and significance scales

Sampling error or standard error

This statistic provides an estimate of the true population mean (i.e. the likely value we would get if we were able to measure all individuals in a population). It is based on two concepts – probability and normal distribution. In general, we would expect in a large population very few very large values and very few very small values. Most values would tend to group around the mean (Figure 4). So, any estimate that we make is likely to be somewhere near the true population mean. Our estimates are less likely to



(a) 68% of values are within ± 1 standard deviation of the mean

(b) 95% of values are within ± 2 standard deviation of the mean

(c) 99% of values are within ± 3 standard deviation of the mean

Figure 4 Normal distribution curve

be very much smaller or larger than the population mean. Thus, it is possible, within certain limits, to estimate where the true population mean lies.

The following example illustrates the point. In a survey of vegetation characteristics on the Isle of Purbeck (UK), a sample of 100 observations found that 50% of the area was farmland, 14% heathland, 12% woodland and 24% other. From these figures, it is possible to state that the true population mean for woodland is somewhere around 12%. The formula for sampling error or standard error is:

$$\sqrt{\frac{P(100-p)}{n}}$$

Where P refers to the proportion of (in this case) woodland

$(100-p)$ refers to the proportion that is not (in this case) woodland

n refers to the sample size.

Thus, our estimate of the proportion of woodland that exists on the Isle of Purbeck is:

$$12\% \pm \sqrt{\frac{12 \times 88}{100}} = 12\% \pm 3.2\% = \text{from } 8.8\% \text{ to } 15.2\%$$

We are stating that we know that our own survey may not be totally accurate and that the true population mean is likely to lie somewhere between these limits.

The larger the sample, the more accurate the estimate. In the above example, if the proportion of woodland were still 12% but the sample size was 1000, the standard error or sampling error would be:

$$12\% \pm \sqrt{\frac{12 \times 88}{1000}} = 12\% \pm 1.0\%$$

Equally, given our results from our sample of 100 we can say that:

- one standard error = 12% ± 3.2% = a range from 8.8% to 15.2%
- two standard errors = 12% ± 6.4% = a range from 5.6% to 18.4%
- three standard errors = 12% ± 9.6% = a range from 2.4% to 21.6%

Confidence limits

Confidence limits are based on the ideas of probability and assume the data being sampled have a normal distribution. They are usually established at the 95% and 99% levels . These levels are found by multiplying the standard error by 1.96 and 2.56 (i.e. 2 and 3 standard deviations above and below the mean – see Figure 4).

Going back to the survey on the Isle of Purbeck, the sample mean was 12% and the standard error was 3.2%. So, at the 95% confidence level, the actual confidence level for the woodland would be:

$$12\% \pm (3.2 \times 1.96) = 12\% \pm 6.27 = 5.63\text{–}18.27\%$$

At the 99% confidence level, the limits would be:

$$12\% \pm (3.2 \times 2.56) = 12\% \pm 8.19 = 3.81\text{–}20.19\%.$$

We could express this in a slightly different way and say that if the actual woodland mean were 12% , we would expect that:

- 95% of the surveys would record the mean as lying between 5.63% and 18.27%
- 99% of surveys would record the mean as lying between 3.81% and 20.19%.

Spearman's rank correlation coefficient (Rs)

Spearman's rank correlation coefficient (Rs) is one of the most widely used statistics in social and environmental sciences. It is relatively quick and easy to do and only requires that data are available on the ordinal (ranked) scale. More complex data can be transformed into ranks very simply. It is called a rank correlation because only the ranks are correlated, not the actual values. The use of Rs allows us to decide whether or not there is a significant statistical correlation (relationship) between two sets of data. In some cases, it is clear whether a correlation exists or not. However, in most cases it is not so clear cut and to avoid subjective comments, we use Rs to bring in a certain amount of objectivity.

Purchasing power parity (PPP) and infant mortality rates (IMR)

Procedure

1. State null hypothesis (H_0) – there is no relationship between IMR and PPP. The alternative hypothesis (H_1) is that there is a relationship between IMR and PPP. (Note that this example uses secondary data.)
2. Rank both sets of data from high to low (highest value is rank 1, second highest 2, and so on) as in Table 6. In the case of joint or tied ranks, find the average rank (if two values occupy positions 2 and 3 they both take on rank 2.5. If three values occupy positions 4, 5 and 6, they all take rank 5).
3. Work out the correlation using the formula:

$$Rs = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

where d refers to the difference between ranks and n the number of observations.

Country	PPP / $	IMR / ‰	Rank / PPP	Rank / IMR	Difference	Difference^2
Afghanistan	800	151.9	10	1	9	81
Bangladesh	1 500	59.2	9	2	7	49
Brazil	10 100	22.5	4	6	–2	4
China	6 000	20.2	6	7	–1	1
India	2 800	30.1	7	5	2	4
Kenya	1 600	54.7	8	3	5	25
Mexico	14 200	18.4	3	8	–5	25
South Africa	10 000	44.4	5	4	1	1
UK	36 600	4.6	2	9	–7	49
USA	47 000	6.3	1	10	–9	81
						320

Table 6 Ranked data for PPP and IMR

$$Rs = 1 - \frac{6\Sigma d^2}{n^3 - n} = 1 - \frac{6 \times 320}{10^3 - 10} = 1 - \frac{1920}{990} = 1 - 1.94 = -0.94$$

4. Compare the computed Rs with the critical values for a given level of significance (normally 95% in ecological studies) in the statistical tables (Table 7). If the

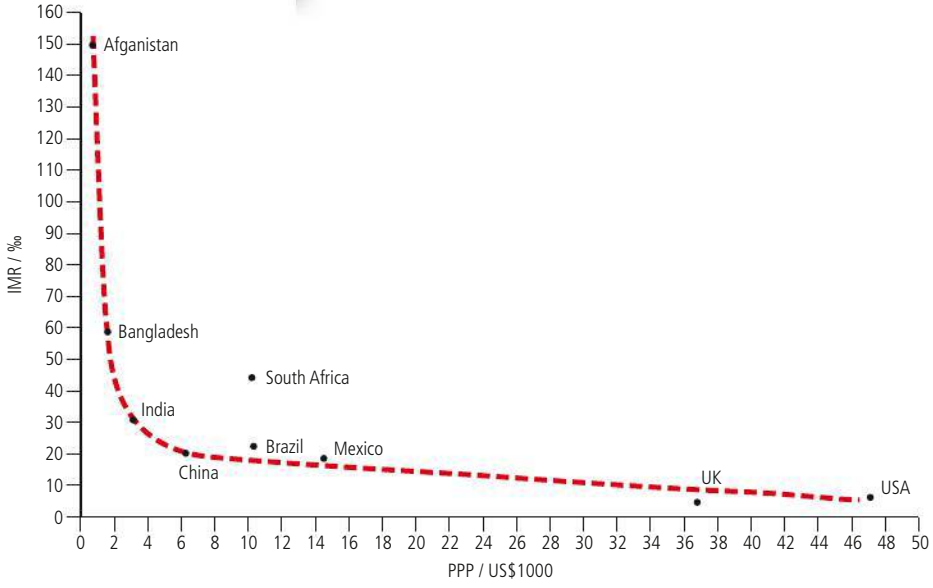
computed value exceeds the critical values in the table, we can say that we are 95% or 99% sure that there is a relationship between the sets of data. In other words, there is only a 5% or 1% chance that there is no relationship between the data.

Table 7 Levels of significance for Spearman's rank correlation

n	Significance level	
	95%	99%
4	1.00	–
5	0.90	1.00
6	0.83	0.94
7	0.71	0.89
8	0.64	0.83
9	0.60	0.78
10	0.56	0.75
12	0.51	0.71
14	0.46	0.65
16	0.43	0.60
18	0.40	0.56
20	0.38	0.53
22	0.36	0.51
24	0.34	0.49
26	0.33	0.47
28	0.32	0.45
30	0.31	0.42

It is convention to accept 95% and 99% levels of significance. From the table, for a sample of 10 (as in our example), these values are 0.56 for 95% significance and 0.75 for 99% significance. In this example, our computed value is –0.94 (the minus sign can be ignored), so there is more than 99% chance that there is a relationship between the data.

Figure 5 Scatter graph to show the relationship between PPP and IMR



The fact that the correlation is negative shows that it is an inverse relationship (as one variable increases the other decreases). Thus as PPP increases, infant mortality rate decreases (Figure 5). The next stage would be to offer explanations for the relationship.

It is important to realize that Spearman's rank has its weaknesses. It has a number of limitations which must be considered.

- It requires a sample size of at least seven.
- It tests for linear relationships (Figure 6a and b) and would give an answer of zero for data such as river discharge and frequency, which follows a curvilinear pattern, with few very low or very high flows and a large number of medium flows (Figure 6c).
- It is easy to make false correlations, as between summer temperatures in the UK and infant mortality rates in India. A significant relation is not necessarily a causal one.
- The question of scale is always important. As shown in Figure 6d, a survey of distance downstream and tubifex worms for the whole of a drainage system may give a strong correlation, whereas analysis of just a small section gives a much lower result.

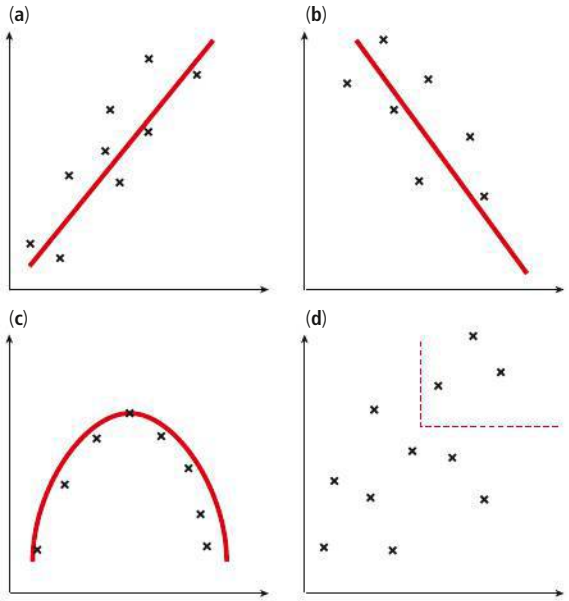


Figure 6 Spearman's rank graphs.
(a) Linear relationship, Rs = +1.0.
(b) Linear relationship, Rs = –1.0.
(c) Curvilinear relationship, Rs = 0.0.
(d) Mixed relationship, Rs = +1.0 for complete data set; Rs = 0.0 for subset.

As always, statistics are tools to be used. They are only part of the analysis, and we must be aware of their limits.

There are other correlation coefficients – the Pearson product moment correlation coefficient is a more powerful correlation but it requires more sophisticated data. However, it is available on many computer packages. The principles are the same as for Spearman's rank, but the data need to be interval or ratio (real numbers) rather than just ranked data. Again, the correlation tests for a linear relationship.

The Mann Whitney U Test

This is one of the most powerful distribution (non-parametric) free tests. Even when only medium sized samples (i.e. 10–20) are involved it has about 95% of the power of Student's *t*-test. It can be used with ordinal (ranked) data, as long as both sets are ranked in a single sequence, or with data on an interval scale that have been allotted ranks in a single sequence. It is used to test whether the mean of two independent samples is statistically different (i.e. that the samples come from different populations). The samples do not have to be the same size – when the samples are of different sizes the smaller of the two is termed n_1. It works best when one of the samples has at least 9 readings – see significance tables.

Procedure

Water temperature upstream and downstream of a sewage outlet (winter):

Upstream 6, 6, 8, 7, 5, 4, 5

Downstream 9, 8, 10, 9, 8, 9, 10, 8, 9

1. The null hypothesis, H_0, states that there is no difference in the means of the two samples. It assumes that the differences between them are the result of chance and are not significant.
2. The alternative hypothesis, H_1, is that there is a significant difference between the two samples, in this case that water temperature below the sewage outlet is significantly higher than that above the outlet.
3. The critical level is 95%.
4. To apply the statistic, the values must be placed in rank order, but kept in their groups. (Conventionally, the smallest value is given rank 1. Where values tie, assign an average rank to each value.)

Upstream 4.5, 4.5, 8.5, 6, 2.5, 1, 2.5 ($\sum = 28.5$)

Downstream 12.5, 8.5, 15.5, 12.5, 8.5, 12.5, 15.5, 8.5, 12.5 ($\sum = 106.5$)

The Mann Whitney formula is:

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

Or

$$U = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$

Where R_1 = the sum of the ranks given to values in n_1, and R_2 = the sum of the ranks given to the values in n_2.

Thus,

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$
$$= 7 \times 9 + \frac{7 (7 + 1)}{2} - 28.5$$
$$= 62.5$$

And

$$U = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$
$$= 7 \times 9 + \frac{9 (9 + 1)}{2} - 106.5$$
$$= 1.5$$

5. Referring to the statistical tables, the lower U value is used, in this case 1.5. In order for it to be significant, it must be lower than the critical values in the table. In the significance tables (Tables 8 and 9) the value for n_1 and n_2 is 16 at the 0.05 level, and 10 at the 0.01 level. Hence, we are 99% certain that given the data above, there is a significant difference in the temperature above and below the sewage outlet.

	n_1=2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n_2=3	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
4	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
5	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
6	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
7	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
8	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
9	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
10	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
11	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
12	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
13	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
14	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
15	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
16	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
17	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
18	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
19	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
20	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139

Table 8 95% level of significance

	N_1=2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
N_2=3	0	0	0	0	0	1	1	2	2	2	3	3	3	4	4	5	5	5	6
4	0	0	0	1	2	2	3	4	4	5	6	6	7	8	9	9	10	10	11
5	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
6	0	0	2	3	4	5	7	8	9	10	12	13	14	16	17	19	20	21	23
7	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
8	0	1	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31	33	35
9	0	2	4	6	8	10	12	15	17	19	22	24	27	29	32	34	37	39	41
10	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	39	42	45	48
11	0	2	5	8	10	13	16	19	23	26	29	32	35	38	42	45	48	51	54
12	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57	61
13	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
14	1	3	7	11	14	18	23	27	31	35	39	44	48	52	57	61	66	70	74
15	1	4	8	12	16	20	25	29	34	38	43	48	52	57	62	67	71	76	81
16	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	87
17	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
18	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
19	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
20	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115

Table 9 99% level of significance

The nearest neighbour index (NNI)

Part of the study of ecosystems (and vegetation) is concerned with distributions in space and over time. The spatial distribution of vegetation in an area can be described by looking at a map. This may lead us to conclude that the some types of vegetation (or ecosystems) are scattered, dispersed or concentrated. However, the main weakness with the visual method is that it is subjective and individuals differ in their interpretation of the pattern. Some objective measure is required and this is provided by the NNI.

There are three main types of pattern which can be distinguished: uniform or regular, clustered or aggregated, and random. These are shown in Figure 7. The points may represent individual trees, etc. If the pattern is regular, the distance between any one point and its nearest neighbour should be approximately the same as from any other point. If the pattern is clustered, then many points will be found a short distance from each other and there will be large areas of the map without any points. A random distribution normally has a mixture of some clustering and some regularity.

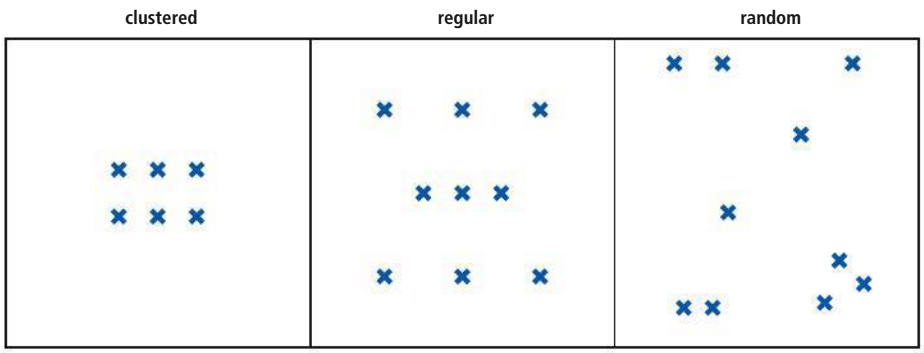


Figure 7 Nearest neighbour patterns

NNI is the technique most commonly used to analyse these patterns. It is a measure of the spatial distribution of points, and is derived from the average distance between each point and its nearest neighbour. This figure is then compared to computed values which state whether the pattern is regular (NNI = 2.15), clustered (NNI = 0) or random (NNI = 1.0). Thus, a value below 1.0 shows a tendency towards clustering, whereas a value of above 1.0 shows a tendency towards uniformity.

The formula for the NNI looks somewhat daunting at first, but, like most statistics, is extremely straightforward providing care is taken.

$$\text{NNI or Rn} = 2\bar{D}\sqrt{\frac{N}{A}}$$

where $\overline{D}$ is the average distance between each point and its nearest neighbour and is calculated by finding $\Sigma\frac{d}{N}$ (d refers to each individual distance), N the number of points under study and A the size of the area under study. It is important that you use the same units for distance and area (e.g. m or km but not a mixture).

For example, a survey of the distribution of vegetation types in the Camley Street Natural Park in London was undertaken to plot the distribution of deciduous trees and marsh species. The results are shown in Figure 8 and Tables 10 and 11. The area of the nature reserve is approximately 13 200 m^2.

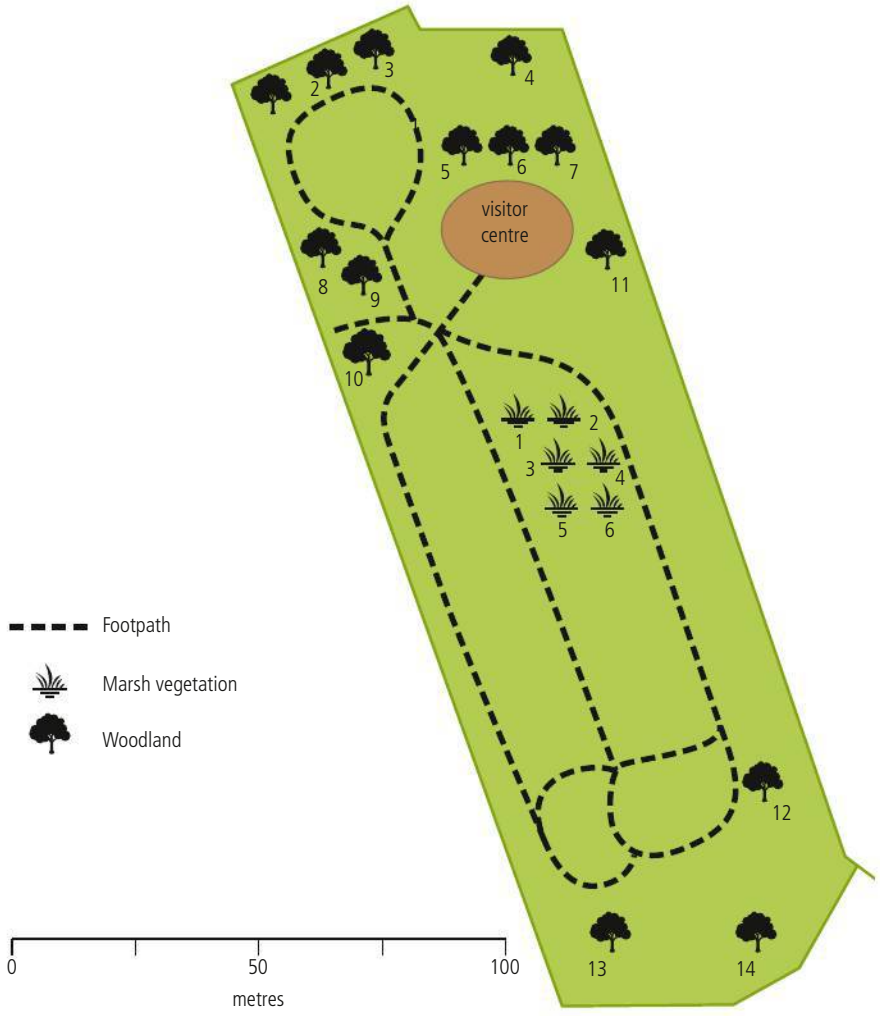


Figure 8 Distribution of vegetation in Camley Street Natural Park

Vegetation	Nearest neighbour	Distance / m
1	2	10
2	1, 3	10
3	2, 4, 5	10
4	3, 6	10
5	3, 6	10
6	4, 5	10
$\sum d$		60

Table 10 Nearest neighbour distances for marshland vegetation

$$\text{NNI or Rn} = 2\bar{D}\sqrt{\frac{N}{A}}$$

$$\bar{D} = \sum\frac{d}{N} = \frac{60}{6} = 10$$

$$\text{Rn} = 2 \times 10 \times \left(\sqrt{\frac{6}{3\,200}}\right) = 0.43$$

This answer suggests a significant degree of clustering (Figure 9).

Table 11 Nearest neighbour distances for woodland vegetation

Vegetation type	Nearest neighbour	Distance / m
1	2	14
2	3	10
3	2	10
4	6	20
5	6	10
6	5, 7	10
7	6	10
8	9	10
9	8	10
10	9	20
11	7	22
12	14	30
13	14	30
14	13	30
$\sum d$		236

$$\text{NNI or Rn} = 2\bar{D}\sqrt{\frac{N}{A}}$$

$$\bar{D} = \sum\frac{d}{N} = \frac{236}{14} = 16.86$$

$$\text{Rn} = 2 \times 16.86 \times \left(\sqrt{\frac{14}{3\,200}}\right) = 1.10$$

This answer suggests regular spacing (Figure 9).

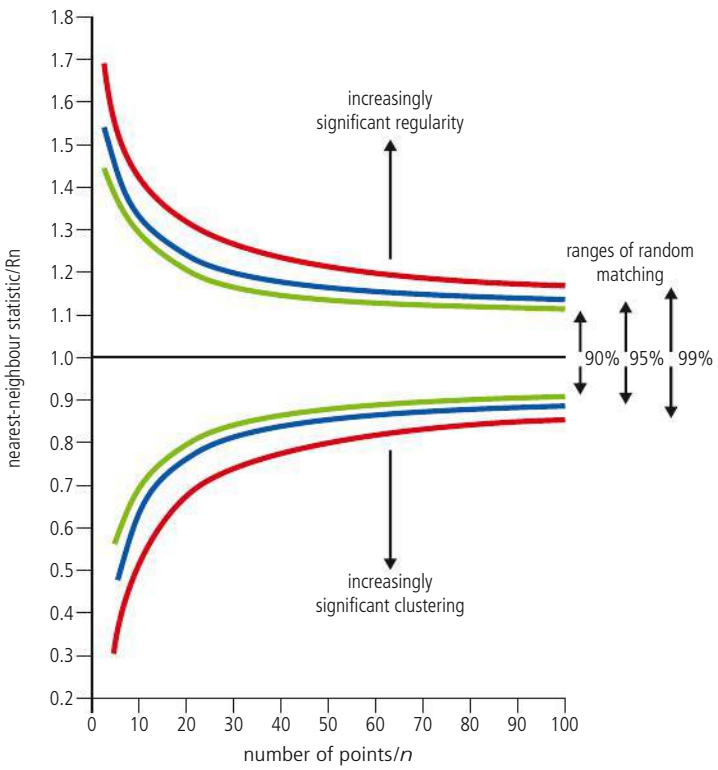


Figure 9 NNI significance ranges

There are important points to bear in mind when using NNI.

- Two or more sub-patterns (one clustered, one regular) may suggest a random result.
- What is the definition of, for example, a tree? Do you include all individuals – or just those above a certain size?
- Why do we take the nearest neighbour? Why not the third or fourth nearest?
- The choice of the area, and the size of the area studied, can completely alter the result and make a clustered pattern appear regular and vice-versa.
- Although the NNI may suggest a random pattern, if a controlling factor (e.g. soil type or altitude) is randomly distributed, the vegetation is in fact anything but randomly distributed.

Graphical techniques: charts

Bar charts

Bar charts are one of the simplest ways of representing data (Figure 10). Each bar in a bar chart is of a standard width, but the length or height is proportional to the value being represented. There is a range of bar chart types. A simple bar chart shows a single factor.

A multiple bar chart can be used to show changing frequency over time (e.g. monthly rainfall figures). A compound bar chart involves the subdivision of simple bars. For example, a bar might be proportional to sources of pollution and be subdivided on the basis of its composition. A combination of compound and multiple bar graphs may show how sources of pollution change over time.

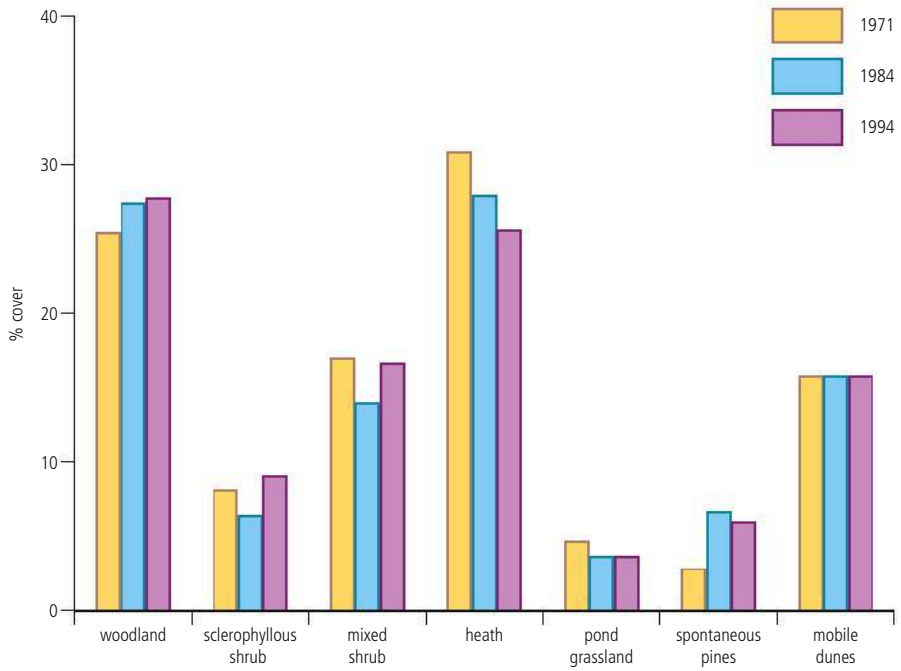


Figure 10 Bar charts showing percentage cover of vegetation in 3 years

Pie charts

Pie charts are sub-divided circles (Figure 11). They are used to show proportional variations in the composition of a feature (e.g. the proportion of sand, silt, and clay in a soil).

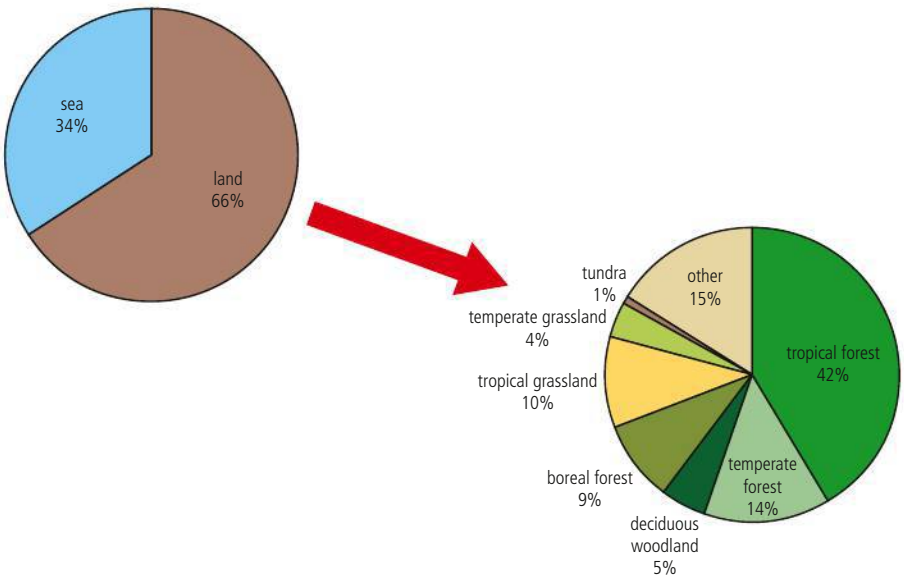


Figure 11 Pie charts showing global carbon fixing

The following steps should be taken when making a pie chart:

1. Convert the data into percentages.
2. Convert the percentages into degrees by multiplying by 3.6 and rounding up or down to the nearest whole number.
3. Draw the appropriately located circles on a map or diagram.
4. Subdivide the circle into sectors using the figures obtained in step 2.
5. Differentiate the sectors by means of different shading.
6. Draw a key explaining the scheme of shading and/or colours.
7. Give the diagram a title.

Figure 12 Dispersion diagram showing lichen cover on east-facing and west-facing gravestones

Dispersion diagrams

A dispersion diagram is a very useful diagram for showing the range of a data set, the tendency to group or disperse, and for comparing two sets of data (Figure 12). It involves plotting the values of a single variable on a vertical axis. The horizontal axis shows the frequency. The resulting diagram shows the frequency distribution of a data set. They can also be used to determine the median value, modal value, and the inter-quartile range.

Kite diagram

A kite diagram is a form of chart which allows you to view the relative distribution of different species along a transect (Figure 13). It is commonly used to show variations in sand dune succession, for example. Distribution is shown on the *y*-axis, species on the *x*-axis, and the abundance of each

species by the width of the columns. First, plot a series of bars representing the relative abundance of each species at each location. Then join the ends of the bars to form the kite shape.

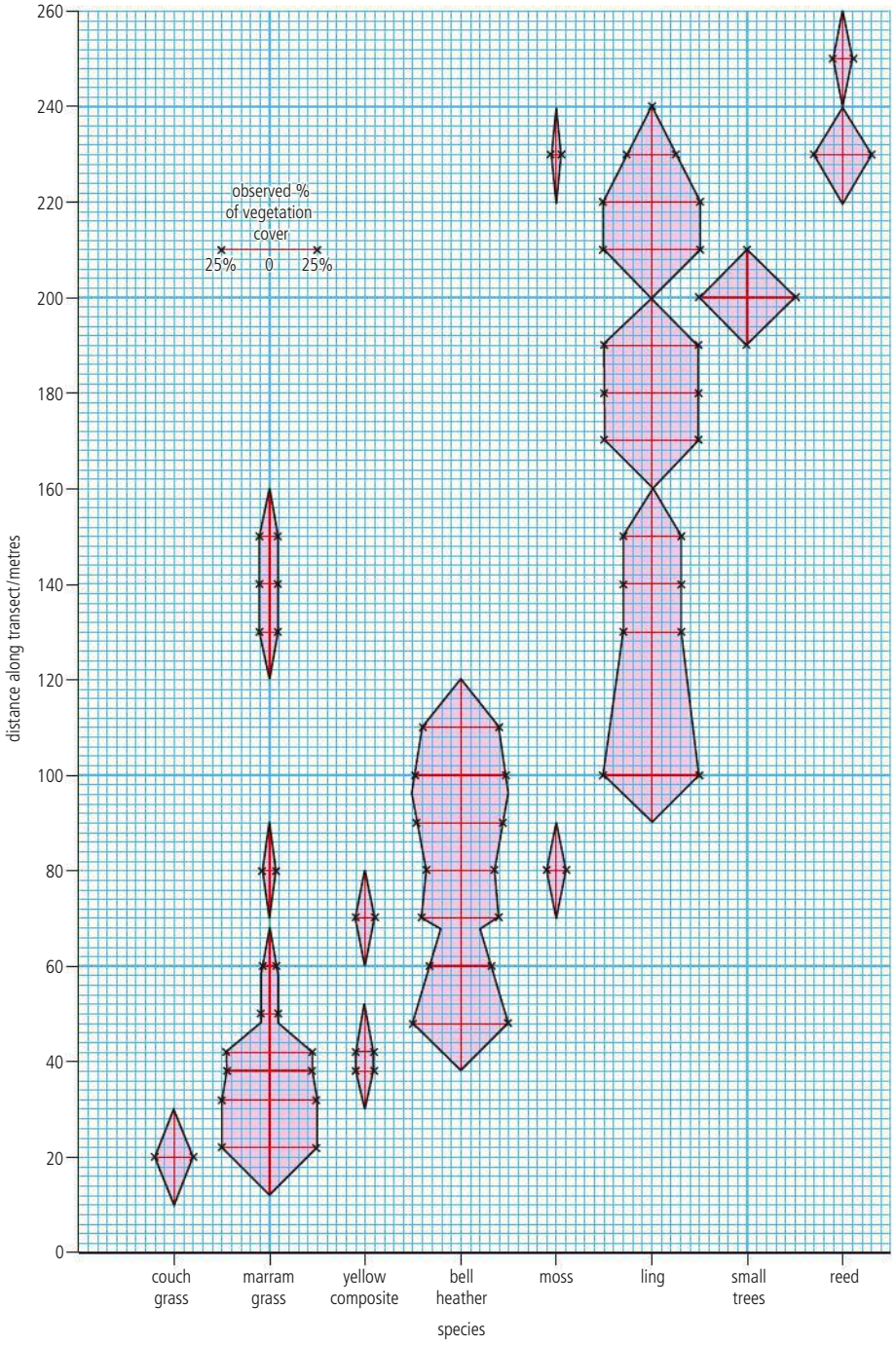


Figure 13 Kite diagram showing vegetation succession on Studland Beach, UK

Graphical techniques: graphs

Line graphs

Line graphs can be quite simple graphs which are used to show changes over time (e.g. temperature change related to the enhanced greenhouse effect) or over distance (e.g. variations in the populations of planktonic krill (Euphausiids), shrimps and crabs (decapods), ostracods, and fish in the North Atlantic (Figure 14).

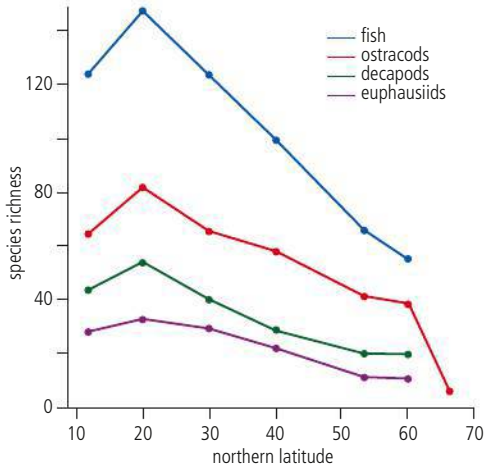


Figure 14 Line graphs showing species richness and latitude

In all line graphs, there is an independent variable and a dependent variable. In this example, the line of latitude is the independent variable and each species is a dependent variable. The independent variable is plotted on the horizontal or *x*-axis while the dependent variable is plotted on the vertical or *y*-axis. Nearer the equator there is more energy, more plankton, and hence more developed food chains.

Multiple or compound line graphs can show changes in more than one variable, for example changes in energy use over time. Such diagrams can reveal interesting relationships between the variables. On such graphs, data can be plotted in a number of different forms – in absolute terms, relative terms, percentage terms, or cumulative terms.

Flow lines

Flow lines show the volume of transfer between different groups or places. A good example is energy flow in an ecosystem (Figure 15). Alternatively, migration rates and direction could be shown using flow lines. In many cases, absolute data are used (Figure 15) but it is possible to use relative data.

As with all graphical techniques, it is important to:

- keep the background as simple as possible so as to avoid clutter
- choose an appropriate scale, so that extreme values can be shown without any loss of clarity
- provide a key, and give a title to the diagram.

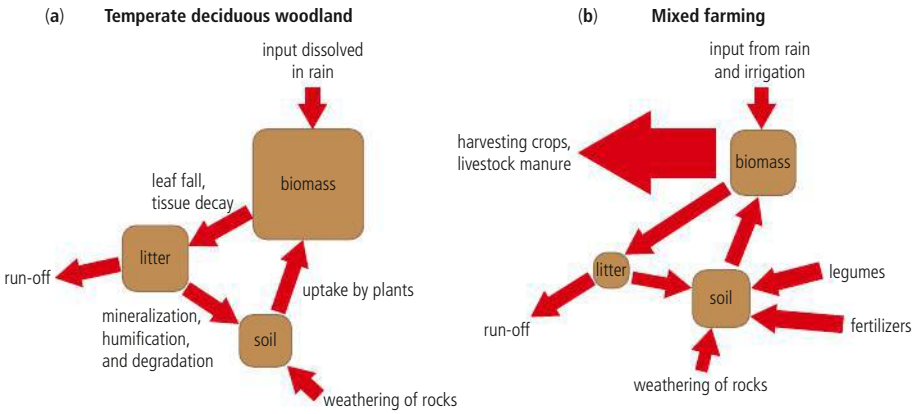


Figure 15 Flow chart showing nutrient cycles for (a) a temperate deciduous woodland and (b) an area nearby where the woodland has been cleared for mixed farming.

Triangular graphs

Triangular graphs are used to represent data that can be divided into three parts (e.g. soil consists of sand, silt, and clay; population consists of the young, adult, and elderly). These graphs require that the data have been converted into percentages, and that the percentages add up to 100%. On Figure 16, point A has 70% silt, 10% sand, and 20% clay. The main advantage of a triangular graph is that it allows a large amount of data to be shown on one diagram. In many cases, once the data have been plotted onto a triangular graph, groupings become evident. In the case of soil texture,

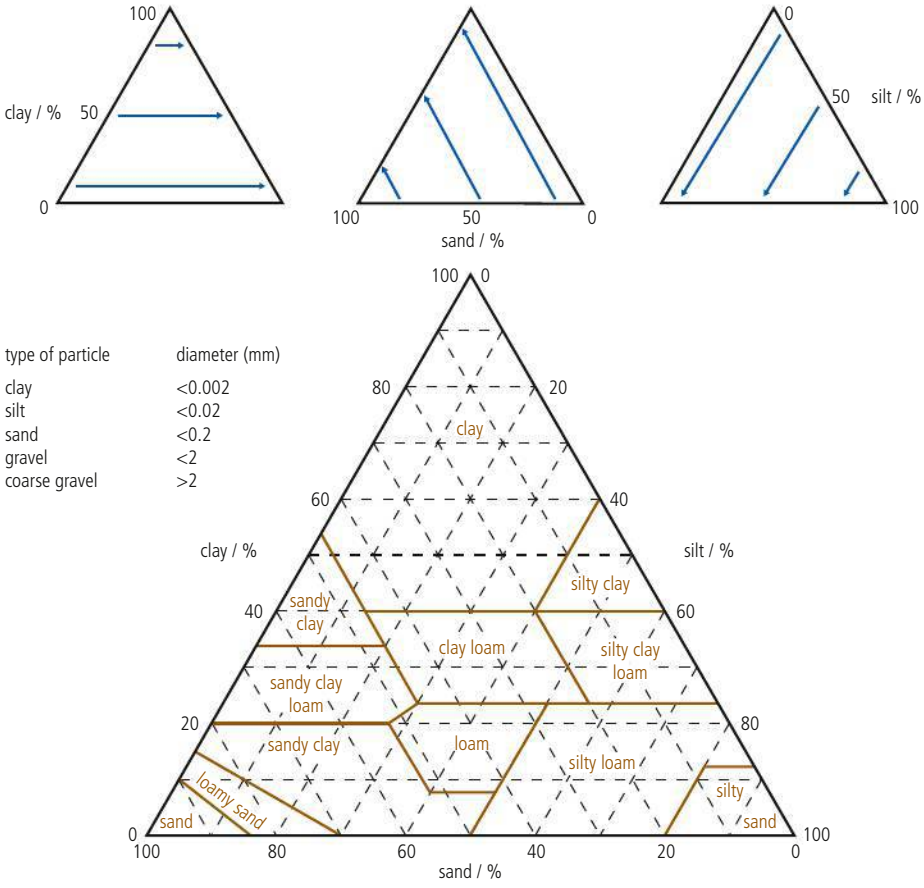


Figure 16 Triangular graphs showing soil structure

there are established soil textural groups. Triangular graphs can be tricky to construct. However, with care they can provide a reliable way of classifying large amounts of data which have three components.

Semi-log and double log graphs

Semi-log and double log graphs can be daunting at first. They allow scientists to compare small-scale features with large-scale ones, and the relative growth over time. This would not be as easy on an ordinary line graph.

The logarithmic scale compresses the range of values. It gives more space to smaller values but compresses the space available for the larger values – look at the space available for large and small values on the line graph and semi-log graph in Figure 17.

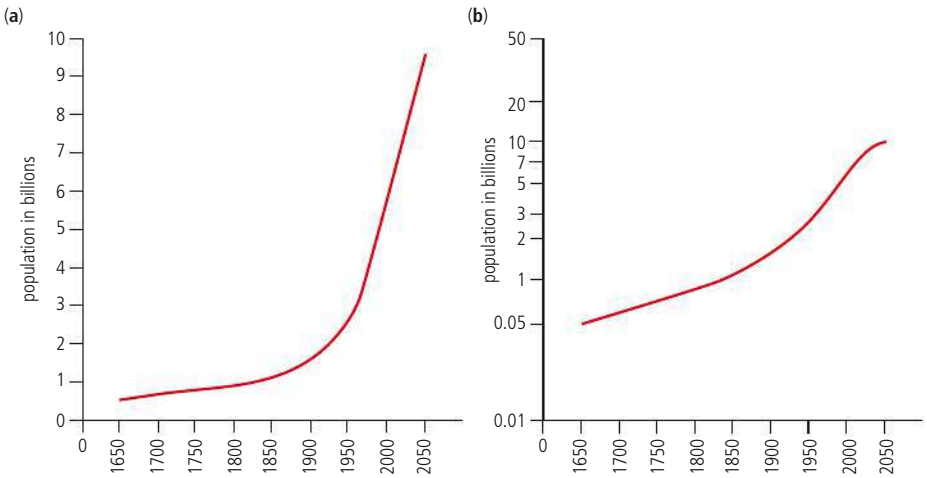


Figure 17 (a) Line graph and (b) semi-log graph showing numbers of survivors and lifespan

In a semi-log graph (aka as log-normal), one scale is logarithmic – usually the vertical one – while the other is a normal linear scale – usually the horizontal one. In the logarithmic parts of the scale, each of the cycles is logarithmic. This means that each cycle on the scale increases by the power of 10. For example, in the first cycle, values may be 1, 2, 3, 4, etc., whereas in the second cycle they would be 10, 20, 30, 40, etc., and in the third cycle 100, 200, 300, 400, etc. and so on.

It is important to realize that the logarithmic axis does not begin at 0 but some factor of 1 (e.g. 0.1, 100, 100 etc.); the horizontal axis can begin at any number and could even be nominal data such as the names of the months of the year.