BIO 109 / GEO 109.

A Brief Guide to Data Handling and Presentation

The following is intended as a quick reference guide – for more information you should refer to the detailed online guides and videos

Basic data analysis techniques

Measures of central tendency - averages – mean, median, mode
The most commonly used average is the 'mean'. This is calculated by summing all the values of a particular sample and dividing by the number of samples.

Sample of neck circumference of a dog (in mm): 280, 300, 280, 340, 280, 300

Sum of all samples = 1780. Number of samples = 6.
Mean = 1780 / 6 = 297 mm.

The mean can be calculated using the '=AVERAGE()' function in Excel.

The median is calculated by placing all the samples in order of size, and selecting the middle value (or the mean of the two middle values, if there is an even number of samples):

280, 280, <u>280, 300</u>, 300, 340

Median = (280 + 300) / 2 = 290 mm

The median can be calculated using the '=MEDIAN' function in Excel.

The mode is the most frequently occurring value. Here the value 280 occurs 3 times, the value 300 occurs twice and the value 340 occurs once. The mode, therefore is 280.

The mode – as well as the mean and median – can be calculated using the Descriptive Statistics option in the Data Analysis menu in Excel.

Which is the best measure of central tendency?
Unsurprisingly, the answer to this depends on the situation. The median is important in a type of statistical test called a *non-parametric test*. The mode can be useful for categorical data (i.e. green is the most common choice). However, in general, if you are unsure – it is best to use the mean.

Histograms and Frequency Distributions

Distributions can be calculated by working out the frequency, or the number of times, each particular value occurs. In the dog neck example above the size 280mm had a frequency of three, where as the value 300 had a frequency of 2. Equally a range of values could be given – so data points between 275 mm and 285 mm could be summed. Frequency distributions can be plotted on a histogram – using the Histogram option in the data analysis menu in Excel. Looking for normal distributions (a symmetrical bell shaped histogram) is important in data analysis, as normal distributions are important assumptions of many statistical tests.

Measures of dispersal and precision – range, variance, standard deviation, standard error and confidence intervals

Sample of neck circumference of a dog (in mm): 280, 300, 280, 340, 280, 300

From this sample, we calculated a set of average values. The mean value was 297mm.

Here is another sample:

Sample 2 of neck circumference of a dog (in mm): 200, 300, 280, 420, 280, 300

The mean of this sample is also 297mm, but the range of values is now much higher.

A very simple way of summarising these samples would be:
Sample 1: mean = 297mm, range = 280 to 340, n = 6 (n = number of samples)
Sample 2: mean = 297mm, range = 200 to 420, n = 6

The values of the biggest and smallest number form the 'Range', but this parameter is sensitive to outliers (extreme values in the sample). It is often better to give the interquartile range – where the values are put in order – like in the calculation of the median – and the top and bottom quarter of the values values are removed.

e.g. Sample: 3, 5, 6, 7, 3, 5, 56, -321, 3, 7, 8, 3, 4, 3, 5, 5, 6

in order of size: *-321, 3* ,3, 3, 3, 3, 4, 5, 5, 5, 5, 6, 6, 7, 7, *8, 56* (top and bottom ¼ underlined)

Interquartile range = 3 to 7.

The interquartile range is normally used in combination with the median. The Rank and Percentile option of the Data Analysis Toolbox in Excel can help with the calculation of interquartile range.

Variance

The variance is calculated by seeing how far away each value in a sample is away from the mean value. For example, the data point **8** is 3 units away from the mean of the sample, which, in this case is **5**. The actual calculation is more complicated than this, but variance can be calculated in excel using the '=VAR()' command.

Standard Deviation

Often called S.D. This is the square root of the variance. S.D. gives a good indication of how variable a sample is – the bigger the more spread out the samples are. In general, the range of data, except for extreme values, will be covered by the calculated range of:

Mean – 2*S.D. to Mean + 2*S.D.

More formally, 95% of the data points fall within the range of the mean +/- 1.96*S.D.

Standard deviation can be calculated by the '=STDEV' function in excel.

Standard Error

While standard deviation measures the 'spread' of data around a mean value, standard error (or S.E.) measures the precision of a sample mean. When a sample is taken – see below – it does not involve taking measurements from ALL individuals in the sample. For example – if the height of all 11,000 students at the University of Gloucestershire were measured, you could calculate the true ***population mean***. If you sample the population and measure the height of 100 students, you can calculate a ***sample mean***. The standard error is a good indication of how confident you can be that your sample mean is likely to be close to the true population mean. Low values of S.E. indicate a high confidence in your estimate of the population mean. (Obviously, without measuring all the population, you don't know what the population mean is – your value of sample mean could be identical – but you could still have a high S.E. This indicates that statistically you can not be confident of your value – and you should take more samples).

S.E. is calculated by dividing the S.D. by the square root of the number of data points sampled. So if your sample consisted of 25 data points and the S.D. was 2.13:

S.E. = 2.13 / (√25)   = 2.13 / 5   = 0.43

Also, S.E. can be calculated using the Descriptive Statistics option in the Data Analysis menu in Excel.

Confidence intervals

Confidence intervals (or C.I.) are a more formal way of indicating the degree of certainty that the population mean is close to the sample mean. A 95 % C.I. can be calculated by multiplying the S.E. by 1.96. Unlike the use of 1.96 * S.D. given above – the calculation of confidence intervals does not require the assumption of a normal distribution. Also, C.I. can be calculated using the Descriptive Statistics option in the Data Analysis menu in Excel.

The range created by:

Mean – 95 % C.I. to mean + 95 % C.I. give the confidence limits of a sample mean. You can therefore be 95% sure that the real population mean lies within these limits.

Confidence limits (the mean + and – the 95% confidence intervals) give rise to a basic form of statistical test. Many statistical tests test for differences between the means of samples (e.g. t-tests and ANOVAs – more next year). However, by considering if the upper confidence limit of one sample is less than the lower confidence limit of another sample, you can determine if their true means are significantly different or not.

It can be easier to plot mean ± C.I. on a graph (see below). If the 'error bars' overlap then the samples are not significantly different. If they do not overlap, then they are significantly different.

Although next year you will learn more powerful ways of testing for significant differences between means, this technique is very visual and can help you understand the results of the statistical tests you are doing.

Sampling

A sample should be representative of an entire population. The best way to sample, therefore is randomly, as this reduces any bias in the choice of what to sample.

Example: You need to know the average height of the population of students at the University of Gloucestershire – which has about 11,000 students. You could allocate each student a number from 1 – 11,000. You could then pick 100 numbers using a random number generator and measure the height of those students.

Example: You need to know the average depth of a rectangular reservoir. You can generate two sets of random numbers. The first gives a distance between 0 and the width of the reservoir, the second a distance between 0 and the length of the reservoir. You then locate the point in the reservoir using the numbers as coordinates and drop a weighted line until it touches the bottom. You repeat this sampling strategy for a suitable number of replicates (perhaps 100 times).

How many samples is enough?
In general, you should take as many samples as possible, as this increases the confidence in the sample mean being close to the true population mean (makes the confidence intervals smaller). However, at sample sizes above 30, since precision depends on the square root of the number of samples, you encounter a law of diminishing returns. You put in a lot of effort for minimal gains. If the S.E. of a sample is still high after 30 samples (another way to know when to stop is to

get to a point where the S.E. is < 10 % of the mean), then it may be better to modify your sampling strategy.

For example – measuring tree diameter in Pittville Park. Is there a difference in the mean diameter of oak and beech trees? You could sample 30 oak trees and 30 beech trees. If the value of your S.E. is still high compare to the mean, you may be better modifying your sampling design. You may wish to consult an old map or photograph that has the positions of trees 50 years ago on it, and only sample trees over 50 years old. This is called 'stratified' sampling, and can be useful. However, you should still ensure that trees are selected at random within the stratified design.

Note - If looking for a difference in the mean diameter of beech and oak trees and the standard error or confidence intervals of each sample mean are high – BUT even so, the confidence limits do not overlap, you have already shown a significant difference between the means occurs. In this case you do have enough samples (although you must be careful, as confidence limits act strangely at very low numbers of samples – you should always perform at least 5 replicates).

Data Presentation

Statistical data

When presenting summary statistics such as means, S.D., S.E. or C.I. you can normally give sufficient details in the text of your results section. Note – results sections of lab or field reports, or scientific papers, should contain *at least* one paragraph of text and refer to all your tables and figures.

Example:

The mean diameter at breast height of beech trees in Pitville Park was 1020mm (S.D. = 298, n = 30).

You should include the mean value (or the median or mode if more appropriate), the units that the mean is measured in (e.g. millimetres) and an indication of the variation or precision of your sample (ONE of variance, S.D., S.E. or C.I. – or interquartile range if the median is given) and the number of samples taken.

If you are presenting a lot of data – for example the diameter of 10 different species of tree, then it may be better to put the data into a table. You must refer to this table in the text. For example:

The mean diameters at breast height of the 10 species of trees measured are given in Table 1.

However, it is often common to make a statement about the values you are presenting in the table:

The mean diameters at breast height of the 10 species of trees measured varied form 480 mm to 1540 mm (Table 1).

Equally, if you were comparing these mean values to see if they were significantly different, you could refer to a figure with error bars showing the confidence limits of the diameter.

The mean diameter of oak trees was significantly higher than that of the other nine species of tree (Figure 1).

Tables and figures

Tables should be either placed at the appropriate position in the text (somewhere in the results), or included at the end of your report (this is largely up to you, and how happy you are with your word processing package – when writing for most scientific journals, the authors are asked to include figures and tables at the end of the manuscript – they are only placed in position by the publishers or printers of that journal – so you won't loose marks for doing this).

Tables and figures should be:

- Clearly laid out
- Normally in black and white – as a rule, if you ***can*** make the figure black and white, then do
- Text should be legible (generally use a non-serif font such as aerial on figures). Make sure the text is big enough to read on the printed out version of your manuscript
- Graphs should have labelled axis
- Graphs should NOT have a title, but should have a figure legend - below the figure.
- Tables should NOT have a title, but should have a table legend – above the table.
- Figure and table legends should contain enough information to understand the figure or table without reference to the text. Normally one sentence is not enough.
- Make sure all units are clearly indicated on the figure or in the table
- Make sure you only use a sensible number of decimal places in tables (normally one decimal place is enough)

Examples:

Table 2. Estimates of savings made based on examples of what can be grown in a typical garden (numbers indicate number of plants grown). Savings are comparisons with supermarket bought cheapest items or cheapest organic items (note negative saving for non-organic comparison in example 1).

Vegetable	Example 1 – pots on patio	Example 2 – small plot in mid sized garden	Example 3 – medium plot in large garden
Tomato	3	5	10
Pepper	0	5	10
Courgette	2	2	5
Cucumber	0	2	5
Runner Beans	4	8	20
Peas	4	8	20
Radish	0	30	50
Onion	0	10	50
Lettuce	0	10	30
Carrot	0	20	50

Savings	Non-organic = -£1.72
	Organic = £15.85
		Non-organic = £3.88
		Organic = £66.66
			Non-organic = £36.86
			Organic = £170.08

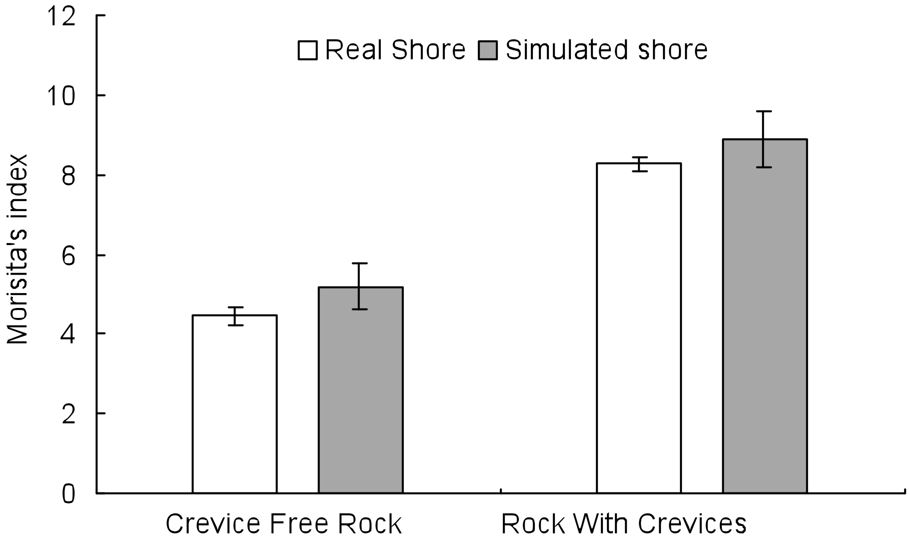

Figure 4. Comparison of distribution patterns obtained from real shores and from the computer simulation. The mean value (± S.E. n = 5) of Morisita's index of dispersal for snail distributions from a section of shore with long thin crevices (Site 1) and a section of shore with no crevices (Site 2).

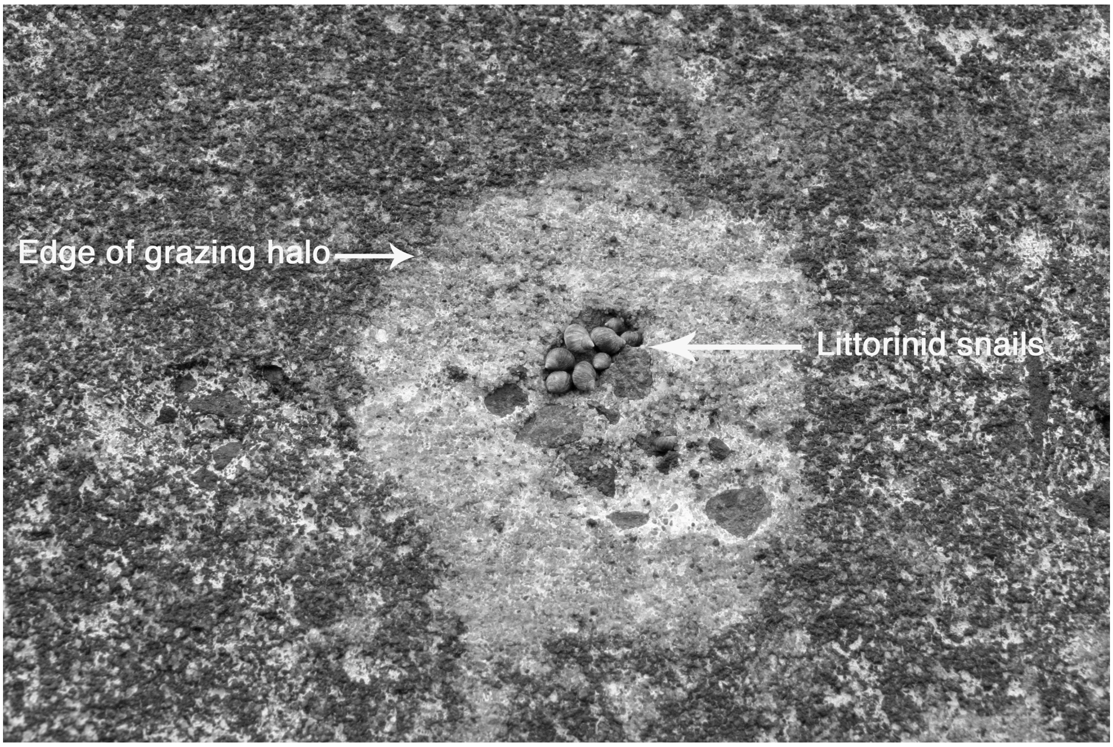

Figure 2. Littorinid snails in aggregations during emersion. Areas of rock close to the snails are visibly different in the amount of biofilm present. This area has been called a grazing halo, and most (> 90 %) of grazing occurs within these halos (Stafford and Davies, 2005).