

What is statistical analysis?

Statistical analysis is a way of looking at sets of data and deciding whether there is a **significant** difference or relationship between them. There are just a few statistical tests which are important in 'A' level Biology and Geography and we aim to make their use as painless as possible!

What do we mean by significant?

Results which appear to be meaningful could happen by chance. Statistical analysis looks at the data and allows you to decide on the **probability** of this happening.

- If there is less than 5% (1 in 20) probability of a chance or random happening then the result is said to be **significant (at the 5% probability level)**.
- If there is more than 5% (1 in 20) probability of a chance or random happening then the result is said **not to be significant**.

What type of data do you have?

Some statistical tests are sensitive to the type of data and you need to identify which type you have **before** you choose a test.

All measurements fall into one of two overriding categories

- **Continuous** - Measurements of length are continuous so fractions can be included. If you have rounded up to whole figures you can still consider your data to be continuous.
- **Discontinuous** - Generally counts of things (frequencies) whereby a fraction of an individual is impossible (although you may only have part of that individual). For example, if an insect with only its head and thorax falls into a pitfall trap it is a count of one, although it is only part of the individual. Remember that values derived from such data should be rounded to whole individuals.

Data then falls into several other subcategories in order of complexity starting with the simplest.

- **Nominal** - according to categories i.e. species of plant or colour of eyes.
- **Ordinal** - categories but in ascending or descending **ranks**.
- **Interval** - describes a data set in which the units are the same size throughout the scale i.e. the difference between 21 and 27 is the same as between 1 and 7. Temperature is a good example. However, they do not have a zero value so cannot be said to be x amount of times bigger.
- **Ratio** - This is the highest level of data complexity and can include all the previous categories. Such scales have zeros and are continuous. Linear dimensions are a good example.

In addition to the above there are other categories that the data may also fall into. It is not necessary to apply these to statistical tests unless stated.

- **Qualitative** - do your data describe a quality rather than a measurement? i.e. species, male/female or colour describe qualities and normally the data that follow are counts for these categories.

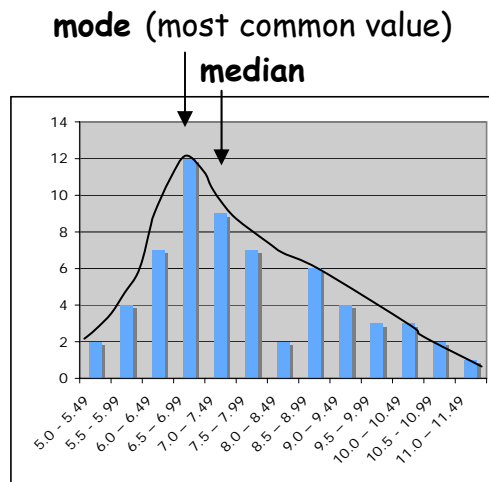
- **Quantitative** - measurements of a variable normally indicate a quantitative variable. Length of femur or weight of gonads are quantitative values.
- **Derived variables** - these are data that have not been measured directly but have been calculated from measurements. The most common derived variables are proportions (ratios and percentages). Proportions often do not follow normal distributions.

Distribution of data

A set of data may have a **normal** or **skewed** distribution. To decide you need to plot a frequency distribution histogram (number of individuals in each category). Here are 2 examples, the first from a population of Meadow Grasshoppers in early summer, the second from adult male bank Voles in autumn:

Table 1 (grasshoppers)

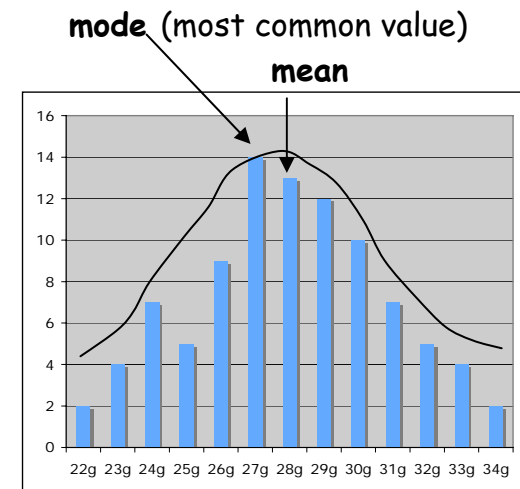
Length of grasshopper femur (mm)	Number of individuals
5.0 - 5.49	2
5.5 - 5.99	4
6.0 - 6.49	7
6.5 - 6.99	12
7.0 - 7.49	9
7.5 - 7.99	7
8.0 - 8.49	2
8.5 - 8.99	6
9.0 - 9.49	4
9.5 - 9.99	3
10.0 - 10.49	3
10.5 - 10.99	2
11.0 - 11.49	1



This is known as a **skewed distribution** and you should use the **Mann-Whitney 'U' test** for the significance of the difference between medians when comparing 2 sets of this type of data

Table 2 (adult male bank voles)

Wt of animal (rounded up to nearest g)	Number of individuals
22	2
23	4
24	7
25	5
26	9
27	14
28	13
29	12
30	10
31	7
32	5
33	4
34	2



This is known as a **normal distribution** and you should use the **t-test** for the significance of the difference between means when comparing 2 sets of this type of data

When the data is skewed you should use the **median** as your average. This is found by arranging the observations in ascending order and finding the middle value. In table 1 the median length of grasshopper femurs is 7.0 - 7.49 mm.

When the data is normally distributed (a more or less bell-shaped distribution) you should use the **mean** as your average. This is found by summing the observations and dividing by the number of observations. In table 2 the mean weight of adult male bank voles is 27.978g (rounded up to 28g for the graph).