Types of data: Statistics

Students need know that data that they collect can be one of several types. It is important to know what type it is because it helps decide how best to collect it and appropriate ways to display it. The first distinction is between:

- **Discrete data:** This can only have specific values (Male, Hot, 3 etc)
- **Continuous data:** This can have any value. This is often called **measurement data**.

Click on the link or use the keywords discrete data OR continuous data for more information.

These two data types can be further subdivided as follows:

	Data type	Definition	Examples
Discrete	**Category data - without order** (also called **Nominal**)	Data has a name only	Street, Road, Way Male, Female
	Category data - ordered (also called **Ordinal**)	Data has order, but does not have a numerical scale	Very happy, Happy, Unhappy, Very unhappy
	Whole number data	Data can have any whole number value	Day 1, Day 2, Day 3, … (simple time series data) Number of books (0, 1, 2, 3, …)
Continuous	**Measurement data**	Data can take any numerical value	Length of a pencil. It can be 8 cm, 9.1 cm, 9.48m. Time in seconds.

Category data (Level 1 and above of the curriculum)

This falls into two types:

Category - without order (Nominal data):

This is data with no order between the different categories

Examples:
1. There is nothing that is half way between a road and a street.
2. There is no sport half way between tennis and swimming (See *Choosing a sport* (ST8032*)*. It is often displayed in pictographs or bar charts. It should **never** be displayed with a line graph, as there is no meaning to the lines between any two consecutive points on the scale.

Category data - ordered (Ordinal data)**:**

This is when the categories can be put into order**.** There is no exact numerical relationship between the categories.

Example:
"Very happy" is not twice as happy as "Happy", but it is definitely **happier**.
Often this data involves a subjective judgement, for example, how do you define happy. For resources of this type click on the link or use the keyword, qualitative data.
It is unusual to draw line graphs or histograms for ordinal data. Bar charts are better.

Whole number data (Level 2 and above of the curriculum)

This is data that can only take on numerical values like 1, 2, 3, etc. Zero (0) is also possible.

> *Example*:
> Number of books must be a whole number (*Reading books* (ST8701)).
> It is unusual to draw a line graph or histogram, because 1.5 has no meaning. Bar charts are better.
> Sometimes statisticians draw line graphs or histograms when there are lots of whole numbers involved (ten or more is a simple rule of thumb) — *Picking peaches* (ST8043).

Measurement (continuous) data (Level 4 and above of the curriculum)

This can take on any value, so line graphs are appropriate.

> *Examples*:
> 1. The length of a pencil, which can take a range of values (say from 3cm to 20 cm, e.g. 13.84cm).
> 2. Race times (*Women's marathon* (ST8059)).

Time series data (Level 3 and above of the curriculum)

This refers to situations where other variables are being measured at different times.

> *Simple time series data* (Level 3 and above of the curriculum)
>
> These should just use whole numbers, i.e., it is treated as **whole number** data. The variable being collected would also be ordinal or whole number data at Level 3.
>
> *Examples*:
> 1. The number of cats taken to a vet in a week (Cats at the vet (ST8707)).
> 2. The hours of sunshine over a week (Hours of sunshine (ST8670)).
>
> *Time series data* (Level 4 and above of the curriculum)
>
> The curriculum considers time as **measurement** data, i.e., it can take on any value.
>
> *Example*:
> Food price changes over time (*Food prices* (ST8718)).
> For resources of this type click on the link or use the keyword time-series.
> Time series data is often plotted with line graphs, but are also plotted as bar graphs where times are clumped together, e.g., days of the week (*Hours of sunshine* (ST8670)).

Points to note:

- The vertical axis (y-axis) of many statistical graphs is frequency data, which is **whole number** data.

- It is acceptable to draw line graphs or histograms if the horizontal axis (x-axis) is **measurement** (continuous) data [or "near continuous", i.e. has a large number — at least 10 numerical categories].

- Often a **measurement** (continuous) scale underlies an **ordinal** or a **whole number** scale:

  1. Shoe size is **whole number** (discrete), but the underlying measure is foot length which is **measurement** (continuous) data. Even half sizes are still not really measurement but "whole number", because there is nothing between size 8 and 8 ½.

  2. Very big, Big, Small, Very small is **ordinal**, but underneath this is length, which is **measurement** (continuous) data.

  3. Money in your wallet is **whole number**. There is nothing between 10c and 20c.
     But banks calculate interest on money exactly, which is measurement data.

  4. In the resource *Wind speed*(ST8747), the mid-points of the bars could be joined to make a line graph, as there is the underlying continuous variable of time. This helps show trends in the data, especially if the underlying variable is time (e.g., *Number of deaths* (ST8041)). This is acceptable, even though there is no data point between 1993 and 1994 for example.

- **Do not** use line graphs with discrete data on the x-axis, especially if the data **is unordered category** (nominal) data. The exception is when you have a large number of categories of whole number data (e.g. number of children in different classes).

- With **whole number** data it is correct to say "The *number* of (people, cars, books etc.)…".
  It is wrong to say "The *amount* of (people, cars, books etc.) …".

- With **measurement** data it is often correct to say "The *amount* of (rain, time, etc.)…".
  It is wrong to say "The *number* of (rain, time, etc.) …".
  Often we use more specific words like 'length', 'weight, 'depth' etc. instead of 'amount'.

- **Measurement** (continuous) data allows inferences to be made with far less data than with category (discrete) data.