# Week 2 glossary

**Here are the words that you will encounter this week. You can also [print the full glossary](https://www.stat.auckland.ac.nz/~wild/d2i/articles/data-to-inzight-glossary.pdf" \o "Download the glossary) of words and refer to them during the course.**

The Glossary pages are here for reference purposes. They are not sets of definitions to be learned. Don’t worry about the meanings of words until after you have met them in the videos, etc. The Full Glossary is provided to be used like a dictionary. Week 2 introduces an unusually large number of glossary words.

(Different subject-fields sometimes use different names for the same statistical or data-analysis idea. This is why you will often see “Alternative names: …” at the end of a definition. These are not things you need to know for the course. They’re just there in case you already know something by another name.)

**Alpha-numeric**: Consisting of numbers and letters.

**Background Variability**: The extent to which the individual values within a group vary when compared to the central value(s) of that group. When looking at centres of data on their own this could mislead, so we tend to look at this in comparison to the background variability. (*See video 2.12.* )

**Bar chart**: A form of graph we use for categorical variables to display the percentages falling into each category. (See Week 2 video, “Categorical Variables” .) (*Alternative names: bar graph, bar plot, column chart.*)

**Bimodal**: When two peaks are evident in a graph of the distribution of a numeric variable. (See the Week 2 article “Features of numeric variables”.)

**Box Plot**: A summary-graph for displaying the distribution of a numeric variable. (See the Week 2 article “Features of numeric variables”). (*Alternative names: box and whisker plot.*)

**Centre**: The idea of where the “middle” of the set of observations is. (*Alternative name: Average, Location.*) (*Special cases: mean, median.*)

**Dot plot**: A form of graph for displaying the distribution of a numeric variable. The form we use is a special case, a *stacked* dot plot (See the Week 2 video, “Numeric Variables”.)

**Frequency**: The number of times a value (or category) occurs. (*Alternative names: count, tally.*)

**Histogram**: A graph made up of vertical rectangles that display the distribution of a numeric variable. The range of the data is divided into class intervals which form the bases of each rectangle. The height of each rectangle is set so that the area of the rectangle represents the relative frequency with which values fall into that class interval.

**Interquartile range (IQR)**: A measure of spread for a distribution of a numeric variable. It gives “the length of the middle half of the data”. Calculated by the difference between the 3rd and 1st quartiles. (See the Week 2 article “Features of numeric variables”.)

**Lower quartile**: (See quartile).

**Mean**: A measure of the centre for a distribution of a numeric variable. The total of all values divided by the total number of values.

**Median**: A measure of the centre for a distribution of a numeric variable. The “middle value”. It splits the data in half with half the observations at or above and half at or below. (See the Week 2 article “Features of numeric variables”.)

**Multiplicative**: A descriptor for comparisons made as in these examples: The second item looks twice as big (or long, or high) as the first, or 3 times as big, or half as big.

**Nominal variable**: A categorical variable in which the categories have no natural order.

**Oddities**: Anything in the data that looks strange or odd. Things that make us wonder, “Is that a mistake?”

**Ordinal variable**: A categorical variable in which the categories have a natural order.

**Outlier(s)**: Value(s) that lie so far away from the bulk of the data that they look odd and make us wonder, “Is that a mistake?”

**Overlap**: A visual notion. The degree to which two boxes, lines, or sets of dots extend over common values.

**Pie chart**: A graph for displaying the relative frequencies of a categorical variable. A circle is divided into sectors according to the relative frequency of each category.

**Quartiles**: Comes from separating a numeric distribution into four groups, each containing equal numbers of values. The 1st quartile (or lower quartile) is the middle of the lower half of the data and the 3rd quartile (upper quartile) is the middle of the upper half of the data. (See the Week 2 article “Features of numeric variables”.) (*Special cases: 1st quartile = lower quartile, 3rd quartile = upper quartile.*)

**Relative Frequency**: The number of times a value, or interval or category, occurs divided by the total number of occurrences (=frequency/number of observations).

**Shape**: Used to talk about the outline (or profile) of a plot of the distribution of a numeric variable. (See the Week 2 article “Features of numeric variables”.)

**Skewed**: The lack of symmetry in a distribution of a numeric variable. Positively skewed is when the data are piled up on the left and the tail extends out to the right. Negatively skewed is when the data piled up on the right and there is a long tail to the left. (See the Week 2 article “Features of numeric variables”.) (*Special cases: Positively (right) or negatively (left) skewed.*)

**Spread**: The idea of the degree to which values of a numeric variable differ from one another (vary), or, visually, are spread out along the axis. (*Alternative names: variability, variation.*)

**Stacked bar chart**: A graph for displaying the relationship between two categorical variables. Constructed by taking a bar graph for one categorical variable and subdividing each bar according to the percentages of the second categorical variable. (*Alternative names: Segmented bar chart.*)

**Subset**: Used in this course in its everyday, nontechnical sense - a collection of things that is part of a larger collection of things.

**Symmetrical**: Something that has the same shape reflected on both sides of some axis. (See the Week 2 article “Features of numeric variables”).

## Upper quartile: (See quartile).

**Variability**: (See spread.) The extent to which we get different values for different individuals (or in some contexts different values at different times).

## Common Questions

***Are “Background Variability” and “Spread” the same as Range, Variance and Standard Deviation?*** (*and similar questions about means and medians*).  
Background variability and spread are general ideas. Range, interquartile range and standard deviation are ways of trying to capture these ideas as “measurement”. Variance is the square of standard deviation. Similarly mean, median, mode are ways of trying to capture the general idea of “centre” as measurement. They all do it differently, tend to produce somewhat different answers and have different strengths and weaknesses; see **Steps 2.7 - 2.9**.