Full Glossary

Every discipline has a vocabulary, and sometimes words have a technical meaning which is different from the everyday meaning we may be used to. Data Analysis is no exception.

(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.)

Here are the words that we encounter in this course.

Additive decomposition: Breaking a time series down into 3 parts, trend, seasonal effects and residuals where it is assumed that there are *constant underlying seasonal swings* (e.g. of the same size every year) *that add to the trend value*.

Alpha-numeric: Consisting of numbers and letters.

Artefacts: Artificial patterns caused by deficiencies in the data collection process.

Association: Variables are said to be *associated* if there is a relationship between them (that is too strong to be likely to have arisen simply by chance). The short article *"Association and Correlation"* (Step 4.5) explains association pictorially for the case of two numeric variables and scatterplots including the following special cases. (*Special Cases: association versus no association, strong association versus weak association, positive association versus negative association, correlation.*)

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.*)

Biased selection processes: An inadequate selection process leads to systematic biases. (See *Selection process*.)

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".)

Bootstrap confidence interval: A confidence interval formed using bootstrap re-samples taken from the data.

Bootstrap distribution: The distribution of the estimates calculated from each of the bootstrap re-samples taken from the original sample (our data). (*Alternative name: Bootstrap re-sampling distribution*.)

Bootstrap re-sampling: Repeatedly re-sampling from the sample (our data) with replacement.

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*.)

Categorical variable: A variable whose values are names or codes for different groups (or categories). (*Alternative names: qualitative variable, factor, class variable*.) (*Special cases: nominal variable, ordinal variable*.)

Causal conclusion: A "this is what made it behave like this" conclusion.

Causation: The process of something making some phenomenon happen.

Cause: Something that makes some phenomenon happen.

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

Cluster: A distinct grouping of values that is separated from other groupings of values. This same basic idea appears in Weeks 2 and 3 in slightly different settings. (See the Week 2 article "Features of numeric variables" and the Week 3 video "Relationships between numeric variables".)

Colour gradient: A visual idea shown and named in the Week 4 video "Diving deeper with more variables" (~0.55-3:50) but already used towards the end of the Week 2 videos "Feature spotting" (~4:00-5:05) and "Comparing groups" (~6:40-7:40).

Comparison interval: An interval around the estimate constructed to enable us to make allowances for sampling error when making visual comparisons between groups. (Explained in Step 6.8)

Confidence bands: A band around a time-series forecast to allow for uncertainty. (They do not allow for uncertainties about the realism of the assumptions that went into the prediction.)

Confidence Interval: An interval around the estimate of a population parameter calculated to allow for uncertainty due to sampling error. We have almost exclusively used 95% confidence intervals which cover the true population value for 95% of samples taken (and miss for 5%).

Confounder: A variable that causes changes in both the outcome variable and the predictor of interest. (*Alternative names: Lurking variable, Confounding variable.*)

Correlation: Correlation between variables is a special case of *association* (see above). The short article "Association and Correlation" (Step 4.5) explains correlation pictorially including the following special cases. (*Special Cases: Pearson correlation versus Spearman's rank correlation, strong correlation versus weak correlation versus no correlation.*)

Cubic curves: (in the context of scatterplots): Useful for capturing trends that are gentle curves with up to two bends. (See the Week 4 video "Lines, curves and smoothers" and the end of the article "Interpreting the slope of a trend line", ~1:08-1:26.)

Decomposition plot: A plot built up of 3 parts, the basic Time Series plot with trend, the Seasonal Swing (additive or multiplicative) and the Residuals.

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".)

Effect size: The size of the difference between treatment-group centres. We need to distinguish between the unknown true effect size and that observed in the data.

Entities: The individual "things" we are recording data about. (*Alternative names: individuals, units, cases.*)

Experiment: A study in which the researcher controls (or manipulates or changes) the conditions experimental units experience.

Forecast: Predict.

Frequency: The number of times a value of a variable, or a 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.

Holt-Winters method: A good general-purpose forecasting-method for time series data where the trend is basically monotone (either rising or falling, not both).

Inference: See *statistical inference*.

Intercept: (See the article Step 4.4 for more detail): The *intercept* of a trend line for a scatterplot which has Y plotted against the vertical axis and X plotted against the horizontal axis is the average Y-value when X=0. This is often not a meaningful value as X=0 is often outside the range of the data.

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".)

Jittering: A technique used in scatterplots to deal with overprinting in which a little bit of random horizontal and/or vertical displacement is added to 'break apart' points sitting on top of one another. (See the Week 4 video "Overcoming perceptual problems", ~1:57-2:10.)

Linear trend: (in the context of scatterplots) A straight line fitted to the data in a scatterplot to try to represent the trend. (See the Week 4 video "Lines, curves and smoothers" (~0:55, 1:22, 3:03-end) and article "Interpreting the slope of a trend line".)

Lower quartile: (See quartile.)

Margin of error: An estimate of the likely size of the sampling error in an estimate of a population parameter ("likely" in the sense that the sampling error is very unlikely to be larger than the calculated margin of error). Confidence intervals can often be obtained by using *estimate* ± *margin of error*.

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.

Measurement Error: The difference between the measured value of something (say a person's height) and its actual (real) value.

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".)

Missing value code: A code that is used to tell the software program that no value has been recorded for this cell (either an empty cell, NA or NULL for iNZight).

Missing value: No information has been recorded for this cell.

Multiplicative decomposition: A time series which has been broken down into 3 parts, trend, seasonal effects and residuals where it is assumed that there are *constant underlying seasonal effects* that *multiply the trend value* rather than adding to it.

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.

No relationship: (See relationship.)

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

Non-significant: If a significance test of a difference produces a *p*-value that is not small (commonly if it is bigger than 5%) it is said to be "nonsignificant". (See the entries for *p-value* and *significance test*.) People often **mistakenly** write that there is *no difference* between groups or no difference between the effects of treatments if a significance test turns out "nonsignificant". (There is a big difference between not having enough evidence to conclude that a true effect exists and having evidence that there is no true effect.)

Numeric variable: A variable for which all of the values are numbers (e.g. from counting or measuring). (*Alternative name: quantitative variable.*) (*Special cases: discrete variable, continuous variable, interval variable, ratio variable.*)

Observational Study: A study where our data comes from observing and recording things as they are in the world, or as they unfold over time, without the investigator actively changing anything.

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.

Outcome variable: The variable of primary interest whose behaviour we may want to explain or predict using one or more predictor variables. (*Alternative names: response variable, dependent variable, output variable*.)

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.

Overprinting: A problem with scatterplots when points sit on top of one another so that we are unable to tell how many points are sitting at a given position. This can lead to very misleading impressions of what the data is saying. (See the Week 4 video "Overcoming perceptual problems".)

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.

Practically significant: A difference is practically significant if it is big enough to have a real-world impact.

Prediction: Often used in its everyday, nontechnical sense. Also used for a value predicted for an outcome.

Predictor variable: A variable that we want to use to try to predict or explain the behaviour of an outcome variable, or just to investigate whether this might be possible. (*Alternative names: explanatory variable, independent variable, exogenous variable, input variable.*)

p-value: The probability of getting a result at least as extreme as the result seen in the data (in our case, assuming no real differences between treatment groups). The *p*-value comes from the tail proportion(s).

Quadratic curves: (in the context of scatterplots): Useful for capturing trends that are gentle curves with up to one bend. (See the Week 4 video "Lines, curves and smoothers" (~1:00-1:07, 1:22) and the end of the article "Interpreting the slope of a trend line".)

Quarterly data: Data that is reported 4 times a year covering periods of 3 months.

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.*)

Random Error: Errors or changes caused by a random process (or sometimes just appearing as if they have been).

Random sampling: The process of selecting a sample from a population where all units have an equal likelihood of being selected for the sample.

Randomisation Test: A procedure used to investigate whether an experimental difference could have occurred purely by the luck of the randomisation draw (purely by chance).

Randomisation variation: The differences in group summaries produced by random allocation to "groups" (or random labelling) and nothing else (random labelling alone).

Randomised Experiment: An experiment where we change conditions purposefully and use a random process to decide who (or what entities) will be subjected to what conditions.

Rectangular data: Data organised and stored in such a way that each row corresponds to an individual entity and each column corresponds to a property recorded for these entities.

Relationship: A pattern that connects two or more variables. (In practice we apply this to patterns we believe are strong enough that they would be unlikely to be generated by a purely chance mechanism.) There is **no relationship** when learning the value of one variable would tell you nothing new about the likely value of the other. (*Alternative name: association.*) (*Special case: correlation.*)

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

Reliability (of measures): The degree to which repeated "measurements" of the same thing give consistent answers.

Representative: A non-technical word used to convey the idea of something we are seeing reflecting real features of what we are interested in. Often used in connection with whether data we have is "representative" of the population we are interested in. If the data we are getting is systematically biased it is not representative.

Re-randomisation distribution: the distribution of the artificial "group" differences generated by repeated random re-labelling.

Re-randomisation: The process of randomly re-labelling experimental units (or entities in a sampling situation).

Re-sampling: Taking samples from the sample itself. In the case of bootstrapping this sampling is done with replacement.

Residuals (in time series): The difference between the original series and what "Trend+Seasonal Swing" gives us.

Running quantiles: (in iNZight): A technique for large data sets only. Curves applied to scatterplots that attempt to run through the 50th percentile (median) and as many of the 10th, 25th, 75th and 90th percentiles of the outcome values for any value of the predictor variable. As many of these curves are drawn as is feasible given the size of the data set. (See the Week 4 video "Overcoming perceptual problems", ~4:10-5:05.)

Sample size: The number of units, individuals or values selected from a population or sometimes just the number of entities we have data on.

Sampling error(s): A sampling error is the difference between the estimate we get from our sample and the corresponding value (say a mean income) in the population. We often use the plural version - the errors we make by using sample quantities to estimate population quantities.

Sampling variation: The way estimates made from samples vary (change) from sample to sample.

Scatter: (in a scatter plot): The extent to which the values of the outcome variable are scattered above and below the trend. (See the Week 3 video "Trend, scatter and outliers".)

Scatterplot: The standard form of graph for displaying a pair of numeric variables. (See the Week 3 video "Relationships between numeric variables".) (*Alternative name: scatter graph.*)

Seasonal patterns: A basic pattern that repeats regularly over a period of time. (E.g. a repeating months-of-the-year, days-of-the-week, or hours-of-the-day pattern.)

Seasonal series: A time series with a seasonal pattern.

Selection processes: In this course, the word is used very broadly to refer to the process by which some things get recorded and make it into our data set while others don't. It could be an accidental process. In survey sampling it is used for the way in which entities are *purposefully* selected to being the sample.

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".)

Side-by-side bar chart: A bar chart for investigating the relationship between two categorical variables where, for each outcome category in turn, we put all of the bars for the predictor categories together "side by side". (See the Week 3 video "Relationships between categorical variables".)

Significance test: A test of whether a group difference seen in the data could be generated by a relevant chance mechanism - like the luck of the randomisation draw. (*Alternative names: Statistical hypothesis test.*) (*Special case: randomisation test.*)

Significance: There are two types of significance. (See the entries for *statistically significant* and *practically significant*.)

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*.)

Slope : (See the article Step 4.4 for more detail): The *slope* of a trend line for a scatterplot which has Y plotted against the vertical axis and X plotted against the horizontal axis is the change in the average value of Y associated with a 1 unit increase in X. (*Alternative name: gradient.*)

Smoother: (in the context of scatterplots): A very flexible way of capturing a trend. In iNZight, the degree of flexibility and ability to take bends is controlled by a user-controlled slider. (See the Week 4 video "Lines, curves and smoothers", ~1:33-2:55.)

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.*)

Statistical Inference: The process of drawing conclusions about population quantities based on a data from a sample, or features of a process based on data from that process. (*Special cases: Confidence intervals, randomisation tests*.)

Statistically significant: A difference is *statistically significant* if it has a low - p-value (see *p-value*), commonly less than 5%. It is interpreted as, "the data provides evidence that a true difference exists." The smaller the p-value is, the stronger the evidence that a true difference exists. It says nothing about the size of a difference (for that you need a confidence interval) or whether it is of practical importance (for that you need to know about the size of the difference and whether differences of that size would have a practical impact).

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

Subsetting: An idea used heavily through Weeks 2-4 in which we divide the entities in our data set into different groups (subsets) on the basis of their values for one or two subsetting variables. We then make separate graphs of the same type for every subset and presented them either as a matrix of tiled graphs, or playing through them like a movie. (See the Week 2 video, "Time travel" (whole movie), the Week 3 video "Changes across subgroups" (~0.50-1:55, 3:56-end), and the Week 4 videos "Diving deeper with more variables" (~4:20-end) and "Our changing health and wealth" (~2:30-the end).)

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

Systematic biases: Consistent biases caused by the way a system or process functions. Not random errors that bounce around going positive and negative but errors that tend to be in the same direction. (*Alternative names: Systematic error*.)

Tail proportions: A measure of how close a result is to the edge of the distribution. The probability of getting a result at or above that observed in the data if we are nearer to the right hand end of the distribution, or at least as small if we are nearer to the left-hand edge of the distribution.

Tile density plot: A tile density plot looks like a crude scatterplot. In a tile density plot, the scatter-plotting region is divided into a set of rectangular tiles. If there is no data in the area covered by a tile it is coloured white. If there is data in the area covered by the tile it is coloured with the depth of colour (darkness) determined by the number of data points in the area covered by the tile. In iNZight, the tile containing the most data points (max) is coloured black. A tile area which has only one point in it is coloured a shade of grey which I'll call floor-grey. Tiles with between 1 and max points are coloured on a scale between floor-grey and black of a darkness determined by the number of points falling into the tile. The sizes of the tiles are controlled by *Add to plot > Change plot appearance* and using the *Size of symbols* slider. The darkness of floor-grey is controlled by *Add to plot > Change plot appearance* and using the *transparency* slider. (See the Week 4 video "Overcoming perceptual problems", ~6:02.)

Time-series data: Data collected over time where we are interested in looking at changes over time and predicting.

Transparency: A technique used in scatterplots to deal with overprinting in which we make the symbols semi-transparent. Where there is a lot of overprinting, the symbols will be darker and where there are single or few points overprinted, the symbols will be lighter. (See the Week 4 video "Overcoming perceptual problems", ~2:20-4:10, 5:32-5:42.)

Treatment group: A group of experimental units who have, or will have, been given the same "treatment".

Treatment (in experiments): A treatment is something that is applied to (done to, performed on) experimental units, e.g. being given a drug.

Trend: A line or curve that tracks "the average value of the outcome variable at or near a given predictor value". This idea is much more obvious visually. (See the Week 3 video "Trend, scatter and outliers", or for time-series contexts, the introductory Week 8 video "Time series data".)

Type-1 error rate: The probability of concluding that there is a true difference between groups in a situation where there is no true difference.

Upper quartile: (See quartile).

Validity (of measures): The extent to which a measure is "measuring the right thing". (*Alternative names: Measurement validity.*)

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

Variable: A property that we record for each entity, e.g. a measurement, or one of a set of group labels. (*Alternative name: feature* [*in machine learning*])