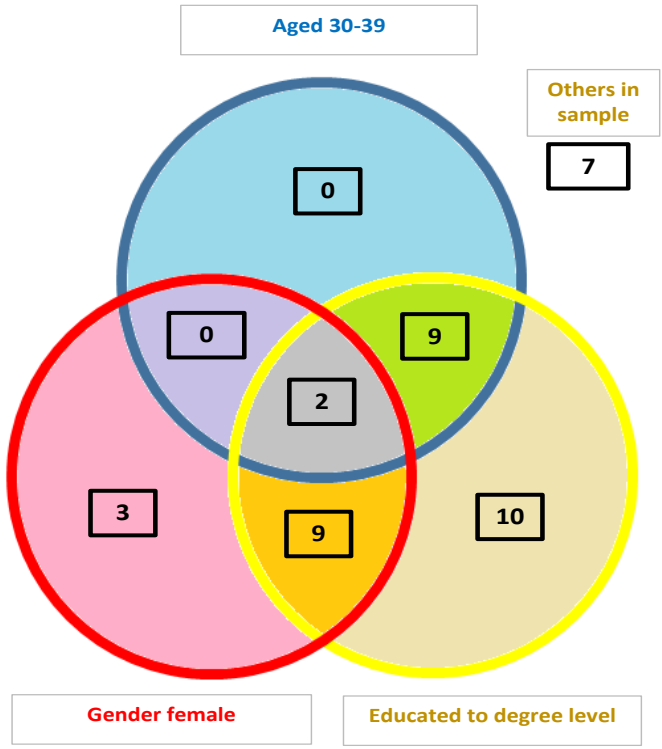
STATISTICS FOR LINGUISTS USING EXCEL



Compiled by Martin Edwardes & Jo Lewkowicz

Contents

This booklet should be read in association with the Excel spreadsheet:

L6-7 - Excel database for practical sessions

Basic concepts

Estimation and Reckoning

Estimation is a useful first way of interpreting your data, and it can quickly reveal areas where your data may be inadequate or uninformative. It is not a reliable indicator of what your data is telling you, and it is subject to unnoticed bias by the interpreter; but it can indicate areas of interest in your data, or areas where you may need to gather more data.

Look at the **Raw Data** tab in the Excel spreadsheet.

- What can you tell just by looking at the data?
- What can you speculate just by looking at the data?

Sheet activities

Renaming a sheet	• Right-click on the sheet name tab of the sheet to be renamed. • Select **Rename**. • Overtype the old name with the new name. Press **ENTER**.
Locking a sheet	• Right-click on the sheet name tab of the sheet to be locked. • Select **Protect sheet**. • You can type in a password at this stage, but be careful: if you forget the password you cannot unlock the sheet. • Click **OK**.
Unlocking a sheet	• Right-click on the sheet name tab of the sheet to be unlocked. • Select **Unprotect sheet**. • If the sheet has a password you will be prompted to give it. (If you don't know the password you cannot unlock the sheet.) • Click **OK**.
Creating a new sheet	• Click on the plus button to the right of the sheet name tabs at the bottom of the screen.
Deleting a sheet	• Right-click on the sheet name tab of the sheet to be deleted. • Click on **Delete**. A sheet deletion cannot be undone, so be careful.
Copying data to a new sheet	• Go to the sheet with the data to be copied. • Select the data to be copied and click **Copy** on the **Home** tab (or press **CTRL+c**). • Open an empty sheet, or create a new sheet. • Paste the date by clicking **Paste** on the **Home** tab (or press **CTRL+v**).
Heading rows & columns	To keep heading rows and columns always visible: Click on the **View** tab, then click on **Split**. In the numbered rows margin (far left) move the row split marker until it is just below the rows you want to keep visible. In the lettered columns margin (top of sheet) move the column split marker until it is just to the right of the columns you want to keep visible. Click on **Freeze Panes** and select **Freeze Panes**.

Sorting

- Select ALL the data in the data series you wish to sort (not just the column of data you wish to sort!) If your data has a row of data headings, you should include this in your selection.
- Go to the **Home** tab and click on **Sort and Filter**. The click on **Custom Sort**.
- In the sort box that appears, make sure that **My data has headers** is ticked if you included a row of data headings. If there were no data headings then make sure **My data has headers** is NOT ticked.
- Select the highest sort level by data heading or by column letter, what you want to sort by, and the sorting order.
- If you wish to have a second level of sorting, click **Add level** and repeat the previous step. You can have up to 10 levels in a sort.

- Sorting is the easiest way to find the RANGE of a data series. Sort the data to find the highest and lowest values.

Formatting cells

Decimal places	• Select the cell or cells to be formatted • Right-click on the selected cells • Select **Format cells** • Select the **Number** tab • Select Category: **Number** • Raise or lower the number of decimal places and click **OK**
Dates	• Select the cell or cells to be formatted • Right-click on the selected cells • Select **Format cells** • Select the **Number** tab • Select Category: **Date** • Select the date format you wish to use and click **OK**
Borders	• Select the cell or cells to be formatted • Right-click on the selected cells • Select **Format cells** • Select the **Border** tab • Select the border format you wish to use and select or unselect the borders you wish to appear. When you have the borders you want, click **OK**
Unlocking a cell	When you lock a sheet, it is possible to leave some cells unlocked and available for input. *You must unlock the sheet to do this* (see **Unlocking a sheet**). • Select the cell or cells to be formatted • Right-click on the selected cells • Select **Format cells** • Select the **Protection** tab • Remove the tick on **Locked** to unlock the cell; click **OK**

Statistically useful Excel functions

AVERAGE	Mean of data series	**The arithmetically central point of a data series.** • Add together all the values in the data series • Divide by the number of items in the data series **=AVERAGE([First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, row 44*
AVERAGEIF	Mean of subset of data series	**The arithmetically central point of a subset of a data series.** • Identify all the items in the data series that are part of the subset • Add together all the values in the subset • Divide by the number of items in the subset **=AVERAGEIF([First cell in data series defining subset]:[Last cell in data series defining subset],"[value defining subset]",[First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, rows 58-63*
CHISQ.TEST	Chi squared test of independence	**Probability the difference between this data series and the general population could have happened by chance** • *The chi squared formula is complex and you do not need to know it for this module. If you really want to know what it is, see Excel Help, CHISQ.TEST* **=CHISQ.TEST([First cell in sample data series]:[Last cell in sample data series],[First cell in population data series]:[Last cell in population data series])** *See **Statistics** sheet, rows 68-71*
CORREL	Correlation coefficient	**A measure of the strength of the relationship between two data series. Values above 0.8 indicate a strong correlation; values between 0.8 and 0.5 indicate a weak correlation; values below 0.5 indicate insufficient evidence for correlation.** • *The correlation formula is complex and you do not need to know it for this module. If you really want to know what it is, see Excel Help, CORREL* **=CORREL([First cell in x data series]:[Last cell in x data series],[First cell in y data series]:[Last cell in y data series])** *See **Correlation**, cell D2*
COUNTIF		**The number of items in a subset of a data series.** • Identify all the items in the data series that are part of the subset • Count the items in the subset **=COUNTIF([First cell in data series defining subset]:[Last cell in data series defining subset],[value defining subset])** *See **Statistics** sheet, rows 51-56*
IF		**Tests the truth of a condition. If it is true then it performs one action, if it is false it performs another.** **=IF([test],[if true do this],[if false do this])** Some typical IF statements:

		• **IF(C3="Yes",1,0)** : if cell C3 equals "Yes" then show 1 in the formula cell, otherwise show 0. • **IF(A8<>7,A8*2,B15*D4)** : if cell A8 does not show 7, show the value of cell A8 x 2 in the formula cell, otherwise show the value of cell B15 multiplied by the value of cell D4. • **IF(A8>52,"Joker",A8)** : if cell A8 contains a value larger than 52, show the word Joker in the formula cell, otherwise show the value of cell A8. • **IF(A8<=52, A8,IF(B8="Tarot",C8,"Joker"))** : if cell A8 contains a value less than or equal to 52, show the value of cell A8 in the formula cell, otherwise: if cell B8 contains the word "Tarot", show the value of cell C8, otherwise show the word "joker". **For more on IF formulae see IF function in Excel Help.** *See **Statistics** sheet, column F; **Venn Diagram** sheet columns C, E, G to K, and N*
INT		**Gives the integer of a real (decimal) number by ignoring the decimal portion. So 8.3, 8.5 and 8.8 all give 8.** **=INT([cell to be integered])** *See **Statistics** sheet, column D*
LINEST	Line of best fit for correlation	**Gives the slope of the line of best fit for a correlation, and the value of y when x=0.** • Select two cells next to each other, and type the formula on the formula line • Two values will be returned: [the slope] in the first cell and [the value of y when x=0] in the second cell • You can use these values to calculate the line of best fit (*see **Correlation** sheet, column C*) • When you look at the formula in the cells it will have {} braces around it to indicate it is a multi-cell formula. • *The skewness formula itself is complex and you do not need to know it for this module. If you really want to know what it is, see Excel Help, SKEW* **=LINEST([First cell in x data series]:[Last cell in x data series],[First cell in y data series]:[Last cell in y data series],,TRUE)** *See **Correlation** sheet, cell E2 and F2*
MEDIAN	Median of data series	**The middle value of a data series if the data series were ordered numerically.** **=MEDIAN([First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, row 45*
MODE	Mode of data series	**The most commonly occurring value in a data series; if the data were displayed on a graph, it would be the tallest bar.** **=MODE([First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, row 46*

PEARSON	Pearson correlation coefficient	**A measure of the strength of the relationship between two data series. For details see CORREL.** **=PEARSON([First cell in x data series]:[Last cell in x data series],[First cell in y data series]:[Last cell in y data series])** *See **Correlation** sheet, cell D4*
SKEW	Skewness coefficient	**A measure of the lack of symmetry in a data series distribution. A negative value indicates a negative skew, a positive value is a positive skew. The size of the value indicates the range of the data series as well as the skewness.** • *The skewness formula is complex and you do not need to know it for this module. If you really want to know what it is, see Excel Help, SKEW* **=SKEW([First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, row 49*
STDEV.S	Standard deviation	**A measure of distances from the mean of the data series. More informative than the variance because it also indicates probability in a normal distribution.** • Square root of the variance (see VAR.S) **=STDEV.S([First cell in data series]:[Last cell in data series])** *See **Statistics** sheet, row 48*
SUM	Producing totals	**The total value of a selected set of data items.** **=SUM([First cell in data series]:[Last cell in data series])** *See **Venn Diagram** sheet, rows 44-45*
TDIST	p-value	**Probability that a data series with more extreme values could have come out of the population by chance. This value is not automatically calculated by Excel, so you need to apply the formula below.** • Calculate the Pearson correlation coefficient (PCC) • Find the count of items in the data series (n) • Calculate the PCC product (PCCP), where: PCCP=(PCC*√(n-2)/√(1-(PCC^2) **=TDIST([Pearson correlation coefficient product (PCCP)],[Count of items in data series (n)],[1 or 2 tailed test, usually 2])** *See **Correlation** sheet, F4 & **UK Popn** sheet, C9*
T.TEST	Student t-test	**Measures the probability that two data series are from the same population. An extension of the chi-squared test.** • *The t-test formula is complex and you do not need to know it for this module. If you really want to know what it is, see Excel Help, T.TEST* **=T.TEST([First cell in x data series]:[Last cell in x data series],[First cell in y data series]:[Last cell in y data series],[1 or 2 tailed test, usually 2],1)** *See **UK Popn** sheet, C19*
VAR.S	Variance	**A measure of distances from the mean of the data series. Variance emphasises larger distances and reduces smaller ones. Larger numbers indicate greater variation.**

- Add together the square of the differences between each data item in the series and the mean
- Divide by [the number of items in the series] -1

=VAR.S([First cell in data series]:[Last cell in data series])

*See **Statistics** sheet, row 47*

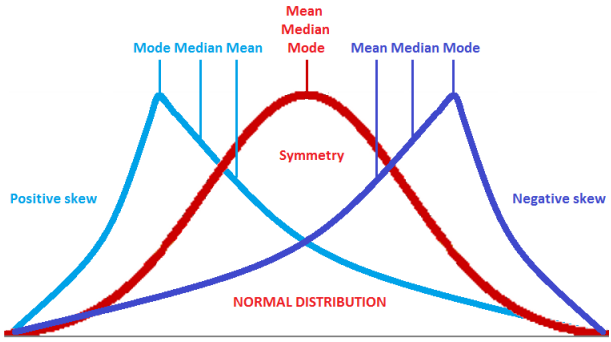
Types of distribution

Normal

A useful statistical fiction which assumes that data series tend towards a standard distribution of data values. The distribution forms a bell-shaped curve around the mean. The normal distribution is useful because about 68 percent of the data values are within one standard deviation of the mean, about 95 percent are within two standard deviations, and about 99.7 percent are within three standard deviations. By assuming a normal distribution you can make predictions about the relationship between your data and the population it is drawn from.

Skewed

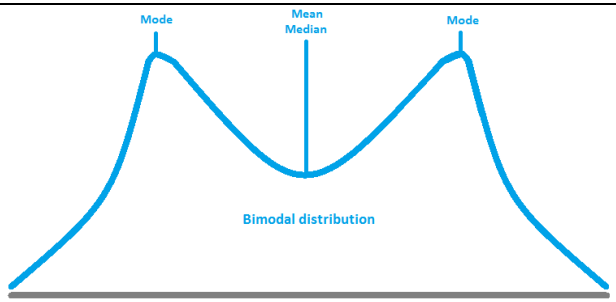
In a normal distribution, the mean, median and mode are all the same value because the data is distributed symmetrically. In a skewed distribution the mode is less than the median which is less than the mean (positive skew), or the mean is less than the median which is less than the mode (negative skew).



Bimodal

A bimodal distribution has two maxima or peaks with a minimum (or valley) between them. This is also known as a marmite distribution. Where the two modes are of different values, the larger is called the major mode and the smaller is the minor mode.

Multimodal distributions can be treated as two distributions by dividing the data series into two subsets based on the series mean. However, this does require some arbitrary decisions about how to allocate data items to one or the other subset; and, while it is a convenient way to apply standard statistical methods, it may not reflect the relationships in the data itself. It is a method that should be used with care.



Some Possible Test Questions (with answers)

Statistics

What is the mean and what can it tell you about a data series?	The mean is the arithmetic mid-point of a data series. By itself it tells us little, but when combined with other statistical measures it can tell us about the shape, size and range of the data series.
What is the median and what can it tell you about a data series?	The median is the logical centre of a data series: if all the data were sorted into order, the median would be the middle-most value. Where there is an even number of items in a data series, the middle two items are added together and divided by 2.
What is the mode and what can it tell you about a data series?	The mode is the frequency centre of a data series. The most frequent value in a data series is the mode. With highly skewed data series, the mode can be some distance from the mean, which gives a measure of skewness.
What does the difference between the mean and the median tell you about a data series?	The arithmetic centre and the logical centre of a data series are often separated. The distance between them can indicate the skewness and range of the data series, and the effect of outliers on the statistical analysis of the data series.
What is a bimodal distribution?	A data series with a bimodal distribution has more than one frequency centre. Bimodal distributions do not respond to standard statistical tools in immediately predictable ways, so measures of distribution and hypothesis-testing should be used with care. Distributions with more than two frequency centres are called multimodal.
What is the range of a data series?	The range is the highest value in a data series minus the lower. It can be expressed as lower(difference)higher, e.g. 5(15)20. This arithmetic range should not be confused with the statistical range, which is a prediction of the range of a population based on the data series as a sample, and assuming the population has a normal distribution. It has limited predictive application.
What does the standard deviation measure about a data series?	It measures the amount of variation in a data series. The calculation is designed so that, when applied to a normally distributed data series, the number of standard deviations from

	the mean reliably indicate the amount of data included by the SD limits. The SD can also be used to measure the probability of a particular data item occurring. Items that are more than 3 SDs from the mean can be treated as data outliers.
When is the standard deviation a useful measure and when is it misleading?	The SD is useful when: • The sample has an approximately normal (bell curve) distribution. • The sample is not heavily skewed. • The sample has a single mode. In all other circumstances the SD can be misleading.
You are teaching the same lessons to two different groups. The previous teacher told you that group A was better than group B. How do you test this?	Use a t-test to compare the two data series. This will identify whether they are likely to be different samples from the same population (the apparent difference is not significant), or samples from different populations (the difference is significant).
What does the coefficient of correlation measure?	It measures the relationship between two (or more) data series: do changes in one series reliably indicate changes in the other? Not all correlations indicate a reliance of one series on the other (correlation does not imply causation), and some apparent correlations are actually coincidences (collocation does not imply correlation); so there has to be evidence beyond the statistical to indicate a true correlation.

Philosophy of Quantitative Methods

How large should a statistically significant subset of a data series be?	The usual claim is that a data series should contain 30 responses (about 75% confidence level at a 10% level of error). At a minimum it should contain 20 responses (about 65% confidence, 10% error). Market researchers quote a magic number of 400 responses (95% confidence, 5% error). However, if the series is to be used to identify subsets in a data set, the series should contain enough items to ensure about 30 items per subset. For example, a data series showing gender (M, F and U) is to be used to analyse a questionnaire question into subsets. Assuming the U classification is not analysed, the data series should have 60 items.
Give a case where qualitative methods are better, and a case where quantitative methods are better.	Qualitative methods work best when: • The data sought is necessarily idiosyncratic (the data cannot be usefully divorced from the data subject). • The data involves reactions which are difficult to quantify or difficult to anticipate (e.g. emotions or beliefs). • The data collection method is discursive. Quantitative methods work best when: • Large amounts of data are needed • The data can be represented numerically

	• The approach to the data is objective rather than subjective
List three advantages quantitative methods have over qualitative methods.	• Comparability: results from different studies can be compared, even where the studies seem quite disparate in terms of their data sets. • Testability: results can be tested in agreed ways, and the reliability of the data can itself be quantified. • Presentation: there are many agreed ways to present numerical data visually, and these visual representations can be much more informative than just numbers and words.
What is the most important thing to remember about your data subjects when conducting quantitative research? List some ways you can manage this.	That your data subjects are real people, and not just ciphers that produce numerical data. This should inform decisions: • At the design stage: questionnaires should be polite, accept and provide for null responses, and anticipate areas of offense. • At the analysis stage: without second-guessing your data subjects, try to understand why they have made the responses they have. • At the reporting stage: discuss the data as a product of human interaction and not just as numbers; and thank the data subjects!
What is the difference between: 1. "The average student achieved 58.27%"; 2. "The average student mark was 58.27%"; 3. "The average mark achieved by the students was 58.27%"? Which is preferable?	1. There is no such thing as "the average student!" Nobody has ever achieved a score of exactly 58.27%. 2. This is not factually incorrect, but it could be read as referencing the non-existent "average student". 3. This refers to the average mark, which is correct (you can average numbers but not people). So this is the preferred form.
What does "correlation does not imply causation" mean?	Correlation can be caused by several things: • Causation: the events represented by one of the data series instigates processes that produce the events represented by the other data series. • Mutual dependence: the events represented by both data series are caused by a third set of events. • Coincidence: the two data series are behaving similarly for no known reason. The reason why two data series are correlating cannot be discovered statistically; and causation is only one of the possible reasons why it is happening.

Questionnaires

Give two advantages and two disadvantages of using Likert scales.	**Advantages:** • Immediately quantifiable;

	• Does not require interpretation for comparisons between data subjects; • Objectifies the data. **Disadvantages:** • Limits the possible responses by the data subjects; • Can aggregate responses made for very different reasons.
Give an example of a Preferential List question.	**EXAMPLE:** Indicate your order of preference for the following root vegetables, rating the most preferred choice as 1 and the least preferred as 5. ___ Parsnip ___ Potato ___ Swede ___ Sweet potato ___ Turnip
What is the difference between a Graded Response question and an Arbitrary Numerical List question?	A graded response gives the data subject a single choice from a list of possible answers which form a continuous set (e.g. Tick the word that best describes your current emotional state: happy, content, discontent, sad). An arbitrary numerical list also gives a single choice, but the answers are discontinuous (e.g. select your favourite colour: blue, green, red, orange, brown, …)
How do Contingency questions affect the completion of a questionnaire?	Because they dictate the next question the data subject should answer, they mean that data subjects will complete the questionnaire differently.
How does a Guttman scale question work?	It allows the data subject to select all answers that apply. It allows for a nuanced analysis of responses, but it can be misinterpreted by the data subject.
What are some problems with questions that require a binary response?	• A binary response may create bias in the data. • They don't allow the data subject to express 'no interest' in the answer. • They don't allow the data subject to express uncertainy about the answer. • They don't allow the data subject to express disquiet with the question. • They can feel intrusive, and strain the social contract between the data subject and the researcher. • This is not a definitive list.
What is the key difference between even-numbered and odd-numbered Likert scales?	An odd-numbered likert scale allows for a neutral opinion to be expressed, and even-numbered scale means that the data subject has to decide on one side or the other. The appropriate type to use should be dictated by what you want to find out about the data you are collecting.
List an advantage and a disadvantage of a five-point Likert scale and an advantage	**Five point scale advantages:** • The range of answers is enough to differentiate grades of positive and negative;

and a disadvantage of an eight-point Likert scale	• The lowest of the "human numbers" (seven plus or minus two); • Not large enough to encourage neglect of the extremes. **Five point scale disadvantage:** • Odd numbered scale allows neutrality, which can indicate responses other than 'no preference'. **Eight point scale advantage:** • Even numbered scale prevent neutrality. **Eight point scale disadvantages:** • Just large enough to encourage neglect of the extremes; • Just large enough for bimodality to appear.

Graphs and Diagrams

When would you use a scatter chart?	When you wish to produce a representation of the correlative relationship between data series.
1. Give an example of data that is better represented by a column chart than a line chart. Why is this the case? 2. Give an example of data that is better represented by a line chart than a column chart. Why is this the case?	Data which is coded into non-continuous subsets (e.g. gender, hair colour) are better represented by column charts. Data which is coded into continuous numerical subsets (e.g. age, or a time series) are better represented by line charts. Some data (e.g. educational level) can be represented equally well by either chart. This is because a line signals continuity, while separate columns indicate discreteness.
What is different about a bubble chart?	It represents more than two data series. While remaining a two dimensional image. The size and colour of the dots can encode two further data series. A bubble chart provides a useful way to show a group of correlating data series.
What is the difference between a histogram and a column chart?	A column chart is for showing data which is subsetted into discrete or regular intervals: only the height of the column is significant, indicating the frequency in the subset. The histogram is for showing a data series with uneven intervals: the width of the column indicates the range of the data subset, while area indicates the frequency.
What is the difference between a column chart and a bar chart?	A column chart shows subsets horizontally, with the subset frequency indicated as a height above (or below) the horizontal baseline of the graph. A bar chart shows the subsets vertically, with the subset frequency indicated as a width distance from the vertical baseline of the graph.
What is a whisker diagram (or stock chart) used for?	It shows the range of one or more data series in graphical form. The box section is usually divided into quartiles, so that the range of the central half of the data is shown as a box, and the first and fourth quartile are shown as lines, or whiskers.

When would you use a radar chart?	To show discontinuous data, like employment profession. The frequency of each data subset is indicated by distance from the circle centre. Radar charts only really work where there are at least three subsets in a data series.
When is it appropriate to use a node diagram? Give an example.	Node diagrams show how different data items link together, so they are good at showing flows between data items. Classic node diagrams are: • City metro system maps. • Organisational charts. • Task Flowcharts. • Maps of interpersonal relationships in a group. • Electrical and electronic circuit diagrams. • Biological systems (e.g. blood flow in a body).

Glossary of Terms

Arbitrary numerical list question	A question which gives the data subject a single choice from a list of possible answers. The answers are discontinuous and, unlike likert scales, cannot be analysed as a numerical series.
Bar chart	This is similar to a column chart (see below), but the frequencies are displayed as horizontal bars instead of columns. Because they are much more popular, column charts are also often called bar charts, or vertical bar charts.
Bimodal distribution	A distribution which has more than two centres of frequency, or modes. If there are more than two it is a multimodal.
Binary response question	A question allowing only a positive or a negative response (e.g. Yes / No).
Bivariate correlation	A correlation between two data series. See **Correlation**, below.
Bubble chart	A scatter diagram (see below) where the size of the dots represents a third data series. The colour or intensity of the dots can represent a fourth data series. *For more, see Excel Help,* ***Bubble and scatter charts****.*
Column chart	Perhaps the most common way of showing data in a graph, it shows each subdivision of the data series as a vertical bar, with the height indicating the number in that subdivision (this number is also known as the frequency).
Contingency question	A question which dictates the path the data subject takes though a questionnaire. The data subject is directed to different next questions, depending on the answer given.
Correlation	A relationship between two (or more) data series. This relationship shows that changes in one series reliably indicate changes in the other. For example, as age rises from 20 to 60, savings also tend to rise; but also as age rises, energy levels tend to fall. Not all correlations indicate a reliance of one series on the other (correlation does not imply causation), and some apparent correlations are actually coincidences (collocation does not imply

	correlation); so there has to be evidence beyond the statistical to indicate a true correlation.
Data item	A single piece of data from a single data series. Also known as a datum.
Data series	A single type of data from a data set, e.g. a data series of ages.
Data set	All the data that is available to be analysed. It can be from a single study or a series of related studies.
Data subject	A person (or animal or thing) who provides some of the data for a data set.
Data subset	A subset of a data series, e.g. the subset of ages between 30 and 39.
Dependent variable	A variable which identifies the outcome of the statistical event being measured. Data questions are often dependent variables, giving the information which answers the research question.
Doughnut chart	A ring graph showing all the subdivisions of the data series as "slices" of a doughnut. The size of the slice represents the value of the subdivision relative to the value of the data series.
Frequency	The number of occurrences in a data series that are included in a particular subset of the series.
Graded response	A question which gives the data subject a single choice from a list of possible answers. The answers are usually (but not always) arranged in an order, which allows the data subject to decide where on a nonlinear scale they want to be.
Guttman scale question	This type of question allows the data subject to select all answers that apply. It allows for a more nuanced analysis of responses than the graded response, and a more accurate analysis of responses than a graded list. However, it is not always easy for a data subject to understand, and a null response cannot be distinguished from a "none of the above" response.
Histogram	This is a column chart for a data series with uneven intervals. The area of the column, not the height, indicates the value of the data subset. For example, if there are headcounts in three age ranges, 21-30, 31-60 and 61-70, then the 31-60 age group is represented on the histogram with a column three times the width of the other two, and 1/3 the height indicated by its value.
Independent variable	A variable which identifies a cause of the statistical event being measured. Demographic questions are one type of independent variable, they indicate features about the data subject which may affect their responses to the data questions.
Likert scale	A Likert scale is a common quantitative question form. The scale consists of a number of points representing a range of possible answers. The number of points can be anything from 3 to 100, although 11 is usually a maximum; and the range of answers can be indicated either by descriptors at each end, at both ends and the middle, or separate descriptors for each point. The likert

	scale allows analogue feelings and emotions to be expressed quantitatively. For more on likert scales, see **L2 - handout 1** on KEATS.
Mean	The arithmetic mid-point of a data series. By itself it tells us little, but when combined with other statistical measures it can tell us about the shape, size and range of the data series. The mean is the primary statistic in the calculation of most other statistics.
Median	The logical mid-point of a data series: if the series is sorted into order then the median is the middle-most value. Where there is an even number of items in a data series, the middle two items are added together and divided by 2. The median has a role, when compared to the mean, in determining skewness in a data series.
Mode	The frequency centre of a data series: the most frequent value in a data series is the mode. With highly skewed data series, the mode can be some distance from the mean, which gives a measure of skewness. Bimodal or multimodal distributions have more than one mode.
Node diagram	This is a graph where individual nodes are linked by connectors. The size, shape or colour of a node can signify the nature of the node, and the length, width and colour of connectors can signify the nature of the connections between nodes. Node diagrams are good at showing flows between data items. A classic node diagram is a city metro system map.
Normal distribution	Also known as the Gaussian distribution or bell curve, this is a theoretical distribution which has features that make it easy to analyse statistically. It is also useful because it approximates to many actual distributions in nature (e.g. human height). However, it is often used as an approximation where the actual distribution is predictably non-normal (e.g. human age). It should be used with care in these cases.
Outlier	A data item which does not fall inside the expected range of the data. A single outlier can distort a statistical analysis, especially for small samples, so you have to decide whether they should be included or excluded from the analysis. If you exclude them you should state which values have been excluded, and why.
p-value	A measure of the probable relation between the data set collected and the population from which it is drawn, expressed as a ratio. p-values cannot prove the relationship, they only indicate its likelihood. The main problem with p-values is not their calculation but their interpretation. See the booklet, the Excel spreadsheet and other stats sources for more on p-values.
Pie chart	A circular graph showing all the subdivisions of the data series as "slices", as if of a pie. The size of the slice represents the value of the subdivision relative to the value of the data series.

Population	The global group of people who could be productively surveyed for the particular research. Not to be confused with the general population, which is everyone.
Preferential list	A question where a list of items should be placed into an order. While it identifies the relative importance of the different items, it does not identify the range of the preference: does the order given reflect best to worst, or least worst to most worst?
Quartile	A data series can be divided into four sub-ranges: the highest-valued 25% of data items, the next-highest 25%, the third-highest 25%, and the lowest 25%. For instance, a data series of 8 items (1, 3, 4, 6, 7, 7, 8, 9) can be divided into: Minimum = 1; 1st quartile = 3; 2nd quartile = 6; 3rd quartile=7; 4th quartile =9. Other subdivisions of a data series include quintiles (5 divisions), deciles (10 divisions) and percentiles (100 divisions). However, quartiles are the most useful subdivision of most data series.
Radar chart	This is a circular plot for displaying discontinuous data, such as political affiliation. The frequency of a data subset is indicated by distance from the circle centre. Radar charts only really work where there are at least three subsets in a data series. For more on radar charts, see *Excel Help,* ***Available chart types: Radar charts****.*
Random sample	A sample of a population which does not predetermine who is sampled and where the sample is taken. The advantage is that it is theoretically likely to be representative of the population. The disadvantage is that, in practice, no sample can be truly random: social and geographical practicalities mean that there is always bias in a sample.
Range	The highest and lowest items in a data series, and the difference between them. A range is only really effective if it has all three values.
Sample	The section of the population that was surveyed, and for which data is available. See also **Random sample** and **Stratified sample**.
Scatter chart / diagram	A graph which plots the values of one data series (the x series) against another series (the y series) to show the level of correlation between them.
Significance	The level at which it is possible to say that one event or set of data is related to another: values above the significance level (or below, depending on the significance measure) are seen as supporting a hypothesis about the data; values below the significance level (or above it) are seen as contradicting the hypothesis. Significance levels cannot prove hypotheses, only indicate to what level of certainty they can be accepted.
Skew	The deviation of a data series from symmetry, in which mean, median and mode are all equal. The difference between the mode and the mean indicates the direction of skew, with positive values (mean - mode) indicating positive skew, and negative

	values indicating negative skew. The median will lie between the mean and the mode. SKEW is the Excel command to calculate the degree of skew in a data series.
Standard Deviation	A measure of variation in a data series. The calculation is designed so that, when applied to a normally distributed data series, the number of standard deviations from the mean reliably indicate the amount of data included by the SD limits. See **Types of distribution: Normal** above.
Stratified sample	A planned sampling of a population. The relevant structure of the population is identified (e.g. social classes, genders, ages, education levels, etc.) and the sample is planned to include numbers which reflect the population structure. The advantage is that it gives a better representation of the population. The disadvantage is that it can be difficult and costly: every subset of every subset should be sampled in the same ratio as the population (e.g. if class AB females aged 30-39 with a degree are 0.8% of the population, they should be 0.8% of the sample).
t-test	A measure of the likelihood that two data series (and, therefore, their data sets) come from the same or very similar populations. t-tests cannot prove the relationship, they only indicate its likelihood.
Trinary response question	A question with a positive and a negative choice, and one or more neutral responses. (e.g. Yes / No / Don't know / Prefer not to say)
Variable	An item of data which can between data subjects. See also **Dependent variable** and **Independent variable**.
Variance	A measure of distances from the mean of the data series. Variance emphasises larger distances and reduces smaller ones. Larger numbers indicate greater variation.
Whisker diagram	Also known as a **box plot** or a **stock chart**, this diagram shows the range of one or more data series in graphical form. The box section is usually divided into quartiles, so that the range of the central half of the data is shown as a box, and the first and fourth quartile are shown as lines, or whiskers. *See Excel Help **Create a box plot** for how to create a whisker diagram in Excel.*