Making the numbers work

An improvement tool to help you make the most of your data

Are you looking at numbers or data? Maybe you're thinking about what the numbers mean - whether they're good, bad or indifferent, how they compare to others, how you might analyse them - and how you might present what they're saying in the most effective way?

It can be hard to get your head around all of the possibilities - should you use the mean, mode or median? What type of chart will work best? And what on earth is polarity? Or perhaps you always stick to the same analysis and miss the opportunity of mining the data for more.

This improvement tool is designed to help you make the numbers work. It helps you through the different aspects using simple summaries, uncomplicated explanations and practical examples. You'll find it isn't that complicated after all. And maybe you'll spot some new ways to use numbers too.

The tool is set out in three sections - summarising data, comparing data and presenting and using data. There is also a helpful section about useful sources of data - and a glossary.

1. Summarising data

Summarising data can help you make the most of numbers by spotting and understanding patterns and pictures.

There are six ways you can summarise data

1. Using averages
2. Using frequencies and groups
3. Looking at data issues – polarity, precision an data types
4. Looking at distribution
5. Using box plots and histograms
6. Making use of sampling

Essential facts about summarising data

1. The average is an important reference point. How something compares to 'the average', is usually the first thing to think about.
2. Different ways of measuring the average end up with different results and you need to think about this before drawing conclusions.
3. The 'mean' average is the most often used type of average - but for some types of data the median, mode or frequencies might work better.
4. Knowing about the average, the dispersion and polarity of the data can help draw conclusions about how well an authority is doing.
5. The way a set of data is distributed can help you understand the make up of the data and where a specific value lies within the data.
6. Often you can draw a conclusion about a data value by finding out which quartile it lies in.
7. Box-plots and histograms are a useful way of quickly illuminating how a dataset is distributed. For example, if it has a normal distribution, or is skewed.
8. Using the standard deviation will allow you to look at the distribution of the data, - and also identify outliers and unusual values.
9. Sometimes there are too many data values to use - so you need to sample. If you do you will need to know about confidence intervals.

Summarising data

...using averages

Often when you first look at a data item, your initial point of comparison is where it is in relation to 'the average'. Is it better or worse than the average? Or is it about the same?

But what exactly does the average mean? There are different ways of measuring the average - and they usually produce different results.

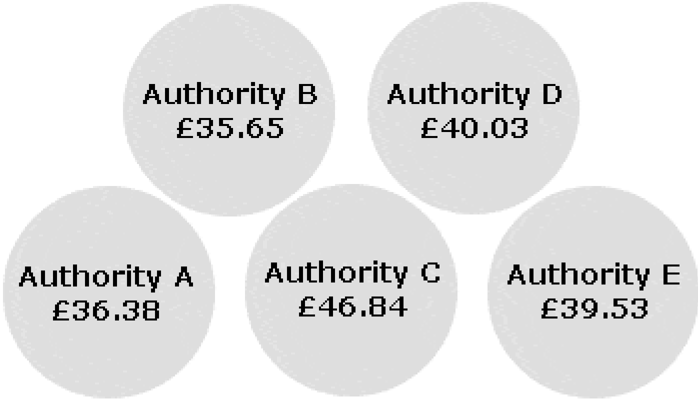
> **Example**
>
> You might read that the **average** salary in the UK is £25,000. Does this mean…
>
> 1. Most people in the UK earn £25,000?
> 2. If we knew what everyone in the UK earned and sorted them in order, the person in the middle would earn £25,000? Or,
> 3. If we took all the earnings in the UK and divided the total amount equally, everyone would get £25,000?
>
> Actually, you could use any of these definitions. But it is unlikely that they would produce the same figure.
>
> That's why it is important to know which definition of 'average' is being used before you can make any judgments on the data.

There are three key types of average you can use to help you make sense of numbers:

1. Mean,
2. Median; and
3. Mode.

Mean

The mean is the total sum of all the data when added together, divided by the number of cases. For example, you might use the mean to work out the average of BVPI 86 (the cost of waste collection per household) for five authorities:



The total sum of the data from the five authorities is £198.43. £198.43 divided by 5 is £39.69. Therefore the mean is £39.69.

You can tell that Authority C is quite a bit higher than the mean, Authorities A and B are lower than the mean. Authorities D and E have values which are similar to the mean - you could say they were 'about average'.

Median

To work out the **median**, the data must be sorted into order. The **case** in the middle is the median (or the 50th **percentile** as it is also known). If the data from above was sorted in order, it would look like this:

Authority B	Authority A	Authority E	Authority D	Authority C
£35.65	£36.38	£39.53	£40.03	£46.84

Authority E is in the middle, the median is therefore £39.53.

We can therefore tell that Authority E is closest to both the mean and is the median. In this set of data the mean (£39.69) and the median (£39.53) are quite similar. But this is not always the case. If you found that the **mean** and **median** were quite different, you may want to look at the dispersion or distribution of the data to see if you can find why this is.

Mode

The **mode** is the other type of average used. The mode is the value that occurs most often. Because much of the performance data from public services are comprised of fairly unique values, the mode is not used as much as the median or mean. For example, we could not work out the mode from the data from the five authorities above because each value only appears once. However for some other performance indicators we might use the mode. For example, in 2005/06 BVPI 199d (flytipping incidents and enforcement actions taken) requires authorities to be rated in **discrete** categories as either '1 = very effective', '2 = effective', '3 = good', or '4 = poor'. For this type of indicator, it would be useful to find out which category authorities most frequently occur in. The category with most cases would be the mode.

Summarising data

...using frequencies and grouped frequencies

This section looks at how to use frequencies and Grouped frequencies (Groups).

Frequencies

Sometimes when you are examining an item of data, it is useful to know how many times the same items appears in a whole set of data. For example, you might like to know how many authorities are county councils (there are 34 county councils) and how many are district councils (more than 200). **The number of times something occurs in a data set is known as a frequency.**

It is useful to know all the frequencies in a set of data, because then it is possible to determine if the specific value we are interested in is unusual or quite common.

Example

If you wanted to look at BVPI 199d (fly tipping incidents and enforcement actions taken) you might already know that the **mode** was 3 (or good), but how many authorities were rated in other categories?

You can find this by calculating the frequency of all the values:

BVPI 199d category	Frequency (number of authorities)	Frequency (percentage of authorities)
1 = very effective	30	7.7
2 = effective	127	32.6
3 = good	150	38.5
4 = poor	83	21.3

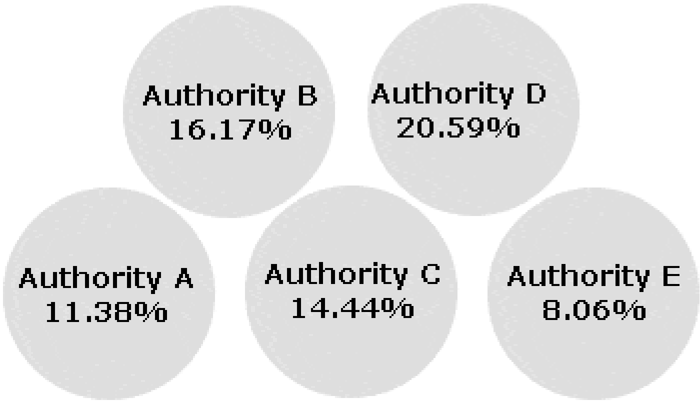
You can also calculate the percentage of authorities in each category. By working out the frequencies of all the categories, you can now tell that although 3 (or good) was the **mode** value (or most frequent) it only accounted for 38.5% of all the cases.

By looking at the frequencies, you are starting to consider the **distribution** of the data. This is very useful when making a conclusion about how well an authority is performing. For example if an authority was rated as 1 for BVPI 199d, you know from the frequencies that it is quite unusual because only 7.7% of authorities have been rated as 1 in the example above.

Grouped Frequencies

Calculating frequencies is easy when you have **discrete** data categories like in BVPI 199d, but what if you wanted to work out the frequency of different values from a set of values like that in BVPI 82a (the percentage of household waste recycled)?

You could produce different **groups** of data according to their values. For example, from the data below you could say that 1 authority recycled less that 10%, 3 recycled between 10% and 20% and 1 recycled more than 20%.



So a grouped frequency table might look like this:

BVPI 82a (% of household waste recycled)	Frequency
Less than 10%	1
10% to 20%	3
More than 20%	1

Obviously there will usually be many more that five pieces of data. Below are the 2003/04 results for BVPI 82a for 378 authorities, which have been grouped using the same grouped frequency classifications as above. You can tell from the frequencies that any authority recycling more than 20% waste in 2003/04 was one of only 5.6% of authorities and therefore quite unusual.

BVPI 82a (% of household waste recycled)	Frequency	% Frequency
Less than 10%	95	25.1
10% to 20%	262	69.3
More than 20%	21	5.6

Summarising data

...using data issues

Data can take many forms. You therefore need to know a few things about the data you are using before you can make a sensible judgment about how good or bad something is. For example you may like to know some of the following:

- Is a high score good, or is a low score good? - See polarity.
- Should the data be rounded to decimal points, or whole numbers? - See precision.
- Some statistical techniques cannot be used if the data is not of the right type, so which data type are you using?

Polarity

Polarity tells you if a low score is good or a high score is good.

For example, BVPI 9 is the percentage of council tax collected and a high score is therefore good.

BVPI 12 is the number of days lost to sickness absence, and a low score is therefore better.

The polarity of the Best Value performance indicators is displayed in the Audit Commission's 2005/06 Comprehensive Best Value Performance Indicator Guidance.

Precision

Often data gets rounded to the nearest whole number, or the nearest decimal place: this is the precision of the data.

Rounding can make a difference to performance depending on if the value goes up or down due to rounding.

An authority that scored 67.6% for one PI might have its score rounded up to 68% . But an authority that scored 67.4% might have its score rounded down to 67% .

Information on how many decimal places Best Value Performance Indicators should be rounded to is shown in in the Audit Commission's 2005/06 Comprehensive Best Value Performance Indicator Guidance.

Data types

Some data is non-numeric (or discrete). Other data is numeric (or continuous).

- **Discrete** data can be split into two different types:
  - **Nominal** data appears as well defined categories that have no order or hierarchy. For example, eye colour (blue, brown, green) is nominal data as there is no reason why blue eyes should come before or after brown.
  - **Ordinal** data does have an order to the categories. For example, survey responses might be categorised as 'very satisfied', 'fairly satisfied', 'neither satisfied nor dissatisfied', 'fairly dissatisfied' and 'very dissatisfied'. Ordinal data can therefore be placed in a hierarchy.
- **Continuous** data is expressed as actual numbers or integers. For example BVPI 12 is the average number of days lost to sickness.

Summarising data

...using distribution

When you look at a performance indicator, you often want to know how it compares with other results. You might start by looking how close the performance indicator was to the **average.** But this might not tell the whole story.

For example, you might want to know if the result you are interested in is close to the top of all the other results when you put them in order? Or is it close to the bottom of the **range** of data items? In other words, you want to know where your indicator **ranks** when compared to others?

Alternatively you could split all the indicators in the entire data set up, and see which **quartile** the result you are interested in falls into. Or you could use something like **standard deviation** to see if the result is unusual or not.

By looking at your specific indicator in comparison to all the other indicators, you are using the **distribution** of the data to make a judgment about performance.

To be able to determine more precisely where a data value sits in a large data set, we might like to look at the data in the following ways:

1. Range
2. Ranking
3. Percentiles
4. Quartiles
5. Quartiles and polarity
6. Normal distribution
7. Squewness
8. Standard deviation

Range

The easiest way to find out about the dispersion of a set of data is by looking at the minimum and maximum value. The range is then the difference between these. The table below shows the minimum, maximum and mean values for BVPIs 8 and 9 in 2003/04.

	BVPI 8 - percentage of invoices paid by the authority within 30 days	**BVPI 9 - percentage of council tax collected**
Maximum value	99.76	99.89
Mean value	90.19	96.85
Minimum value	31.72	79.30
Range	68.04	20.59

Both PIs have a similar maximum value - close to 100%. However the minimum values differ quite a lot. BVPI 8 has a minimum value of 31.72 whereas 9 has a minimum value of 79.30. By subtracting the minimum value from the maximum value we can calculate the **range** of the data. We can tell that BVPI 8 has a much larger range than BVPI 9. In other words, the performance of authorities for BVPI 8 seems to be quite spread out, but for BVPI 9 authorities are much closer together.

Sounds simple - but the problem is that the minimum and maximum values are sometimes not very representative of most of the other data values. They might be extreme. For example in the data on percentage of invoices paid in 30 days, one authority might have performed extremely well to produce the maximum figure and another authority might have had some kind of problem which produced the minimum figure. But most of the other authorities would be somewhere in the middle and not very close to these two.

In this particular example, the mean is much closer to the maximum value than the minimum value. This implies that there could be a possible extreme minimum value, for example one authority may have performed particularly poorly. However this is only a guide, and further analysis of the dispersion would be required.

To be able to determine more precisely where a data value sits in a large data set, you might like to look at the data by ranking, percentiles or quartiles.

Ranking

An easy way of showing where a data value is within the distribution is to **rank** the data. **The data values are sorted, and a number is assigned to each value corresponding to where it appears in the order.** For example, below is a table of 10 local areas, which have been ranked according to population density (number of people per square kilometre):

Population Density	Rank (where 1 has the lowest population density)
71.0	1
153.1	2
158.2	3
159.1	4
204.8	5
241.2	6
902	7
2359.6	8
2451.1	9
2702.4	10

Ranking is an easy and effective way of determining where a data value occurs in a data set. For example, you might hear that somewhere was 'the second most popular tourist attraction in the UK'.

But ranking does have certain drawbacks.

Drawbacks of ranking

- **Ranking doesn't tell you much about the scale of the differences within the data values.** In the table on population densities the first rank is a value of 71. The second is a value of 153 - that is more than twice the density of the first rank. The third rank (158) is only 5 bigger than the second. But the ranks of these three different values just appear as 1, 2 and 3.
- Also **people viewing the information need to know how many values are being considered in the ranking.** A local authority might be ranked as 20th best for a certain PI, but out of how many authorities? It could be all authorities of its type (which could be quite small) or it could be all authorities in the country (which could be quite large).

Percentiles

With **ranking** it is important to always know how many values have been included in the ranking. But if **percentiles** are used, then this becomes less of an issue. **Percentiles show where a data value would lie in the data range in terms of a percentage of all the sorted values** - i.e. the maximum value is the 100th percentile, the value in the middle (the median) is the 50th percentile etc.

Example

Department for Communities and Local Government (formerly ODPM)'s Indices of Deprivation

The 2004 Index of Multiple Deprivation provided data at relatively small geographical units called Super Output Areas (SOA).

Each SOA was given a deprivation score and a rank.

A rank of 1 meant it was the most deprived SOA in the country and a rank of 32,482 was the least deprived SOA.

When there are over 32,000 data values, it can be difficult to comprehend the magnitude of differences between cases. For example, if you said that an area was 'the 2,821st most deprived SOA in the country' it may not sound like it is very deprived. If, on the other hand, you said the SOA was 'one of the 10% most deprived areas in the country', the message is clear.

The table below shows some SOAs and their Index of Multiple Deprivation scores. The rank is provided for each score and the percentile of the rank. Because there are so many cases in the Index of Multiple Deprivation, the percentiles have been rounded to one decimal place so that differences within percentiles can be seen.

IMD SCORE	**RANK OF IMD (where 1 is most deprived)**	Percentile
79.38	31	0.1
62.68	717	2.2
47.17	2821	8.7
28.24	8939	27.5
27.84	9117	28.1
26.32	9897	30.5
19.85	13974	43.0
16.03	17144	52.8
10.63	23130	71.2
7.01	27711	85.3
2.87	31844	98.0

When there are over 32,000 data values, it can be difficult to comprehend the magnitude of differences between cases. For example, if you said that an area was 'the 2,821 st most deprived SOA in the country' it may not sound like it is very deprived. If, on the other hand, you said the SOA was 'one of the 10% most deprived areas in the country', the message is conveyed.

Quartiles

A common way to summarise Best Value data is using quartiles. Many authorities aim to be in the 'best quartile'. But what does this actually mean?

Below is the final Premier League Table for the 2004/05 season. Normally the table is only divided up to show which three teams are in the relegation zone and which could qualify for European football next season. But here it is divided into quartiles. The Premiership has 20 teams, so we can split the teams into 4 quarters of 5 teams each according to how many points they scored. The teams with the pink background were the highest teams - **the upper quartile**. The teams with the green background finished with fewest points - they are in **the bottom quartile**.

Final league position	**Team**	**Total points**		
1	Chelsea	95		**<Maximum value**
2	Arsenal	83	**<Upper quartile teams**	
3	Man Utd	77		
4	Everton	61		
5	Liverpool	58		**<75th percentile**
6	Bolton	58		
7	Middlesbrough	55	**<2nd quartile teams**	
8	Man City	52		
9	Tottenham	52		▲
10	Aston Villa	47		The inter-quartile range
11	Charlton	46		▼
12	Birmingham	45	**<3rd quartile teams**	
13	Fulham	44		
14	Newcastle	44		
15	Blackburn	42		**<25th percentile**
16	Portsmouth	39		
17	West Brom	34	**<Bottom quartile teams**	
18	Crystal Palace	33		
19	Norwich	33		
20	Southampton	32		**<Minimum value**

So the **maximum** value of points scored in the 04/05 Premiership was 95 and the **minimum** was 32. (95-32 = a **range** of 63).

You can also work out the value of each quartile:

- Because there is not one team occupying the 25th percentile, its value will be half way between the two values either side of the quartile ((39+42)/2 = 40.5)
- Because the two data values either side of the 75th percentile are the same, the value of the 75th percentile will be the same (58).

The difference between the 75th and 25th percentile is called the **inter-quartile range** (58-40.5 = 17.5). It is useful to know what the inter-quartile range is as it tells us the range of scores that the middle half of the teams achieved. The **inter-quartile** range might be used instead of the **range**, because the range might be distorted by the minimum and/or maximum values being very different from most of the rest of the data (i.e. they are 'extreme' values).

One thing to remember when using quartiles is that the **quartile thresholds change over time**. If the majority of authorities improve their performance year on year, then the quartile thresholds will also improve. This means that an authority will have to improve its performance just to stay in the same quartile.

Using a league table split into quartiles, similar to the football league example, is a useful way of seeing how well something is doing, but it does have its limitations.

- It doesn't really show graphically the extent of any variation within the quartiles. For example Chelsea finished with 95 points, 37 points ahead of Liverpool. That's more than half the range of all the points scored in the whole league (which was 63), but the table above only shows Chelsea to be four places above Liverpool.
- Presenting the data in a table is good when you only have 20 teams; but a table of nearly 400 local authorities would be too large and time consuming.

However, you could get around these issues by presenting the data graphically in a **box-plot**.

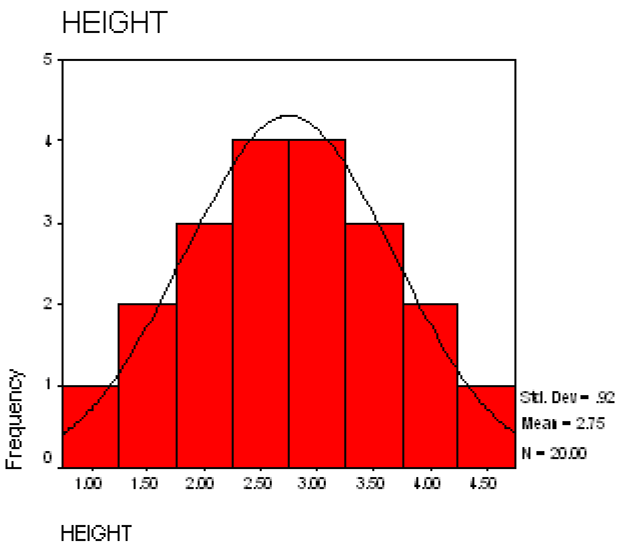
Quartiles and Polarity

Usually when data which does not have a **polarity** is summarised as quartiles, the *first* quartile is the one with the lowest results (i.e. the 1 st to the 25th percentile), the *second* quartile is the next highest (i.e. 26th to 50th percentiles).

When a polarity is included then the terms 'best' and 'worst' are used, regardless of the direction of the polarity. The '2nd'quartile is always the 2nd best quartile.

Normal distribution

if you took a random group of people and plotted their heights, you might expect to get a small number of short people, a small number of tall people, and a larger number of people of average height. The histogram you plot might have the same shape as the one below.



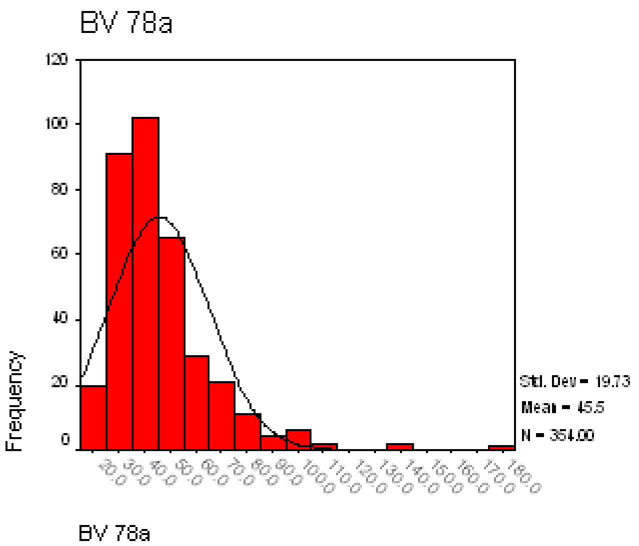
The red bars show the number of people in each height range. The shape is very symmetrical - it is high in the middle with both sides sloping away with the same shape. The mean, median and mode of this data will all be very similar values and will be in the middle of this distribution. A black curve has been plotted over the chart to provide a clearer understanding of the shape. This curve has a bell shape called a **normal distribution**.

Skewness

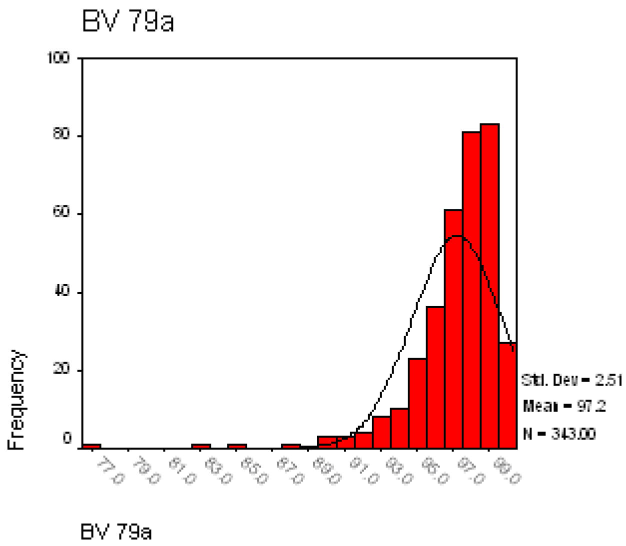
Data does not always have a normal distribution - often it is skewed. **Skewness shows you if the bulk of the values in a set of data are concentrated at one end of the data range or the other.** The more skewed a distribution is, the greater the difference between the mean, median and mode.

The following two **histograms** show the 2003/04 results for BVPIs 78a (the average time taken for processing new benefits claims) and 79a (the accuracy of processing benefit claims). The majority of the cases for 78a appear towards the left - the distribution is **skewed to the left**. The majority of cases for 79a are **skewed to the right**.

BVPI 78a average time for processing new benefits claims:



BVPI 79 accuracy of processing benefits claims:



The two histograms show that most authorities tend to be quick at processing new claims (skewed to the left), and also accurate (skewed to the right).

Implications of working with skewed distributions

- A skewed distribution is concentrated at one end of the scale.
- This means that if you put the data into quartiles, one quartile is likely to have a much larger range than the others.
- It also means that an absolute change in an authority's performance of just a few percentage points, could result in a very large relative change.
- On the other hand, quite a large absolute change could result in a comparatively small relative change; depending on which way the data is skewed.

Standard Deviation

Simply put, **the standard deviation is the expected deviation around the mean.** A large standard deviation indicates that the data are far from the mean, and a small standard deviation indicates that they are clustered closely around the mean.

Standard deviation can be used to calculate how unusual an individual data item is. You can work out how many standard deviations the individual result is away from the mean. The more standard deviations the individual score is away from the mean, the more unusual it is.

In a **normal distribution** you would expect 68% of all the result to be within 1 standard deviation from the mean. You would expect 95.4% of all the results to be within 2 standard deviations; and 99.7% would be within 3 standard deviations. Therefore if you found a result to be 3 standard deviations away from the mean, it would be a very unusual result.

Example of calculating standard deviation

The following set of data could be BVPI scores from 5 authorities:

18		**20**		
	21			**22**
		24		

The mean of these five values is 21. So one thing you could do is work out how far each value is from the mean. 18 is 3 from 21 (21-18=3), and 24 is also 3 from the mean. 20 and 22 are 1 from the mean, and 21 is 0 from the mean.

So the distance from the mean for each value is: 3,1,0,1,3, and the sum of these values is 8. If you divide this by 5 (the number of cases) to calculate the mean then **the mean deviation from the mean is 1.6.**

It can be argued that those values which lie furthest from the mean should have a disproportionately large influence on the measure of how dispersed the data is. One way to do this is by squaring the values (or multiplying them by themselves). So 3x3=9, 1x1=1 and 0x0=0. Your squared deviation values are therefore: 9,1,0,1,9, and the sum of these values is 20. If you divide this by 5 to calculate the mean then **the mean squared deviation (also known as the variance) from the mean is 4.**

Using the variance could lead to some very large numbers; so often the square root of the variance is used. This is the standard deviation.**The standard deviation is the square root of the deviation.** In our example it is 2.

Below are two more distributions and their standard deviations. You will see that the spread of distribution B is larger, and the standard deviation is much higher.

	Example A	Example B
	24 26 27 34 37	10 24 46 63 78
Mean	29.6	44.2
Standard Deviation	**5.6**	**27.7**

As well as being a measure of dispersion, standard deviation is also useful because it allows us to work out values that are **outliers.** The section on **box-plots** introduces the concept of outliers. If a value lies **more than 2 standard deviations from the mean** then it is described as an outlier or unusual value.

Summarising data

...using box plots and histograms

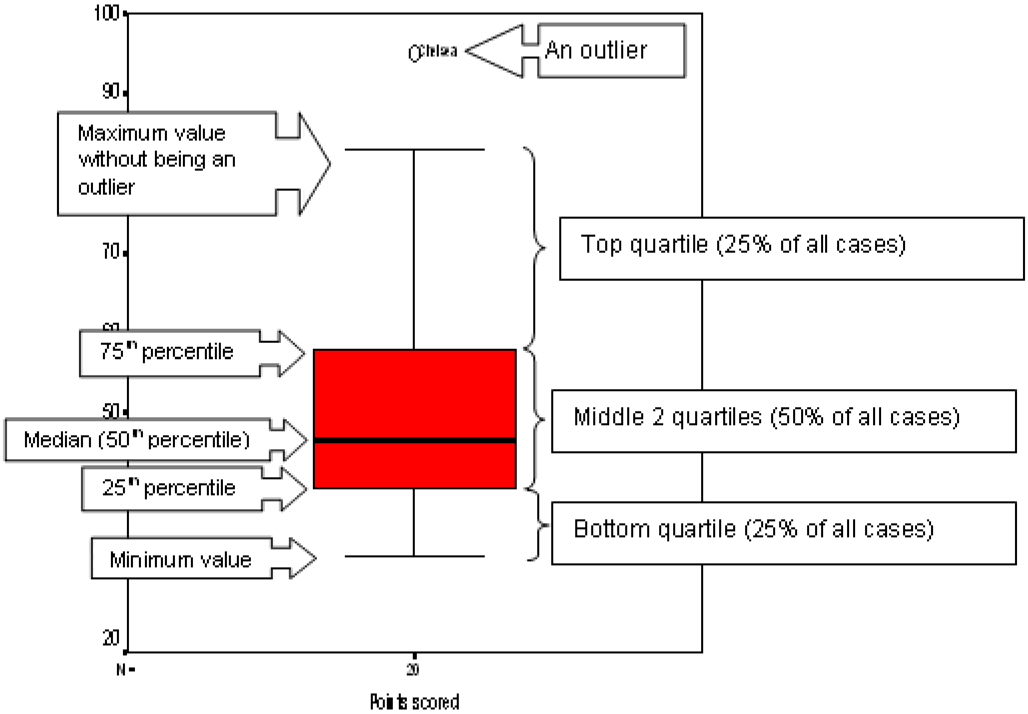
Before you draw a conclusion about a performance indicator, you often want to look at how it compares with other results in the data set. You can calculate **quartiles**, and you can consider the shape or **distribution** of all the data. But often it is easiest to use charts to represent this.

A chart can show all the results in a complete data set. You can then see where the specific result you are interested in fits into the chart, to determine where it lies in the distribution.

When looking at the **distribution** of a dataset, you might use box-plots and histograms. You might use other charts if you wanted to compare your results **over time** or **with other variables**.

Box-plots

A box-plot provides a visual presentation of the quartiles of the data. They are a good way of visualising how a dataset is distributed and identifying outliers. Below is a box plot of the final points scored in the 2004/05 Premiership season.



What does the box plot show?

The red box shows the **middle two quartiles**, the black line in the middle separating them is the median value (or the 50th percentile).

Coming out of the top and bottom of the box are two fences (or whiskers as they are also known). The fences show the **range of the upper quartile on the top and the bottom quartile on the bottom**.

The **size of the range** of each quartile varies, but the number of cases in each quartile remains the same.

You can tell from the box plot above that the bottom quartile and the second bottom quartile are quite small - the range of these quartiles is therefore quite small. This shows that there was not much variation in the performance of the teams in the bottom two quartiles.

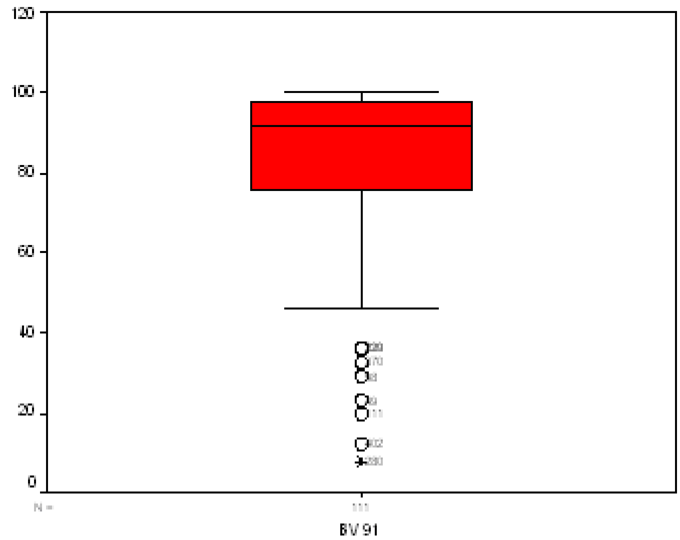
However the size of the range of the top quartile is quite large - this means that there was quite a big difference in the number of points scored for those teams in the top quartile.

One team (Chelsea) falls outside the upper fence. This is known as an **outlier**.

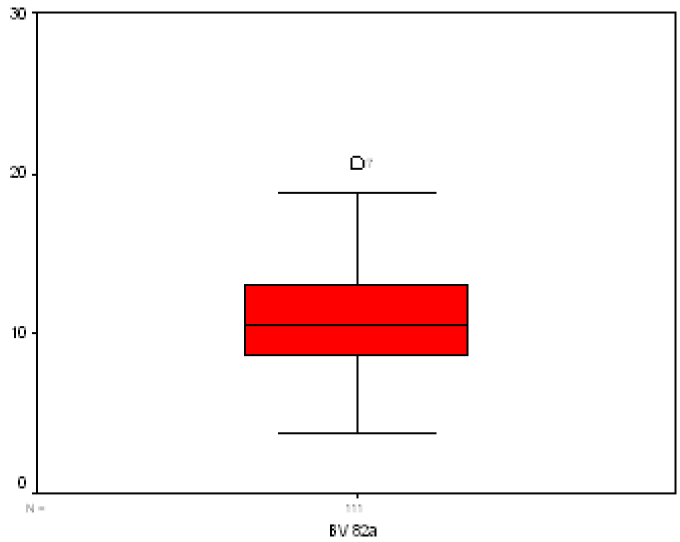
Normally about 95% to 96% of all the data items would be within the fences of the box-plot. Or in other words 95-96% of data falls within 2 **standard deviations** from the mean. An outlier is therefore more than +/- 2 standard deviations from the mean and may warrant further investigation to determine why the performance is unusual.

The two following box-plots show 2003/04 data for Best Value Performance Indicators 82a (the percentage of household waste recycled) and 91 (the percentage of residents served by a kerbside collection of recyclables).

BVPI 91 Residents served by kerbside recycling:



BVPI 82a household waste recycled:



What do the box plots show?

There are different size scales down the left of the plots. The median (the black bar in the red box) is about 10 for 82a (household waste recycled), whereas it is about 90 for BVPI 91 (residents served by kerbside recycling).

The fences for BVPI 91 run from 100 down to about 40, whereas for 82a they go from about 4 to about 20.

So the first thing that could be concluded from the box-plots is that BVPI 91 (residents served by kerbside recycling) has a **larger range** than BVPI 82a (household waste recycled).

The next thing you might notice about the two plots is their **shape**.

BVPI 82a has a fairly even shape: The two red rectangles in the middle of the plot are roughly the same shape, and so are the two fences at the top.

The shape of BVPI 91?s box-plot is very different. It is squashed towards the top. The top fence (the upper quartile) is very small, while the bottom fence (the lower quartile) is large. The same applies to the middle two quartiles: the upper middle quartile is smaller and more squashed than the lower middle quartile.

You know that 25% of all the values are in each quartile, so you can conclude that the values in the upper two quartiles for BVPI 91 are more concentrated or closer together than they are in the bottom two quartiles. You could say that the distribution of the data for BVPI 91 is **skewed**.

The box-plots also identify **outliers** from the distribution.

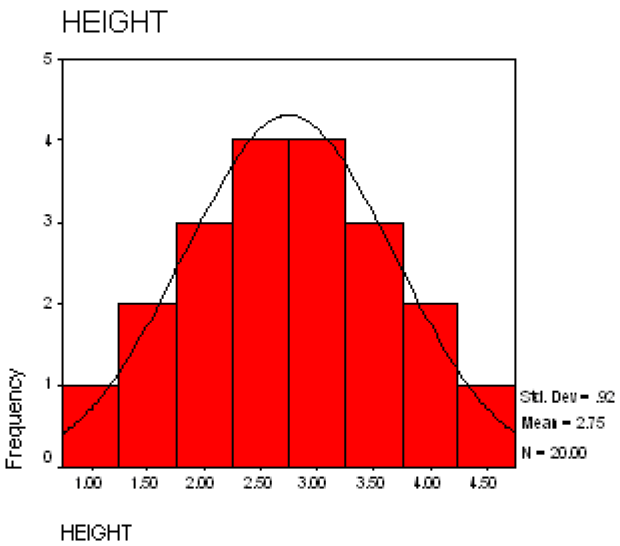
BVPI 82a shows one outlier above the upper fence. This is an authority where the performance is unusually high.

BVPI 91 shows several outliers below the bottom fence. These are authorities where the level of performance is unusually low compared to the rest of the distribution. The box plot

Histograms

Histograms are a good way of displaying the distribution of a set of data. Histograms show the size of the data value being considered at the bottom and the frequency of the cases by the side.

Histograms are useful because they demonstrate the shape of the data distribution. For example, if you took a random group of people and plotted their heights, you might expect to get a small number of short people, a small number of tall people, and a larger number of people of average height. The histogram you plot might have the same shape as the one below.



The red bars show the number of people in each height range. The shape is very symmetrical - it is high in the middle with both sides sloping away with the same shape. The mean, median and mode of this data will all be very similar values and will be in the middle of this distribution. A black curve has been plotted over the chart to provide a clearer understanding of the shape. This curve has a bell shape called a **normal distribution**.

Summarising data

...using sampling

It is not always possible to measure or determine the value of every possible case - i.e. the entire **population**. For example, consulting an entire population is unrealistically expensive; therefore nearly all surveys consult a **sample** of the population. The opinions and results given by the sample are then used to generalise or infer the opinions of the population.

There are a variety of sampling procedures, which fall into two groups - **probability samples** and **non-probability samples**.

Probability and non-probability samples

- A **probability sample** is one in which every individual within the survey population has a known chance of being selected for the sample.
- **Non-probability samples** are based in part, on the judgement of the researcher or interviewer, and are therefore open to bias.

To get **statistically reliable results**, a form of random probability sampling must be used. This is because random sampling limits selection bias (i.e. some individuals are more likely to be consulted than others) and allows the calculation of **confidence intervals**.

What is a confidence interval?

The confidence interval shows how accurate the results are. The smaller the confidence interval, the more accurate the result is. Every survey should give some kind of indication as to how accurate it is, or what its 'margin of error' is. This shows how likely the same result is to be found if the survey was repeated under the same conditions.

Example 1 - Confidence intervals

Authority A asked residents, 'how satisfied are you with the area as a place to live?' in a survey.

89.6% of the respondents said that they were satisfied or very satisfied.

Performance Specialists inspecting the council were pleased that the proportion was so high, but wanted to know how accurately the result reflected the entire population.

The result had a **confidence interval of plus or minus 1.7%** at the **95%** confidence level. The 95% confidence level is generally used in social research. It means that if the survey was conducted 100 times (under the same conditions and at the same time), we would expect the result to appear within the confidence interval margins 95 times - or 19 times out of 20

So for this example we could expect the result to be 89.6% plus or minus 1.7% 19 times out of 20; or in other words somewhere between 87.9% and 91.3%.

Example 2 - Confidence intervals

Two districts wanted to test the level of satisfaction with street cleaning.

District A found that 60% of residents were satisfied with a **confidence interval of plus or minus 5%**.

District B found that 67% of residents were satisfied, but the **confidence interval was 10%**.

So District B got a higher score, but District A's survey was more accurate.

District A's 'true' percentage is somewhere between 55% and 65%. District B's 'true' percentage is somewhere between 57% and 77%.

Because the confidence intervals overlap, **we cannot be certain that there is a difference between the two authorities.** We would need to use a statistical test (for example a t-test) to see if there was a difference.

2. Comparing data

Once you've got a hang of summarising the data you need to think about using the lessons in a practical way to compare data, make judgements - and make the most of the numbers.

There are five key ways to compare your data:

1. Using averages
2. Using quartiles
3. Using benchmarking and ranking
4. Looking over time
5. Using other variables

Essential facts about comparing data

1. Using averages is the easiest way to compare.
2. If you want a more detailed comparison using quartiles might be the answer.
3. Comparing against a group of authorities with similar characteristics can be useful - this is benchmarking.
4. Looking over time or direction of travel is also crucial. But whilst an authority maybe improving absolutely - it may not be improving relative to others.
5. Sometime variables are related. Plotting variables on a scatter plot and calculating the correlation coefficient are useful ways to check whether there is a relationship.
6. But a strong correlation only shows that there is a relationship - it doesn't explain it.

Comparing data

...using averages

The most obvious and easiest way of comparing how well an authority is performing is to compare to the **average.**

The example below is taken from the CPA Information Packs. The indicator is the percentage of local residents claiming unemployment benefit (an indicator where a low score is best).

This authority had an unemployment claimant rate of 3.7% in 2003, 3.3% i in 2004 and 3.2% i in 2005. The decrease in unemployment shows encouraging improvement. But when the rate of the local authority is compared to the average rate in the region and in England, it is clear that this authority still has quite a high unemployment rate.

Unemployment (NOMIS)								
			This LA	**Increasing ▲ Decreasing ▼**		**Regional Average**		**England Average**
Claimant Count Rate	Mar 2003		3.7			2.2		2.2
	Mar 2004		3.3	▼		2.0		2.1
	Mar 2005		3.2	▼		2.1		2.0

But you do need remember that whilst comparing an authority against the average (usually the **mean**) is useful, it only tells us if the authority is better or worse than the average. What it doesn't show is how much better or worse. One way of find out more about the authority's performance is to find out which **quartile** the authority is in.

The table also lets us compare the local authority's rate of improvement with regional and national averages or rates. In this example although the authority has a higher rate than the region or England; it is reducing faster than both as well.

Comparing data

...using quartiles

If comparing against the average isn't giving you the detail you need you might want to use **quartiles**.

If an authority's score is in the top 25% when compared to other authorities, it is in the **upper quartile**. If it is in the bottom 25% , it is known as **lower quartile.** Sometimes the polarity of data is low (for example the number of days lost to sickness absence) so to avoid confusion the terms upper and lower, best and worst are often used instead.

Some example tables from CPA Information Packs are shown below:

> ### Want help with reading the tables?
>
> - The 'this LA' column shows the score that the individual authority achieved for each year
> - The 'improving' column shows if the authority's performance has improved or deteriorated in terms of this result - arrows pointing up indicate improvement, downward arrows indicate deterioration
> - The 'quartile Position' column shows which quartile the authority was in for that particular year
> - The 'best, median, worst' columns show the thresholds, which determine which quartile the authority will be in. By using these columns you can work out how close the authority was to being in a different quartile.

In the example for BVPI 184a, the authority had 73% of local authority homes that were non-decent in 2002/03. In 2003/04 the authority improved to 65% . This improvement is encouraging, but we can see that the authority is ranked in the worst quartile for both years. Although the authority improved in terms of percentage in 2003/04, it needed to achieve less than 53% to not be considered worst quartile - with a score of 65% it did not succeed!

However in 2004/05 it has continued to improve, achieving 47% and reaching the 3rd quartile; even though the threshold for the worst quartile also decreased from 53% to 48%.

BVPI 184a percentage of LA homes that are non decent										
Comparator Group	**Year**		**This LA**	**Improving ▲**		**Quartile Position**		**Best**	**Median**	**Worst**
Single Tier and District Councils	2002/03		73			Worst Quartile		25	40	54
	2003/04		65	▲		Worst Quartile		21	36	53
	2004/05		47	▲		3rd Quartile		19	32	48

BVPI 62 Percentage of private sector housing made fit or demolished										
Comparator Group	**Year**		**This LA**	**Improving ▲**		**Quartile Position**		**Best**	**Median**	**Worst**
Single Tier and District Councils	2002/03		4.50			Best Quartile		4.46	2.75	1.55
	2003/04		4.67	▲		Best Quartile		4.32	2.66	1.44
	2004/05		4.54	▼		Best Quartile		4.50	2.70	1.56

In the example for BVPI 62 (an indicator where high performance is good performance), the authority is ranked in the best quartile each year, even though the authority's performance deteriorated during 2004/05.

Comparing data

...using benchmarking and ranking

It can be useful to compare or **benchmark** an authority against other authorities with **similar characteristics**. You could, for example, benchmark all Metropolitan boroughs or district councils against each other.

A way of determining how well an authority is performing in relation to other authorities is to calculate where it is ranked in terms of performance for a certain indicator. For example, you may say that the X London Borough is ranked 19th of all the 32 London Boroughs for BVPI 12. Other useful benchmarking groups are the CIPFA groupings or ONS's nearest neighbour comparisons.

Comparing data

...over time

It is always interesting to know if things are getting better or worse over time. This **Direction of Travel** is an important requirement of the new CPA methodology.

Determining if something has improved looks a relatively straightforward task. For example, if an authority achieved 60% for BVPI 157 (electronically enabled interactions with citizens) in 2003/04 and scored 65% in 2004/05, then you would say that it had improved.

But what if all the other authorities had improved as well? And what if they had all improved more? You might find that the authority was ranked 109th of all authorities when it achieved 60% in 2003/04. But because all the other authorities improved by a greater amount, it dropped to 125th in 2004/05 even though it improved.

We could say that this authority had shown an **absolute improvement** because its score increased between 2003/04 and 2004/05. But it had shown a **relative deterioration** because its position in **relation** to other authorities had dropped.

We can consider the absolute and relative improvement or deterioration of an authority's performance by plotting its performance and the quartiles on a chart.

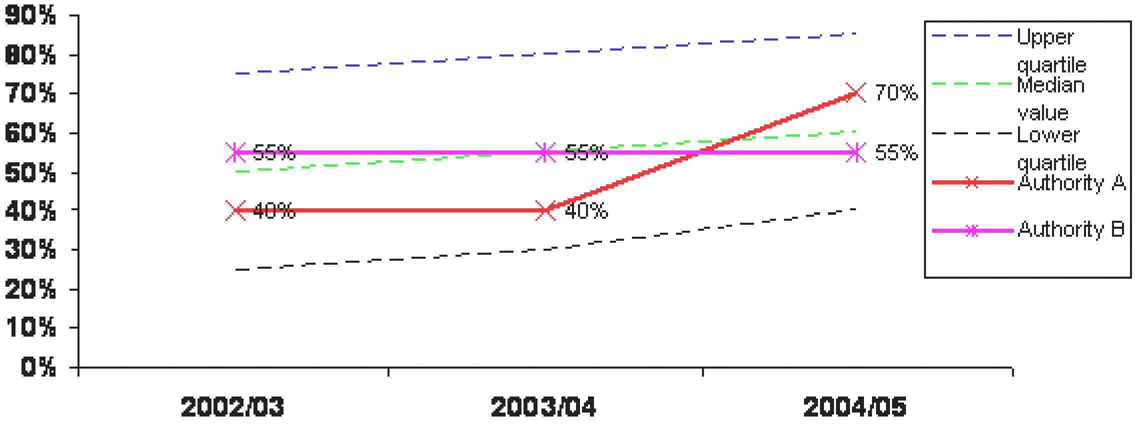
The chart below shows the performance for two authorities.

Example

- **Authority A** achieved 40% in 2002/03 and 2003/04; for these years it was ranked below the median but above the lower quartile, it therefore appeared in the third quartile

  In 2004/05 it achieved 70% (which is an absolute improvement). In 2004/05 its improvement was great enough to move higher than the median. It is now ranked in the second quartile and therefore shows a relative improvement
- **Authority B** achieved 55% in each of the three years. Although it was ranked in the second quartile for the first two years, the improvement in other authorities means that the median value also rose

  Authority B's performance in 2004/05 is now lower than the median which means it now appears in the third quartile. Authority B has therefore shown a relative decline in performance even though its absolute performance has remained the same



The issue of absolute and relative performance arose when the results of the **2000/01 and 2003/04 Best Value General surveys** were compared. When the mean results for authorities were compared from the two sets of surveys, they showed a drop in satisfaction with the overall way that the authority runs things. Therefore those authorities who achieved the same level of satisfaction in 2000/01 and 2003/04 had done quite well as they had shown a relative improvement. Those authorities that also managed to show an absolute improvement had done very well.

Comparing data

...using other variables

You might want to compare one set of data with another set to see if there is a relationship between two **variables.** Scatter plots and correlation are important here.

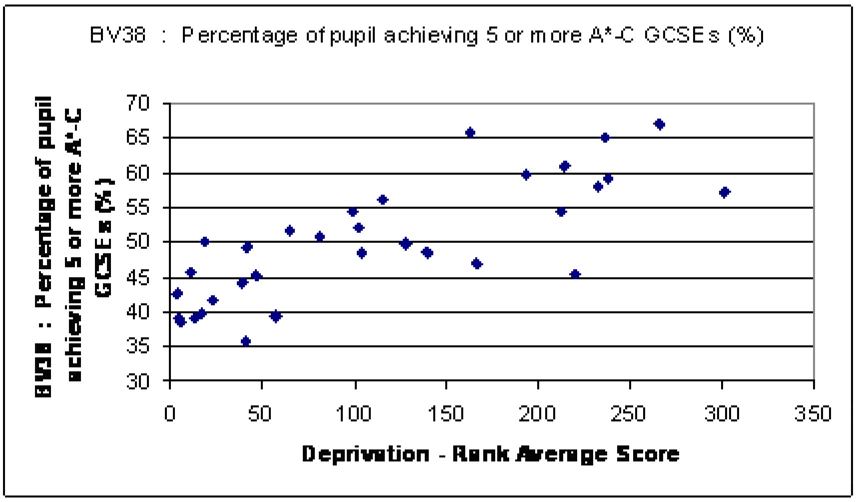
Scatter plots

In a scatter plot the two variables are plotted on the chart - with one variable using the x-axis, and the other the y-axis.

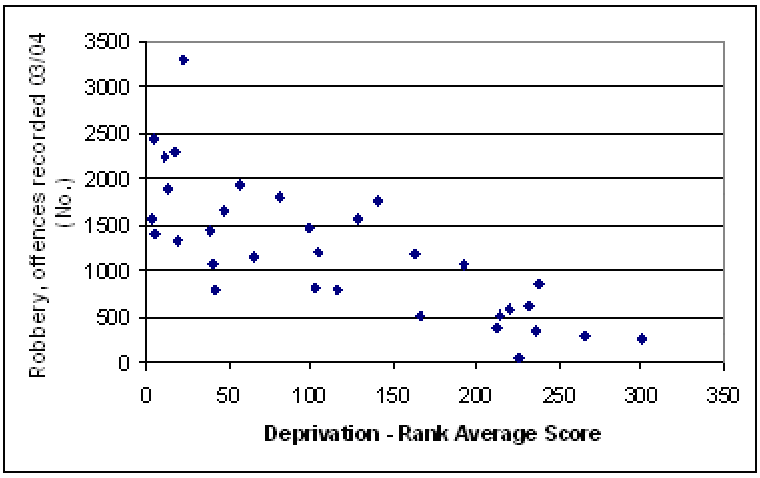
The scatter plot below shows the results of BVPI 38 (percentage of pupils achieving 5 or more A*- C GCSEs) up the side (**the y axis**) and deprivation rank along the bottom (**the x axis**).

The plot shows that there is a relationship between these two variables because the plots vary around a line. You could draw a line through the middle of the plots on the chart - this is called the **line of best fit**.

The chart shows that those authorities that have a high rank (the higher the rank the less deprived the area) also have a high score for BVPI 38. Authorities with a low score for deprivation are also likely to have a low score for BVPI 38. This type of relationship, where one variable increases if the other variable increases, is known as a **positive relationship.**

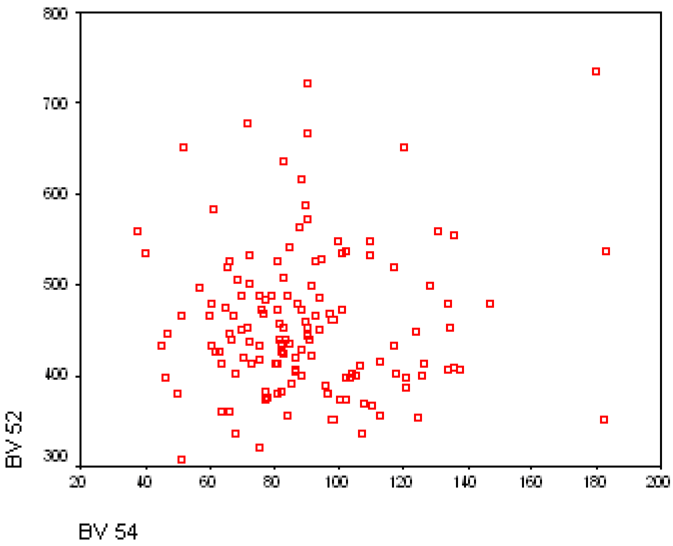


This second scatter plot shows the number of robbery offences recorded against deprivation rank.



This time the line slopes down, suggesting that authorities with high levels of deprivation tend to have more robbery offences recorded (remembering that the higher the deprivation ranking the less deprived the area). This type of relationship, where one variable decreases as the other increases, is known as a **negative relationship.**

This final scatter plot shows two more variables. This time BVPI 52 (the cost of services for intensive social care for adults) has been plotted against BVPI 54 (the percentage of over 65s helped to live at home). This time there is no real line on the chart. It would seem that there is **no real relationship between the two variables**.



Correlation

If you think there is a relationship between two variables, you may like to look at the ***degree of association between the two variables***. To do this you can use correlation.

Various software packages (See presenting and using data) can easily calculate the degree of correlation between two variables. ***The measure of correlation is known as the correlation coefficient.***

The correlation coefficient

- The correlation coefficient will be a number somewhere between 1 and -1.
- A coefficient of +1 shows a **perfect positive correlation** where one variable increases exactly as the other variable increases.
- A coefficient of -1 shows a **perfect negative correlation** where one variable decreases exactly as the other increases.
- A score that is **close to 0 shows no correlation** or association between the two variables.
- The nearer the coefficient is to +1 or -1, the stronger the degree of association is. As a general rule of thumb, correlations with a coefficient of more than +0.7 or less than -0.7 are said to be **strong.**

You could calculate the correlation coefficient by inputting the numbers from the scatter plots already shown into excel. A scatter graph would be generated, and the correlation coefficient easily calculated, see the section about using Excel for more information on this.

The correlation coefficients of the variables used in the scatter plots previously are:

- BV38 (5 GCSEs A*-C) and deprivation rank = 0.786
- Robberies recorded and deprivation rank = -0.768
- BV52 (Cost of intensive social care for adults) and BV54 (over 65s helped to live at home) = 0.042

So there is a strong positive relationship between BVPI 38 and the rank of deprivation score, a strong negative relationship between robbery offences recorded and the rank of deprivation score, and no relationship, for practical purposes, between BVPIs 52 and 54.

Do remember though that the correlation coefficient only shows the extent of a relationship between the two variables. **Correlation does not explain why these relationships exist.**

3. Presenting and using data

Once you've done the analysis through summarising and comparing the data you need to continue to make the most of the numbers by creating and presenting the data using tables, maps and charts.

There are three ways you can present and use the data:

1. Using Word
2. Using maps
3. Using Excel

Essential facts about presenting data

1. You can present data using tables, charts or maps.
2. The presentation you choose may be influenced by the type of data you are using, or the purpose of the analysis.
3. Presenting tables and charts to a good standard has become easier as software has developed.

Presenting and using data

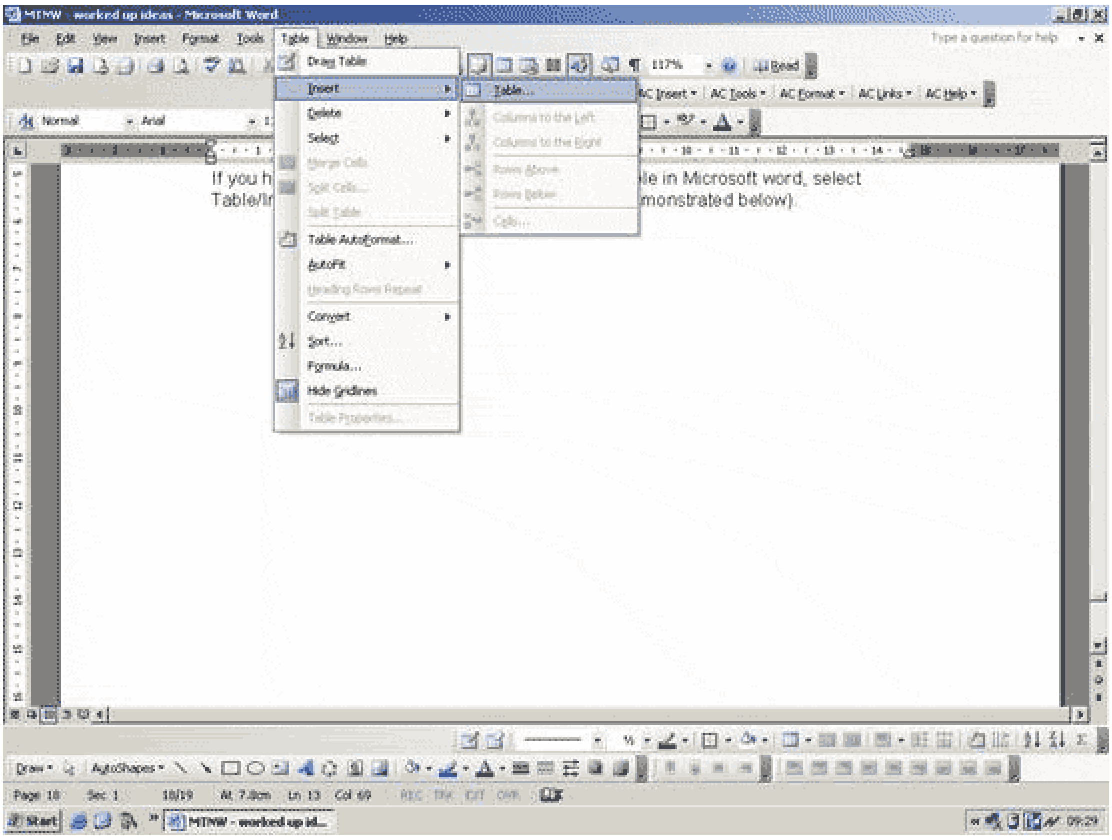
...using Microsoft Word

This area of the tool has three sections:

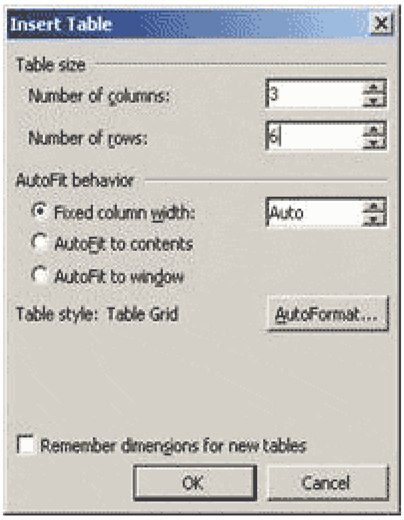
- Creating tables;
- sorting data; and
- creating charts using Microsoft Word.

Creating tables in Word

If you have a set of data you wish to present as a table in Microsoft word, select Table/Insert/Table from the drop down menus (as demonstrated below).



A dialogue box (see below) will appear asking how many rows and columns you would like in the table. Adjust this to suit the data you wish to present.



When you click 'OK', a blank table will appear in the document into which you can enter data. The example below compares five authorities' performance from two customer satisfaction surveys.

Table 1. (BVPI 89) The percentage of people satisfied with cleanliness standards in their area:

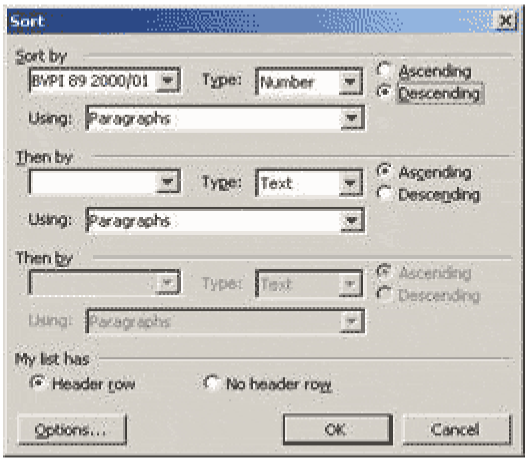
Authority	BVPI 89 2000/01 (%)	BVPI 89 2003/04 (%)
Authority A	55	64
Authority B	37	34
Authority C	46	52
Authority D	21	25
Authority E	46	37

The table has a clearly referenced title, which appears in bold type. The column headings are also clearly labelled in bold text and show the units that the data have been reported in. The alignment is clear, making understanding of the data easier. To aid this further, some text commentary could be inserted referring to the data.

Sorting data

Making sense of a few cases of data as in the table above is simple. But if the table presented hundreds of cases of data, it may be easier to **'sort'** the table so that high performing authorities can be easily identified.

You can do this by setting the cursor somewhere in the table and selecting Table/Sort from the drop-down menus. A dialogue box will appear:

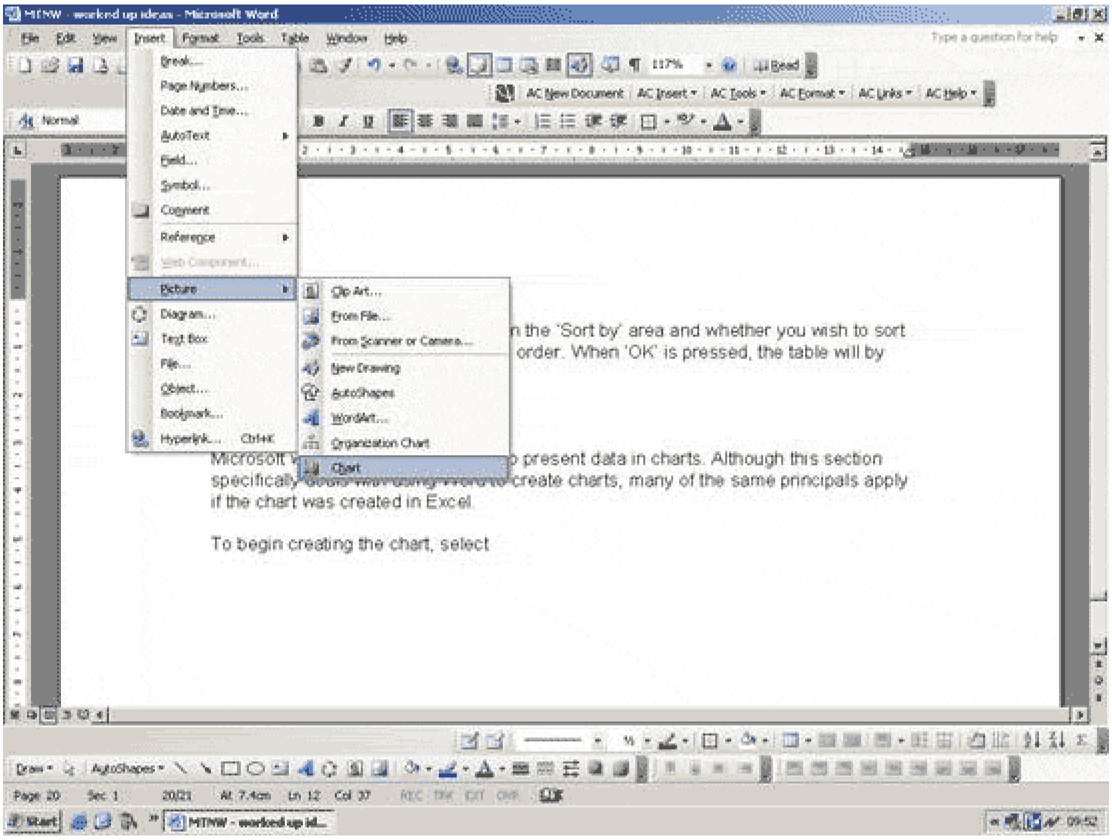


Select the field you wish to sort by in the 'Sort by' area and whether you wish to sort the table in descending or ascending order. When 'OK' is pressed, the table will be sorted in order.

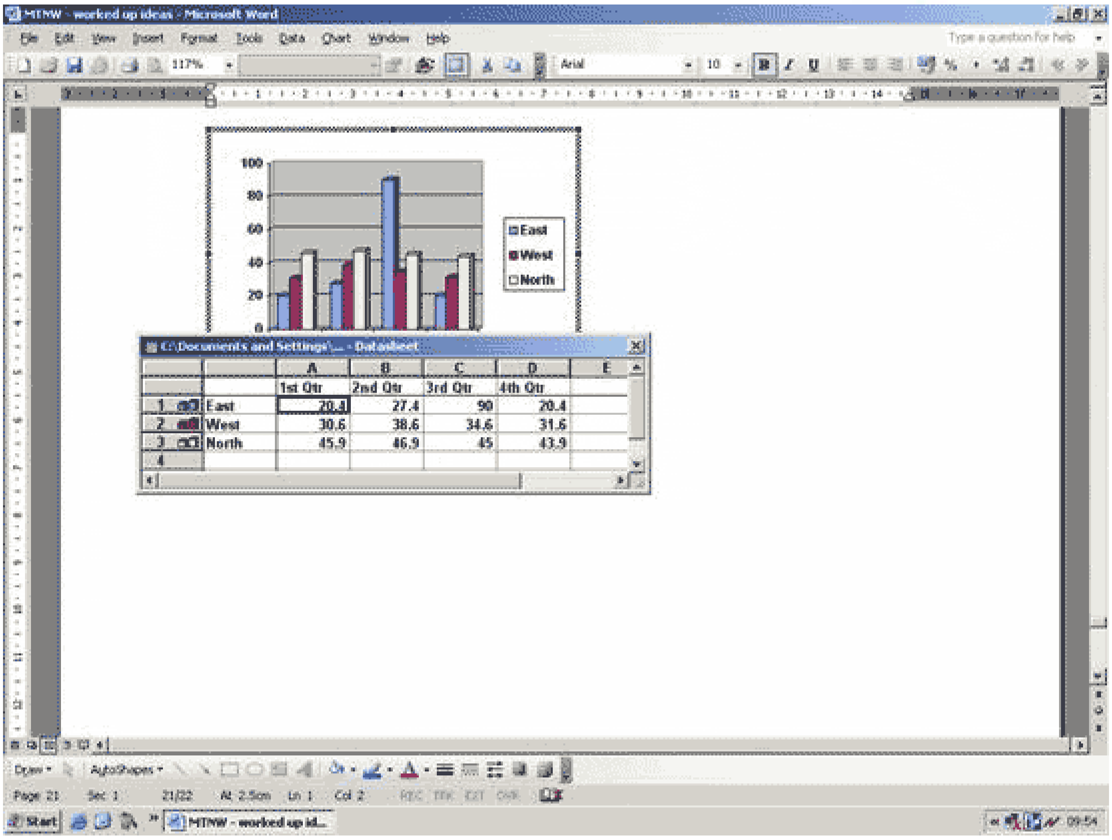
Creating charts in Word

You can also produce **charts** in Word. Although this section specifically deals with using Word to create charts, many of the same principals apply to creating charts in Excel.

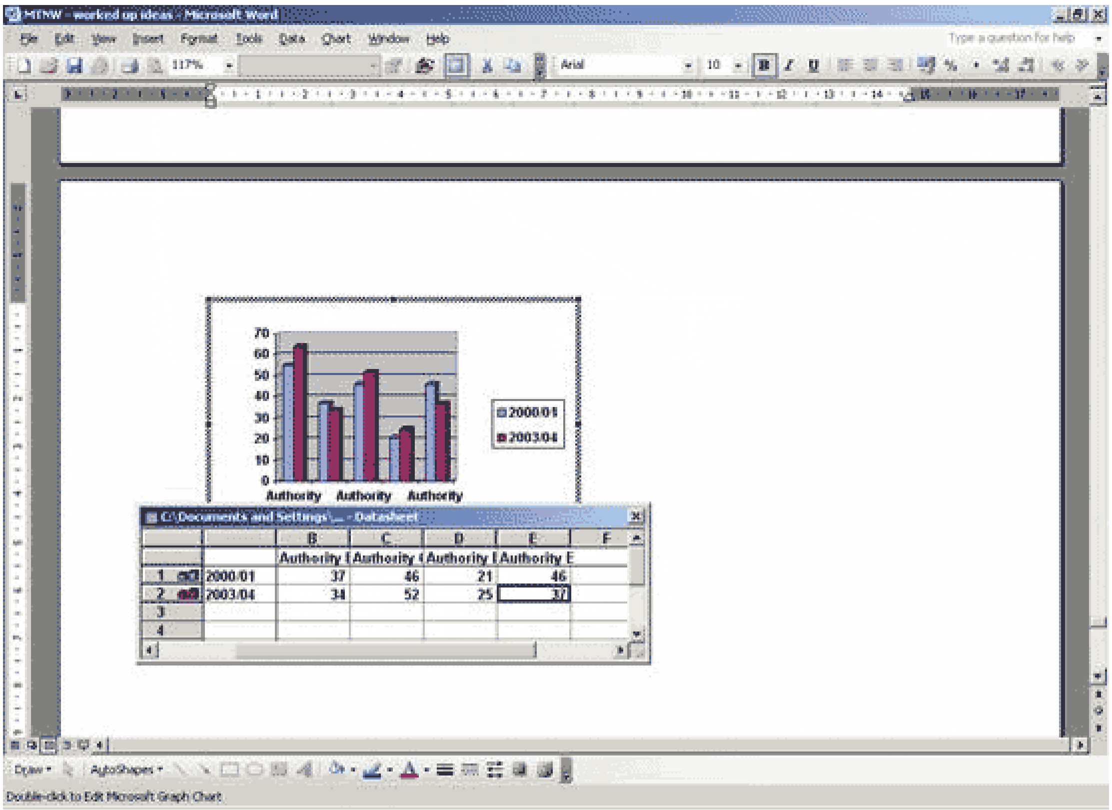
To begin creating the chart, select Insert/Picture/Chart from the drop down menus.



When 'OK' is pressed, the following example chart is inserted into the word document:

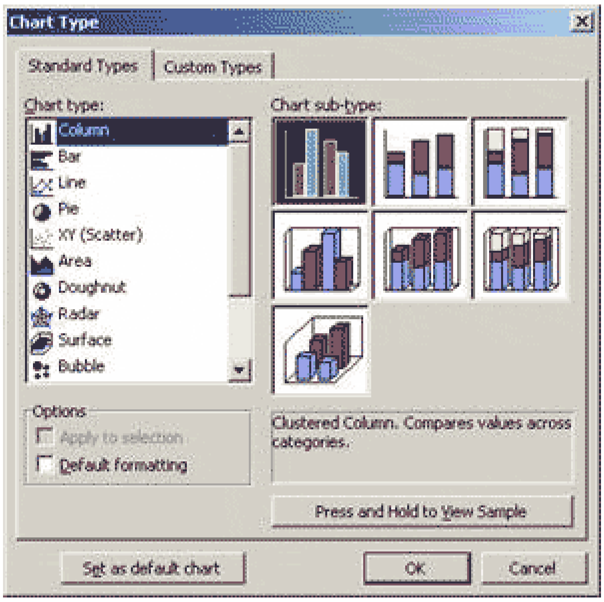


You can insert your data into the datasheet that appears on the screen.

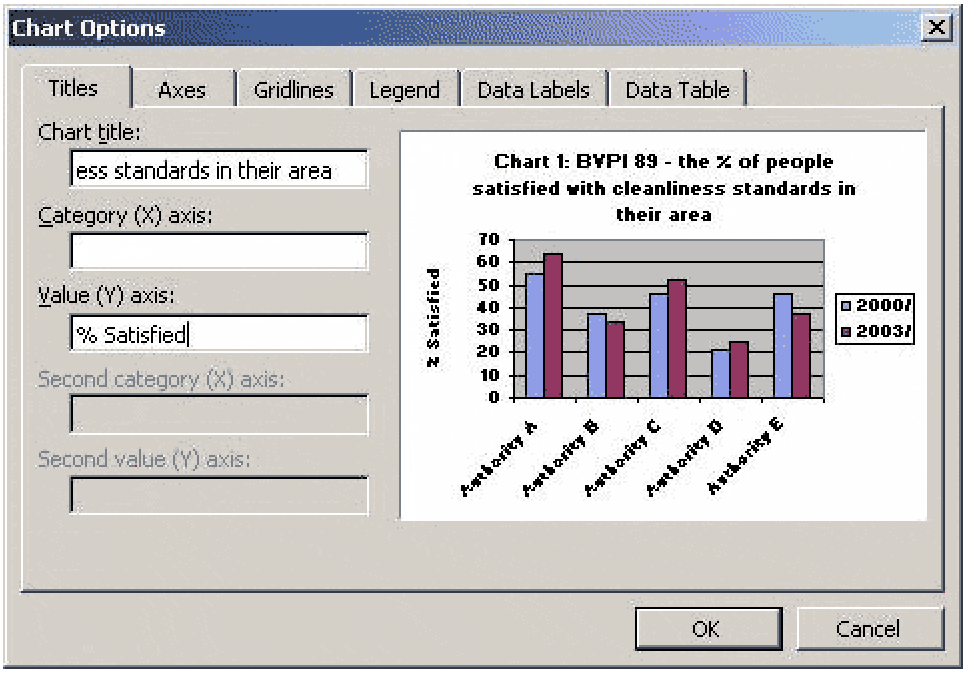


When you have finished entering your data, you will have an unformatted chart set in the default style. To start **formatting,** you will first need to **resize** the chart to its most appropriate size. This can be done by clicking on the chart and dragging it until it is a more appropriate size.

You may next want to change the **chart type**. Word provides you with a large variety of choices for types of charts. There are bar charts, column charts, pie charts, scatter plots - and each has a choice of sub-categories. You will have to choose the type of chart that is most appropriate for your data. But remember that charts should be kept fairly simple. Selecting a chart type can be done at any stage of the process, so you may wish to experiment with different chart types.



With the chart selected, you can now select Chart/Chart Options from the drop-down menu. This will give you a range of editing options for your chart, which you can select by clicking on the tabs along the top of this dialogue box:

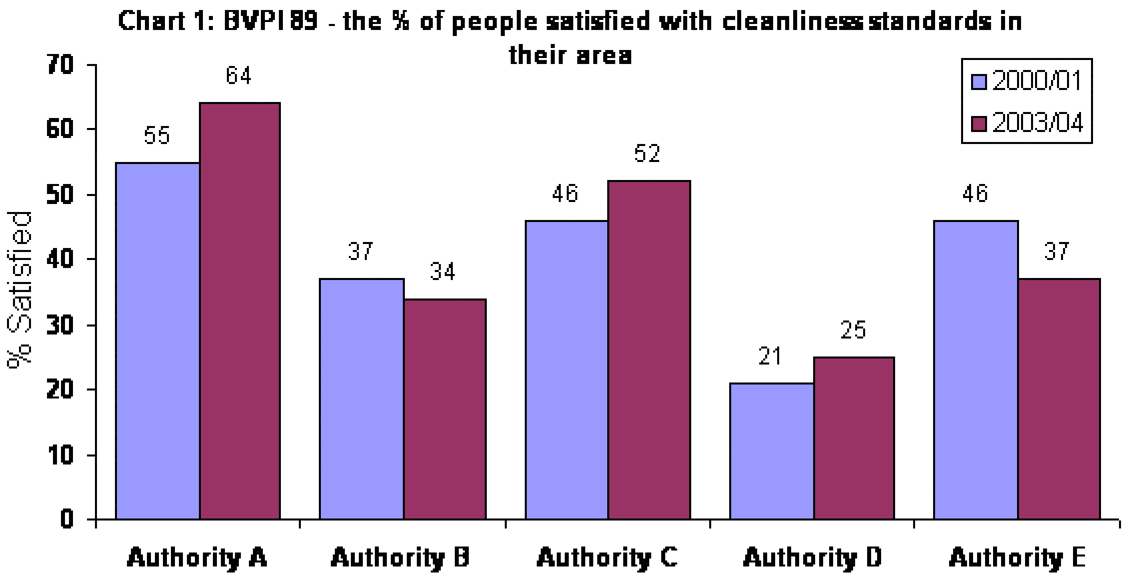


The first thing to do is give the chart a **clear title**, and **label the y-axis**. These are important so that anyone using the chart can easily identify what the chart refers to and what exactly has been measured.

Using Chart Options, you may also wish to delete the gridlines from the chart, move the legend or display the data values. These choices are really up to your personal preferences and the data you are displaying.

You can **edit** almost any part of the chart by double-clicking on that part of the chart you wish to alter. For example, you may wish to turn off the chart's background, alter the font of the labels or axis; you could also change the colour of the bars of the chart or the gaps between them by clicking on them. One good idea is to try and make sure the chart is as big as it can be within its outside border so that as much information as possible can be displayed.

The chart below is an example you may like to follow.

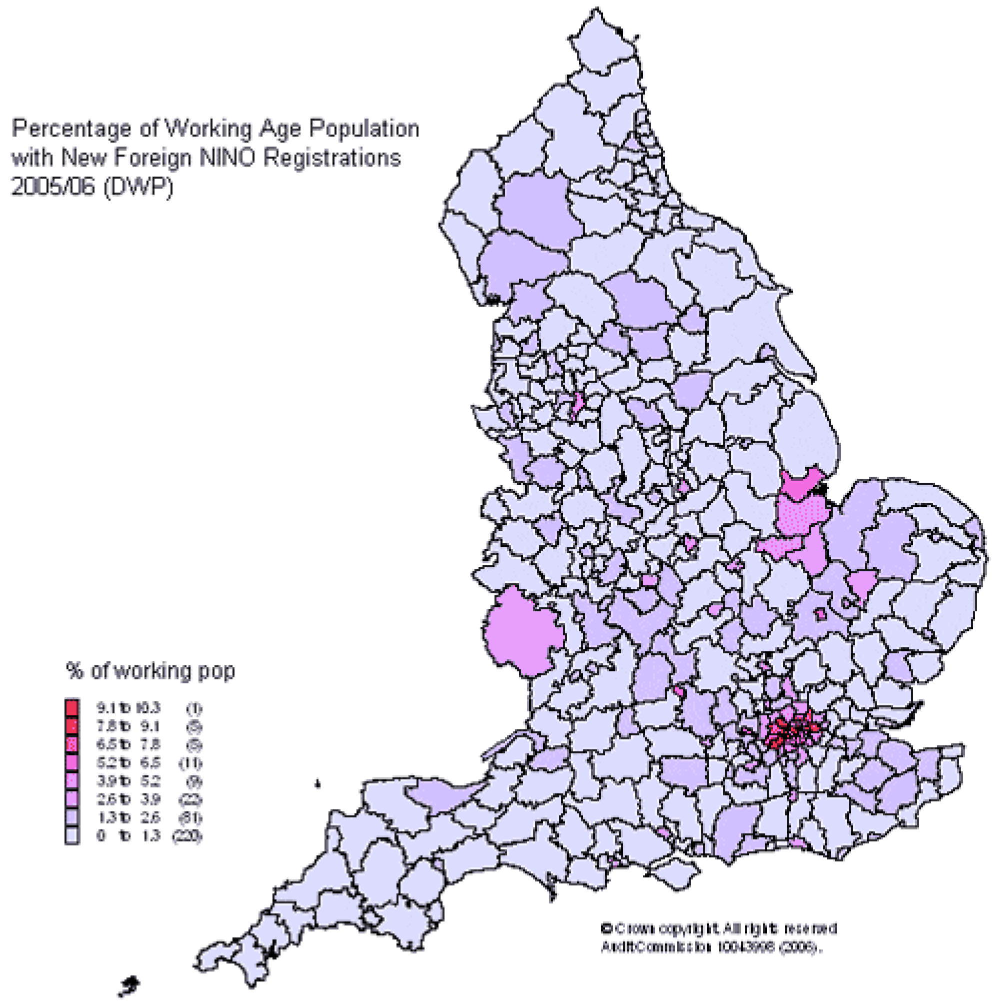


Presenting and using data

...using maps

Maps are a good way of displaying data as any geographical variations or patterns in the data can be shown easily. As GIS (Geographical Information Systems) software has become increasingly available during recent years, more and more data is being displayed in mapped form. There are now several websites which allow users to download and map data online.

The image below is an example of mapped data. It shows the percentage of working age population with new foreign National Insurance Number (NINO) registrations during 2005/2006, as recorded by the Department of Work and Pensions (DWP).



Presenting and using data

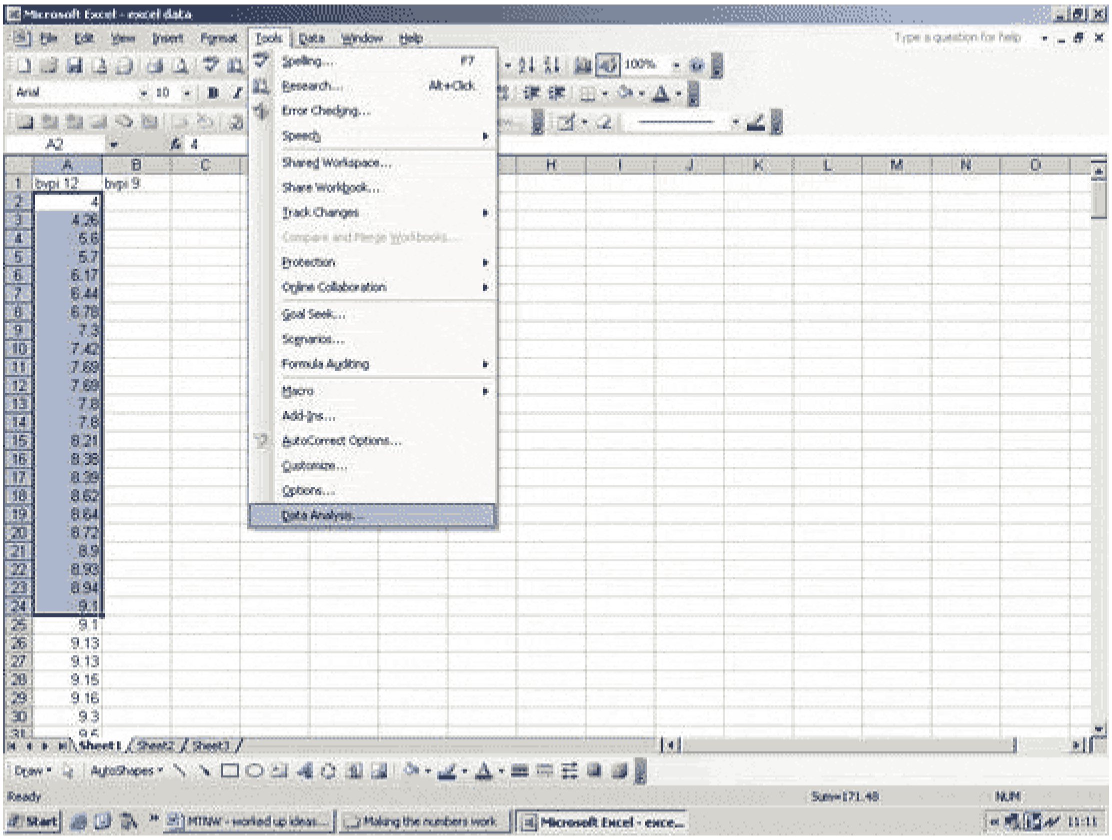
...using Microsoft Excel

This topic is divided into five sections:

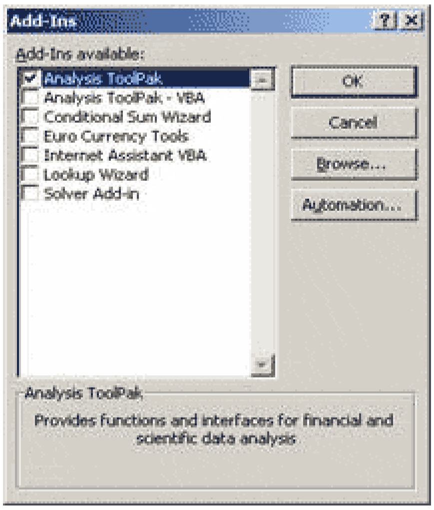
- Locating Excel's Data Analysis Tool;
- analysing descriptive statistics;
- calculating quartile statistics;
- producing a histogram; and
- correlation in Excel.

Locating Excel's Data Analysis Tool

A lot of analysis can by done using **Excel's Data Analysis Tool**. This can be found by selecting Tools/Data Analysis.

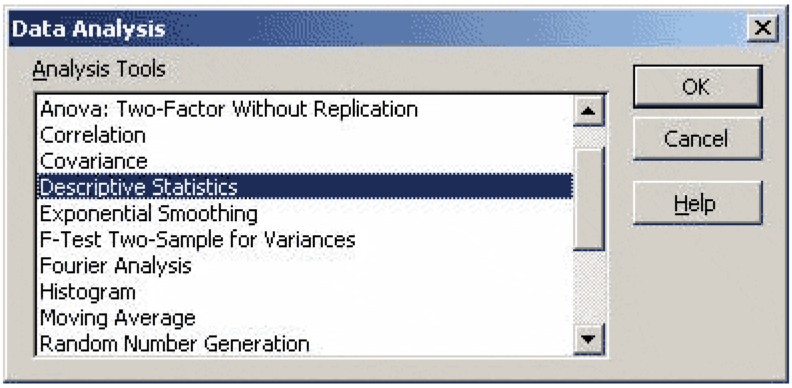


If Data Analysis does not appear under the Tools section of Excel, you can get access to it by selecting Tools/Add-Ins. The following dialogue box will appear. Check the **Analysis ToolPak** option and the Data Analysis tool should now appear under 'Tools'.



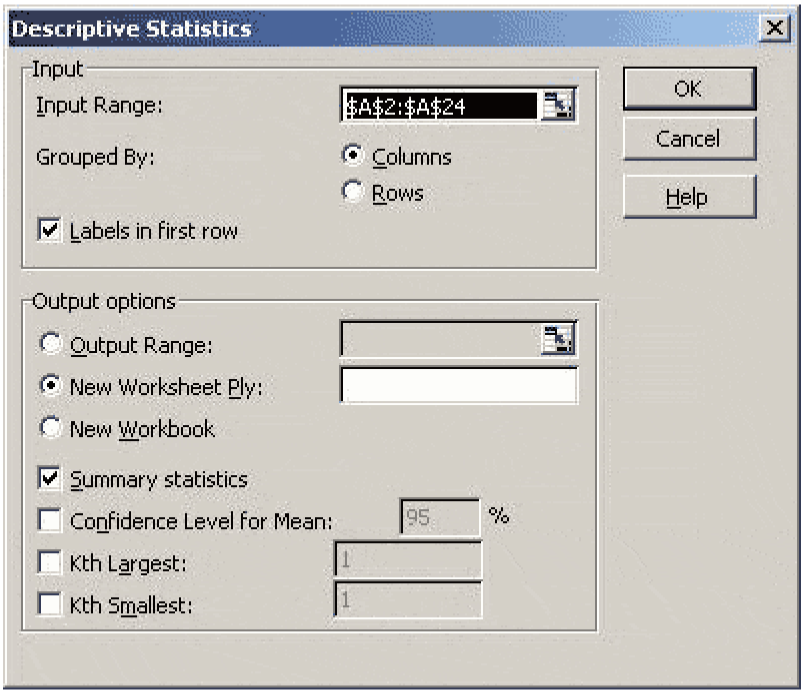
Analysing descriptive statistics

Several of the themes covered in Summarising Data can be explored by selecting Descriptive Statistics from the Data Analysis tool (see example below).

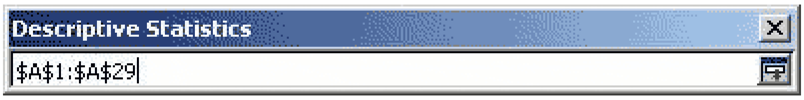


By selecting this option you can quickly determine **mean, median and mode; standard deviation, variance, kurtosis, skewness, range, minimum value and maximum value.**

When you select Descriptive Statistics from the Data Analysis Tool, the following dialogue box will appear:



Click on the 'Input Range' box, and the following box will appear:



Select all the data in the field or fields you wish to obtain summary statistics for, and click on the small box on the right of the input field.

Ensure that the 'Summary Statistics' check box is ticked. If you select 'New Worksheet Ply' the statistics will appear in a new sheet in your spreadsheet when you click'OK'.

The example below shows the descriptive statistics for BVPI 12 for a group of authorities:

Bvpi 12	
Mean	7.755357
Standard Error	0.280741
Median	8.295
Mode	7.69
Standard Deviation	1.485544
Sample Variance	2.206841
Kurtosis	0.693843
Skewness	-1.19873
Range	5.16
Minimum	4
Maximum	9.16
Sum	217.15
Count	28

NOTE: The standard deviation above is an unbiased estimator of the standard deviation. That is, it assumes that the data we have analysed is a sample, and calculates the unbiased estimator using n-1. If you wish to find the standard deviation and y our data is a population, you need to write a formula

Standard deviation formula

1. Select a blank cell near to the data you wish to obtain the population standard deviation for.
2. Type '=STDEVP' in the formula bar (which is near to the top of the screen next to 'fx'.
3. Now type an open bracket '('.
4. Now highlight the data you wish to calculate the population standard deviation for. If the data is not all together, highlight the first part, the type a comer ',' then highlight the second part and type a comer and so on until all of the data has been highlighted.
5. Then type a close brackets ')'.
6. Then press enter. The standard deviation of the population will now appear in the cell.

Calculating quartile statistics

A useful summary statistic that is not covered by the Descriptive Statistics tool in Data Analysis is quartile information. To calculate quartiles, you will need to write a formula

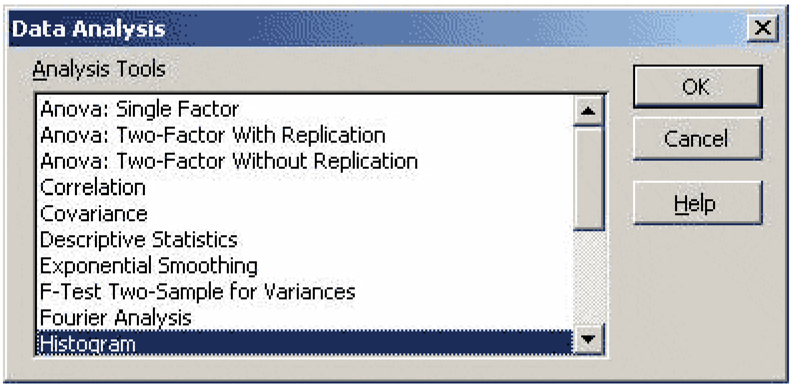
Quartiles formula

- Select a blank cell near the data you wish to obtain quartile information for.
- Type '=quartile' in the formula bar (which is near the top of the screen next to *'fx'*.
- Now type in an open bracket '('. A small pop-up box should appear which says 'QUARTILE(**array**,quart)'. This box is designed to help you put the right information in the correct order to allow excel to calculate the quartile. Array refers to the range of data you wish to get quartile data for. Quart refers to the quartile you wish to find out where '1' is the lowest quartile or 25th percentile, '2' is the middle quartile or median value, and '3' is the highest quartile or 75th percentile.
- Under the array section formula, highlight the data you want to find quartile information for. Your formula should now look something like: '=quartile(A1:A27'.
- Place a comma next to the array in the formula and the number corresponding to the quartile you wish to identify. Close the brackets. In the following example, we wish to find the lower quartile data (hence '1') for the data that is occupies cells A1 to A27. =quartile(A1:A27,1).
- Press enter, and the quartile information will appear in the cell.
- You can now change the quartile number in the formula (from 1 to 2 or 3) to obtain the other quartile statistics.

Producing a Histogram

You already know about histograms from the Summarising Data section. You can also do this using the Data Analysis tool in Excel.

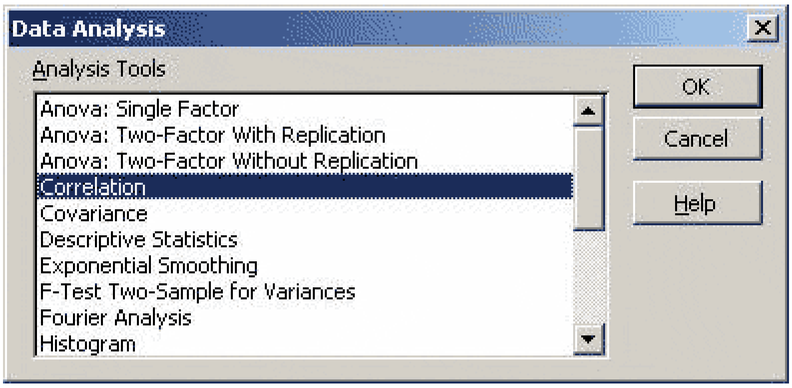
Select 'Histogram' from Data Analysis and click 'OK':



Highlight the data you wish to use for the histogram in the 'input range' and ensure that 'Chart Output' is checked. Click 'OK' and a histogram will appear showing the distribution of the data in another sheet of the spreadsheet.

Correlation in Excel

You can also use Data Analysis to perform **simple correlation** in Excel. Select 'Correlation' from the Data Analysis tool.



Click 'OK' and highlight the fields you wish to correlate in the 'input range'. Click 'OK' and the correlation coefficient showing the degree of association between the two variables will appear in a separate sheet of the spreadsheet.

When calculating a correlation in Excel, the variables you wish to correlate must be next to each other in the spreadsheet. So you may need to format the spreadsheet before you start.

4. Useful sources of data

Having learnt some of the issues of basic statistics and how to perform some analysis using Excel, you may like to explore some of the sources of data available to undertake further analysis.

This section of the tool is regularly reviewed and updated. The following is a brief introduction to where data can be found and used:

- The Office for National Statistics
- Neighbourhood Statistics
- NOMIS
- CIPFASTATS
- Area Profiles

The Office for National Statistics

The Office for National Statistics (ONS) is the government department that provides UK statistical and registration services. ONS is responsible for producing a wide range of key economic and social statistics which are used by policy makers across government to create evidence-based policies and monitor performance against them. The Office also builds and maintains data sources both for itself and for its business and research customers. It makes statistics available so that everyone can easily assess the state of the nation, the performance of government and their own position.

From the website you can access a wide range of economic, social, environmental and population data.

Neighbourhood Statistics

The Neighbourhood Statistics website is part of the ONS website. It allows users to access a range of demographic, social and economic data at a local area. Users can search by datasets and download them, or can search by area and find a wealth of information about small areas.

NOMIS

The NOMIS website gives access to detailed labour market statistics throughout the UK. Data which can be found on the site includes: Labour market and related population data for local areas from a variety of sources including the Labour Force Survey (LFS), claimant count, Annual Business Inquiry (ABI), New Earnings Survey (NES), and the Census of Population. Data from official government sources (mostly National Statistics). The latest published figures and time series data, in some cases back to the 1970s.

Nomis offers two main ways to access data: labour market profiles and detailed statistics. Labour market profiles give you summary data from a range of sources for a single area; the detailed statistics options give you access to the full range of data allowing you to query a single data source in greater depth and for multiple areas.

CIPFASTATS

The Chartered Institute of Public Finance and Accounting (CIPFA) provides a range of financial statistics via their website.

A password is needed to access the site.

Area Profiles

The Audit Commission's Area Profiles website provides sources of data and information instrumental to improving the quality of life and local services. It also provides tools to help you to construct an Area Profile.