Representing Confidence Intervals in Excel

Meic Goodyear
Public Health Intelligence Specialist
NHS Lewisham

Representing Confidence Intervals in Microsoft Excel

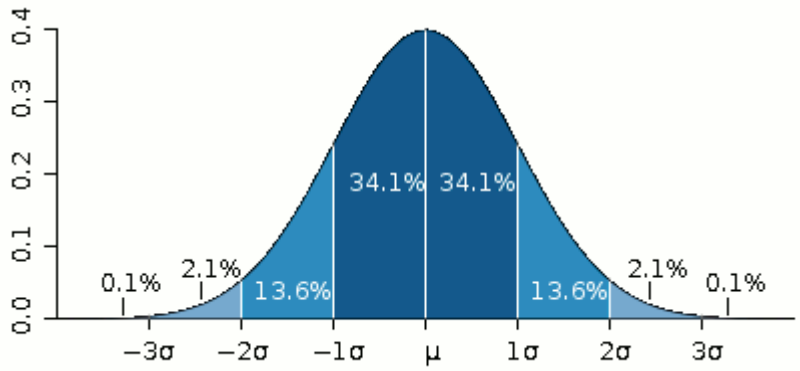
Introduction

Most data used in public health is not complete or precise. It generally involves some sort of sampling, and/or some sort of estimation. It will follow some sort of a probability distribution. It is highly desirable to have a way to include this uncertainty in statistics and in graphical displays. The most common way of doing this is by the use of *confidence intervals* (CIs), or for Bayesian statisticians their Bayesian equivalents, *credible intervals*. These represent the value, together with higher and lower limits within which we can be confident the "true" value of the variable lies within a specified range of probability. Typically, the value is 95%, meaning that if the value were measured 100 times then 95% of measurements would be within the stated range. Sometimes 99% is used, in which case the intervals are wider.
When the form of the distribution is known exact intervals can be calculated, but often this is not the case, and various methods have been devised as approximations according to the type of measurement or processing done. If the underlying distribution is symmetrical (e.g., a normal distribution) then the CI is also symmetrical and can be represented as

Value ± ½CI ,

otherwise the lower and upper CI bounds have to be separately specified.

Most distributions exhibit a degree of central tendency. That is, the closer to the average the more observations cluster. In technical terms, the value of the probability density function is greater nearer the centre. This is illustrated in the classic Gaussian (Normal) distribution's bell-curve for which the 95% confidence interval is just under 2 standard deviations from the mean (actually, 1.96 times the SD)



The normal distribution is symmetrical about its mean. Most public health data is not symmetrical, and the upper and lower CIs are often derived using approximations. This paper is not concerned with how CIs are calculated, but rather with how they can most effectively be displayed.

There are several ways CIs can be represented in Excel.

1 Stock charts

These, originally designed to represent maximum, minimum, and closing values of share prices, adapt easily to representing CIs. They require data to be presented in strict format:
ID, High, Low, Close or in health data terms,
ID, Upper CI limit, Lower CI limit, Observed value.

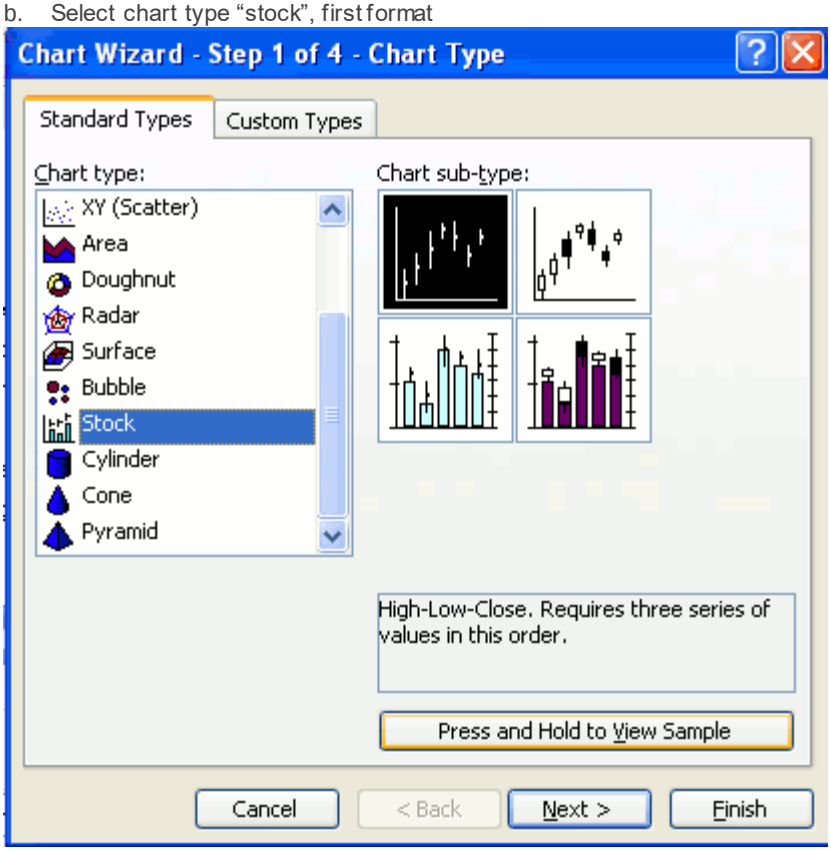
See the following example, using data for a cluster of schools for a particular indicator

School ID	Upper CI limit	Lower CI limit	Value
School1	0.360	0.052	0.150
School2	0.333	0.047	0.136
School3	0.404	0.018	0.100
School4	0.404	0.018	0.100
School5	0.335	0.061	0.154
School6	0.408	0.106	0.222
School7	0.328	0.031	0.111
School8	0.444	0.142	0.267
School9	0.613	0.233	0.409
School10	0.207	0.018	0.065
School11	0.593	0.234	0.400

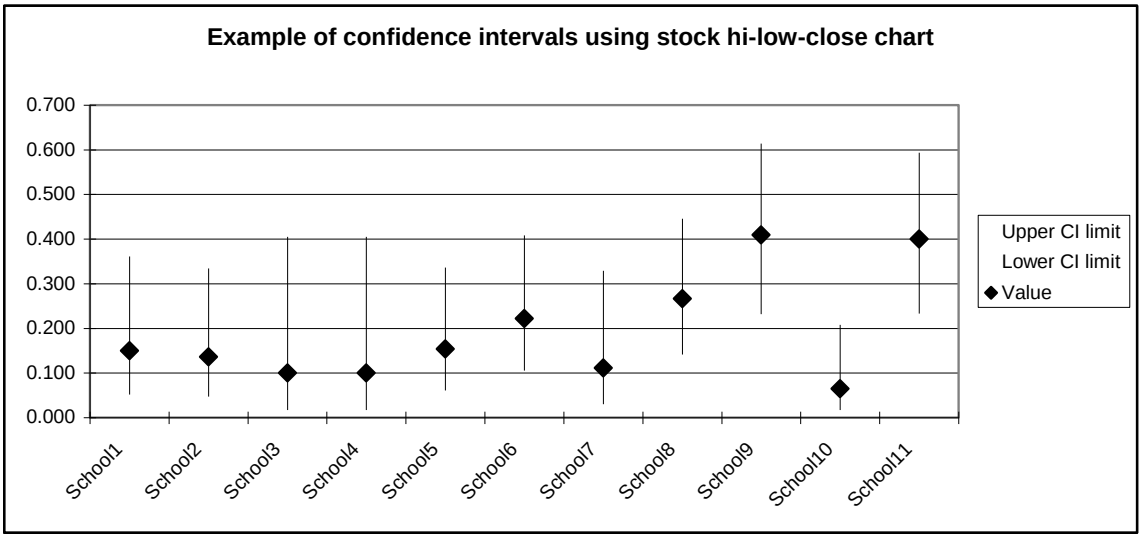
To produce the chart

a. Highlight the data and click on the Chart icon

b. Select chart type "stock", first format



c. Format the chart as you prefer.

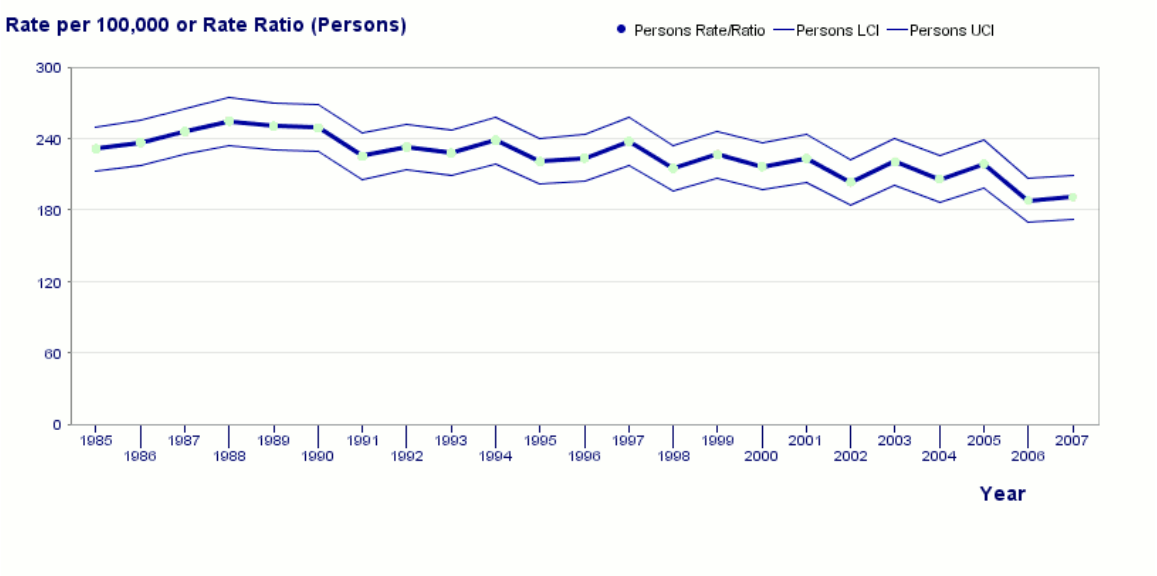


You have control over the point symbol for value, and can if you wish add marker symbols for the extremes of the CI bars. However, you cannot control the presentation of the vertical bars.

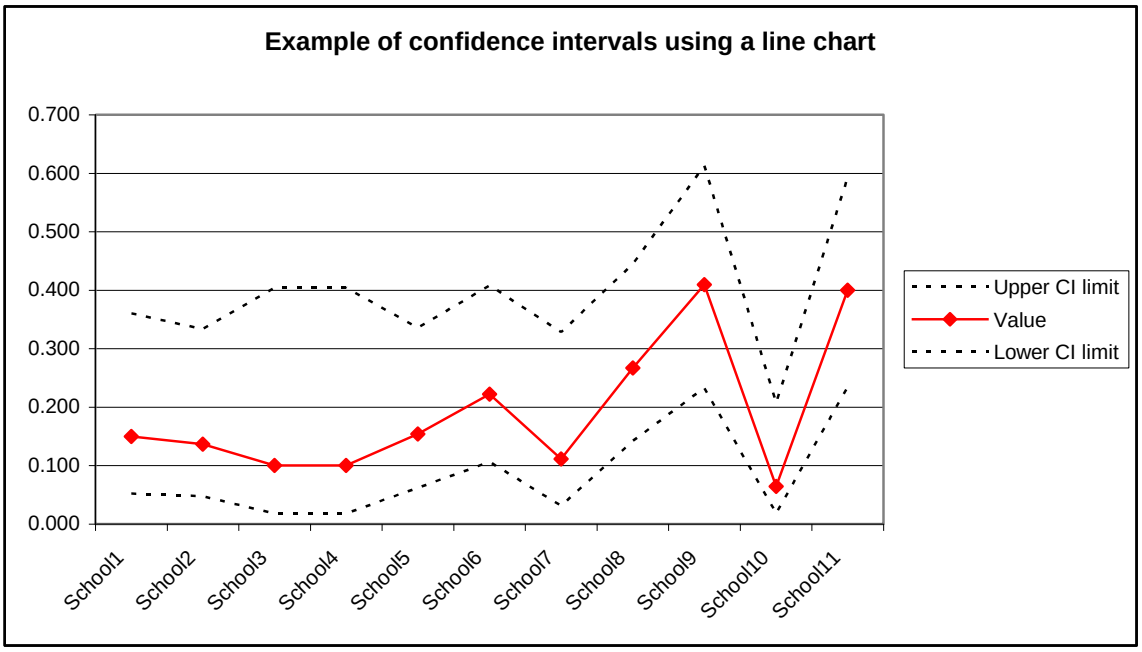
This presentation has the advantages that it is easy to produce, and clearly keeps categories separate. It is useful for creating caterpillar charts. It has the disadvantage that there seems to be no way to display more than one series per category (e.g. male and female) side-by-side on the same chart.

2 Lines

This is the method used in the National Cancer Database:



The value is represented by markers joined by lines, and the CIs by lines without markers. The figure for upper CI, lower CI, and observed value are charted as simple line charts, needing no instruction here.; the order of the series does not matter for the chart, but the columns should be arranged - Upper CI, Value, Lower CI - if the order of the Chart Legend is to reflect the position on the chart.



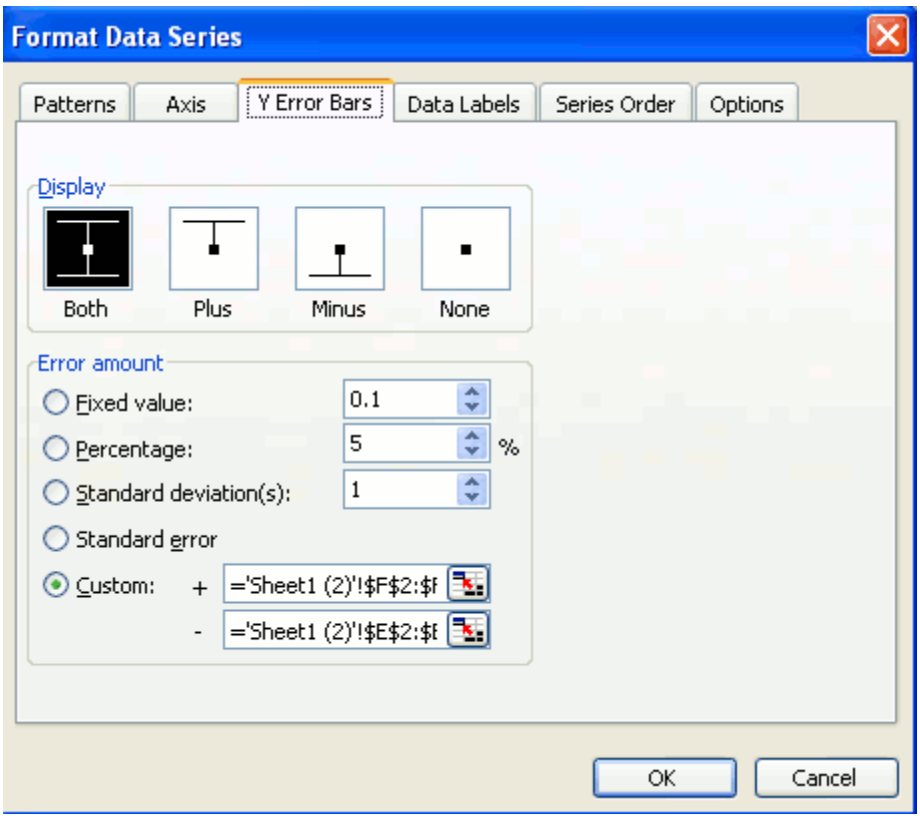
This presentation is effective when the x-axis is a continuum, for example when used in funnel plots, but it is seriously defective when the data is categorical or the values represent aggregates over a period. In the Example above there is no valid reason for connecting lines between schools, whose data is ordered alphabetically by ID. Even worse, in the example for the National Cancer Database the individual values relate to whole year aggregates, but a natural reading of the graph suggests it is possible to interpolate along the lines between values. This is very bad practice.

A better use of line charts involves the use of error bars. For this we need extra columns for the width of the upper and lower CIs. The error bars are calculated as

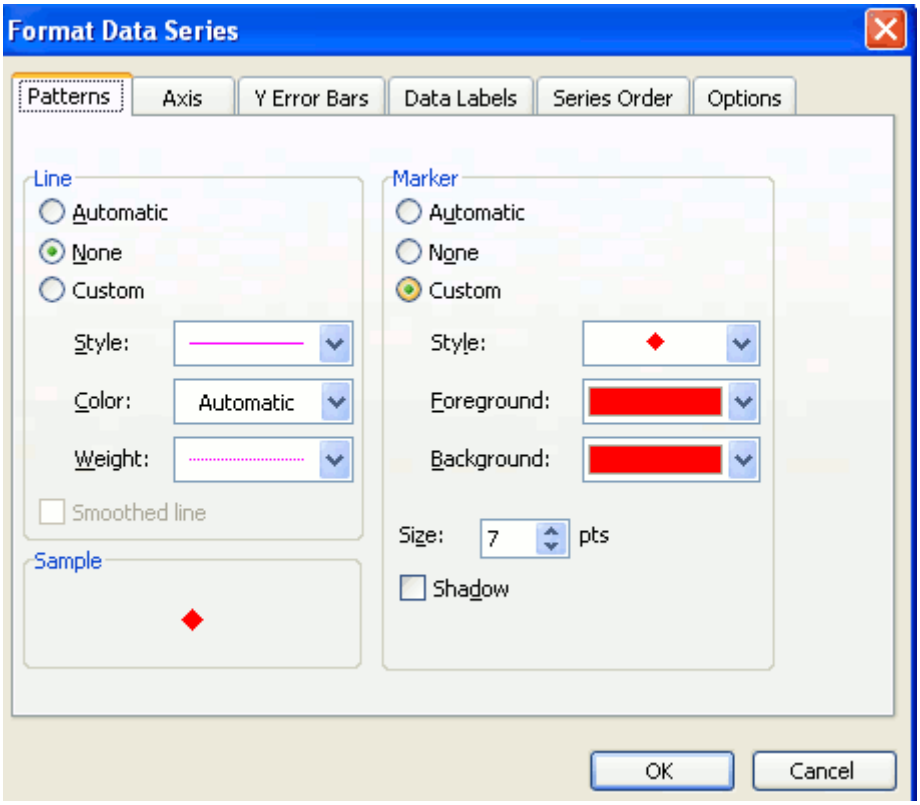
low error bar := Value – Lower CI limit
high error bar := Upper CI limit - Value

School ID	Upper CI limit	Value	Lower CI limit	low error bar	high error bar
School1	0.360	0.150	0.052	0.098	0.210
School2	0.333	0.136	0.047	0.089	0.197
School3	0.404	0.100	0.018	0.082	0.304
School4	0.404	0.100	0.018	0.082	0.304
School5	0.335	0.154	0.061	0.092	0.181
School6	0.408	0.222	0.106	0.116	0.185
School7	0.328	0.111	0.031	0.080	0.217
School8	0.444	0.267	0.142	0.125	0.178
School9	0.613	0.409	0.233	0.177	0.204
School10	0.207	0.065	0.018	0.047	0.143
School11	0.593	0.400	0.234	0.166	0.193

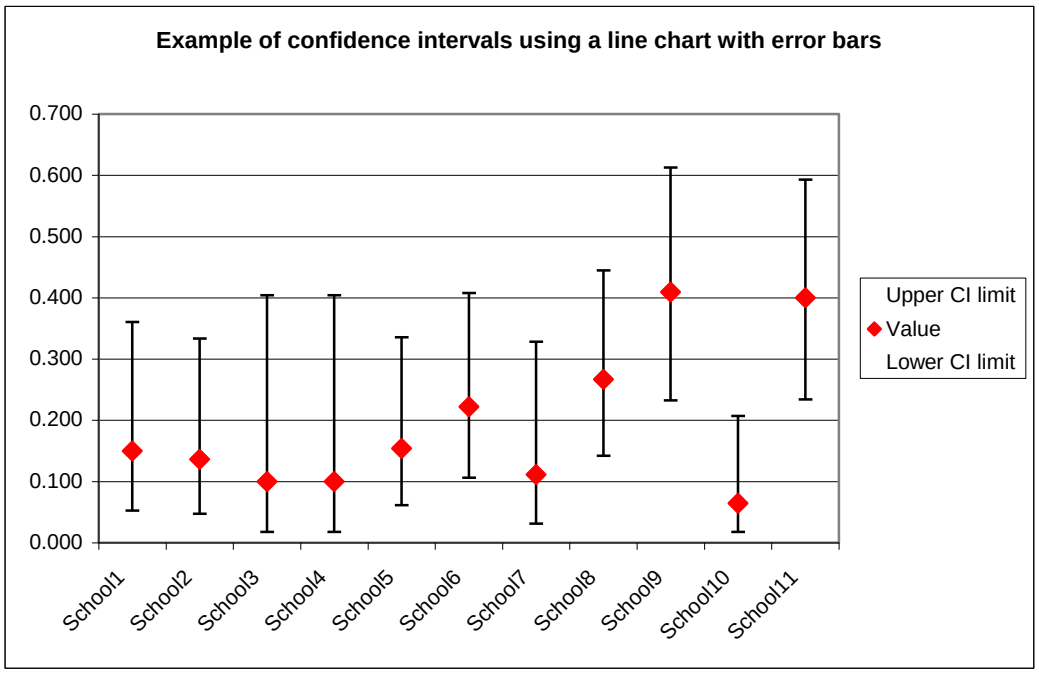
In the line chart above, right-click on the Value line, and click “Format Data Series” then click on tab Y Error Bars



Check the radio-button Custom. Click in the + box then highlight the data only (not the column heading) of the high error bar fields. Click in the – box, and highlight the low error bar data fields. Click on the Patterns tab and select the None radio button in the line box to get rid of the misleading line joining the values.



The result looks like this:



The effect is similar to the stock chart, but the user has control over the format of the error bars (colour and line pattern), though only limited control over the CI markers (no marker is not an option). It is possible to include more than one series, but the error bars become almost unreadable, so this cannot be recommended.

These formats are adequate if all that is to be considered is statistical significance at a particular level. However, by giving no indication of central tendency they lose some information that might usefully be communicated if a method of including this could be identified. Excel does not a facility for incorporating the outline of the probability density function alongside its error bars, but some of this information can be represented by using colour density in bar charts,

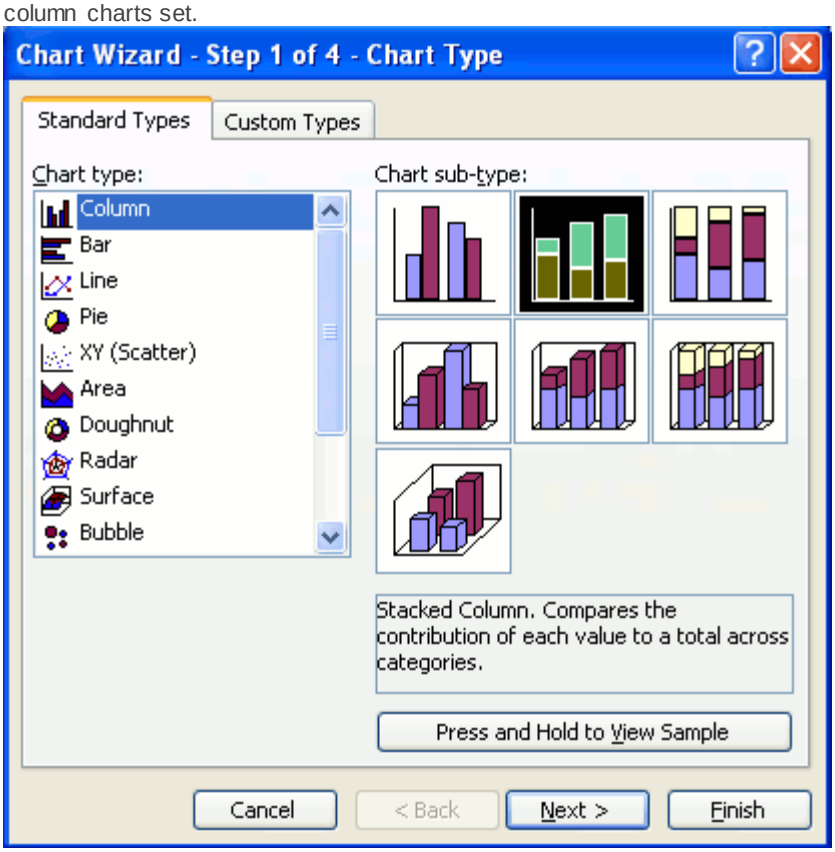
3 bar charts

Given that we start with value, lower CI limit and upper CI limit, we need two extra columns, the width of the lower CI and the width of the upper CI. These are calculated in the same way as the error bars in the line chart section:

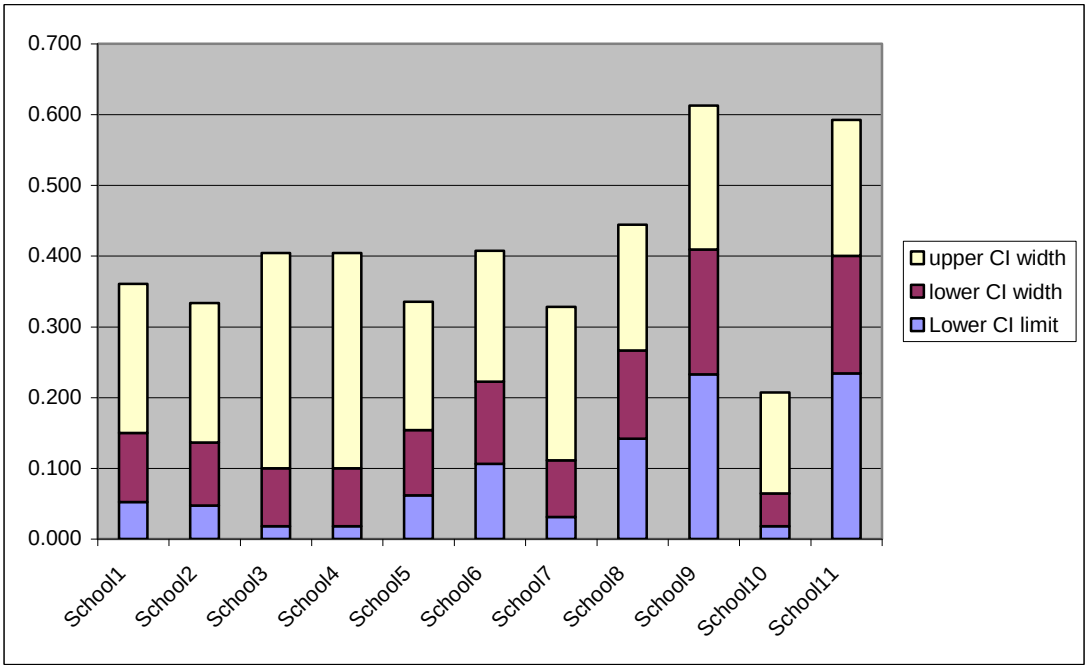
lower CI width := Value – Lower CI limit
higher CI width := Upper CI limit – Value

School ID	Lower CI limit	Upper CI limit	Value	lower CI width	upper CI width
School1	0.052	0.360	0.150	0.098	0.210
School2	0.047	0.333	0.136	0.089	0.197
School3	0.018	0.404	0.100	0.082	0.304
School4	0.018	0.404	0.100	0.082	0.304
School5	0.061	0.335	0.154	0.092	0.181
School6	0.106	0.408	0.222	0.116	0.185
School7	0.031	0.328	0.111	0.080	0.217
School8	0.142	0.444	0.267	0.125	0.178
School9	0.233	0.613	0.409	0.177	0.204
School10	0.018	0.207	0.065	0.047	0.143
School11	0.234	0.593	0.400	0.166	0.193

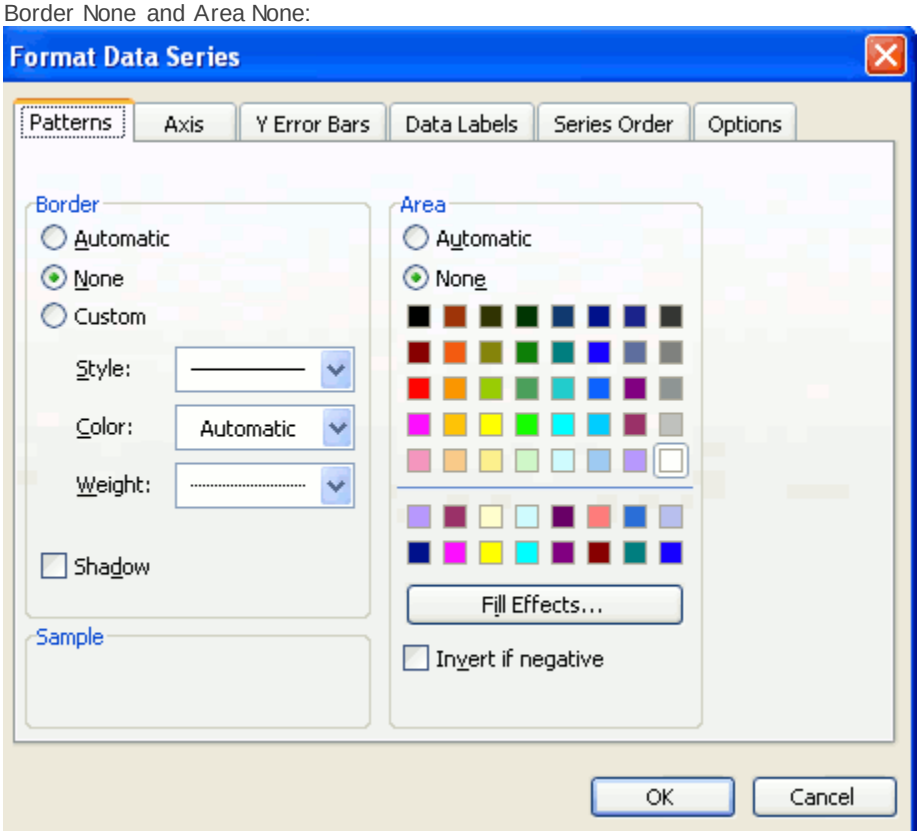
First create a stacked bar chart of ID, Lower CI limit, lower CI width, upper CI width. Highlight the columns including the headers (hold down the Ctrl-key and highlight using the left mouse button). Click the chart icon, and select the second Chart sub-type in the column charts set.



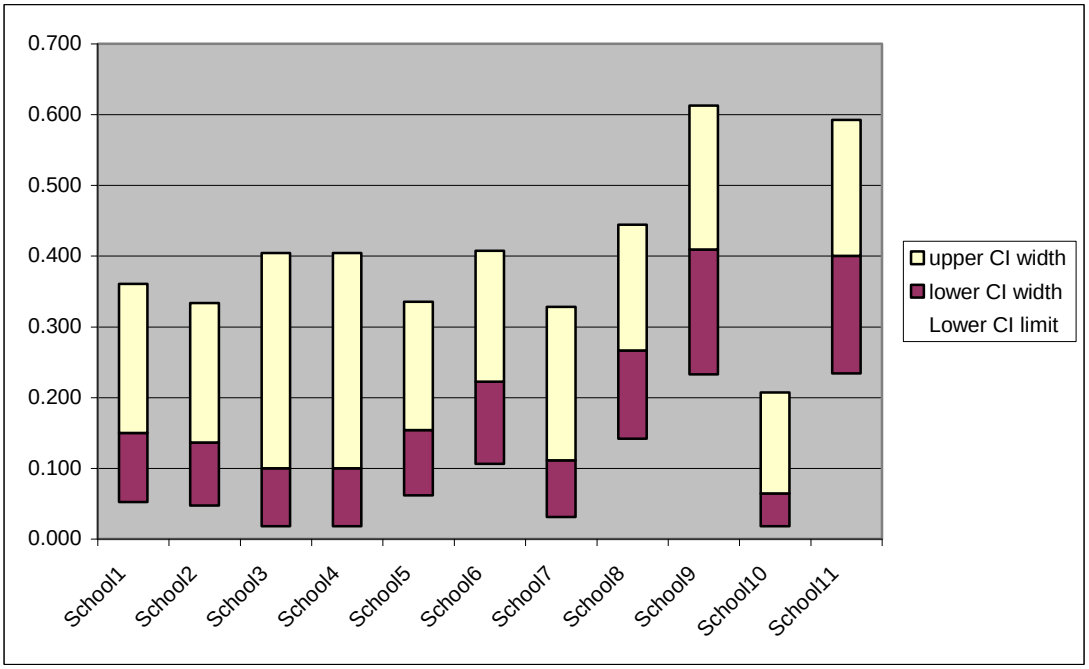
Either click on next to add chart titles etc or click finish. The chart will look something like this:



The next stage is to make the bottom series (zero – lower CI limit) invisible. Point to any of the bottom (blue) level, right click, and click Format Data Series. Select the radio buttons Border None and Area None:

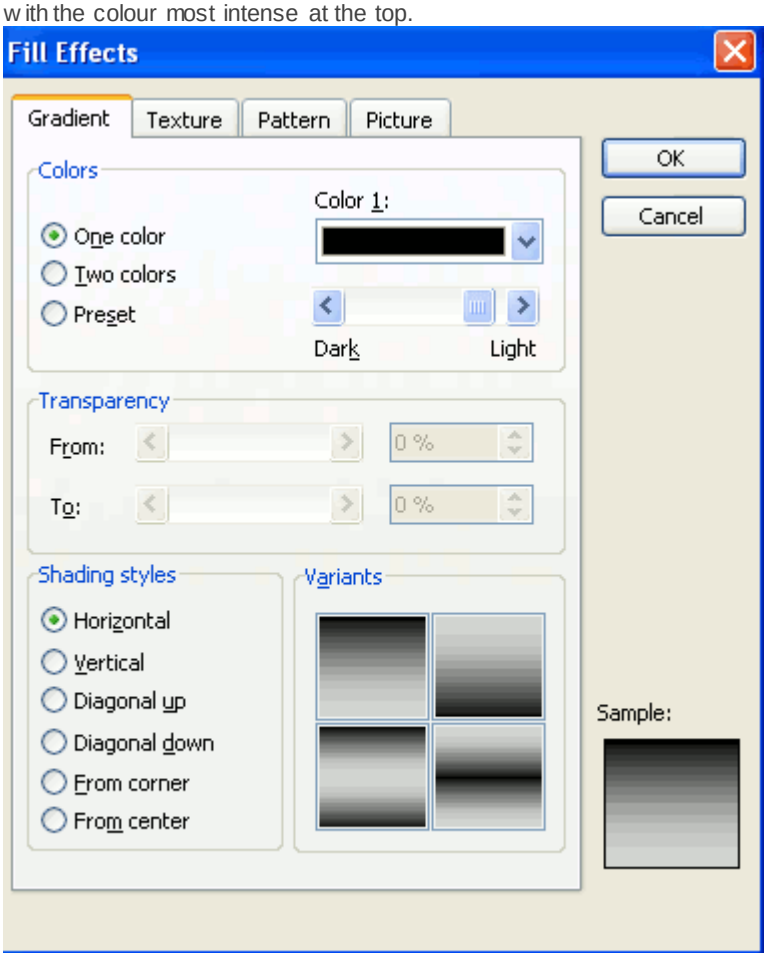


The chart will now look like this:



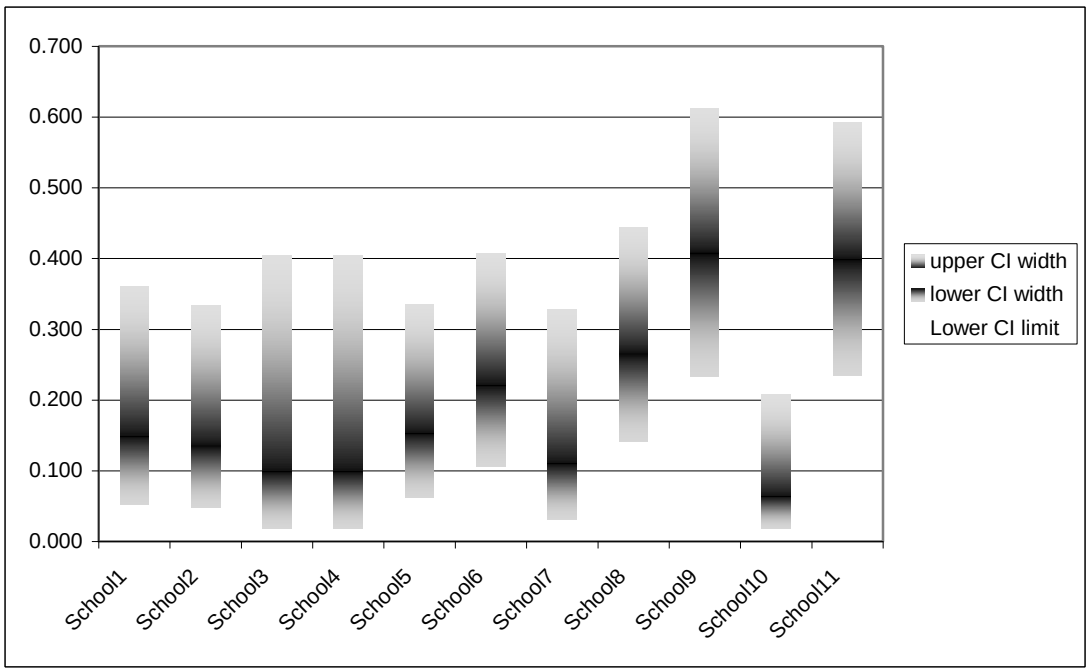
Now we want to do the bit with the shading. Highlight the lower CI width series, right-click and format selected data series.

In the Patterns tab click the None radio button in the Borders box, then click on Fill Effects in the Area box.
In the Gradient tab click on the One colour radio box, select your preferred colour, and move the slider between Dark and Light until the sample reflects the shading you prefer, with the colour most intense at the top.

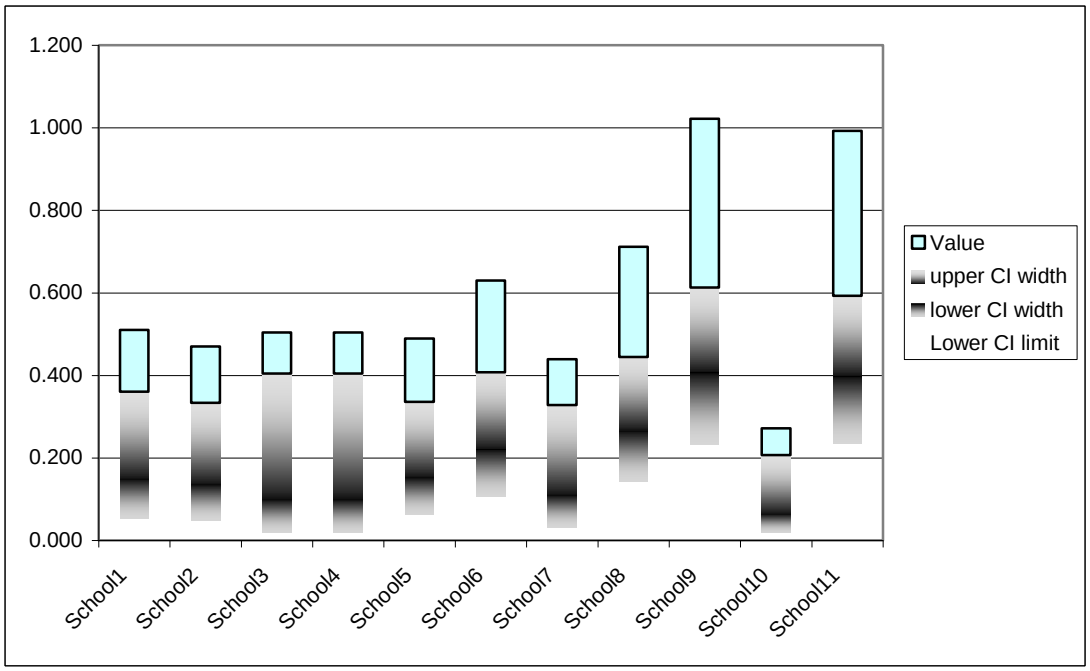


Repeat the process with the Upper CI width series, but with the colour intensity highest at the bottom.

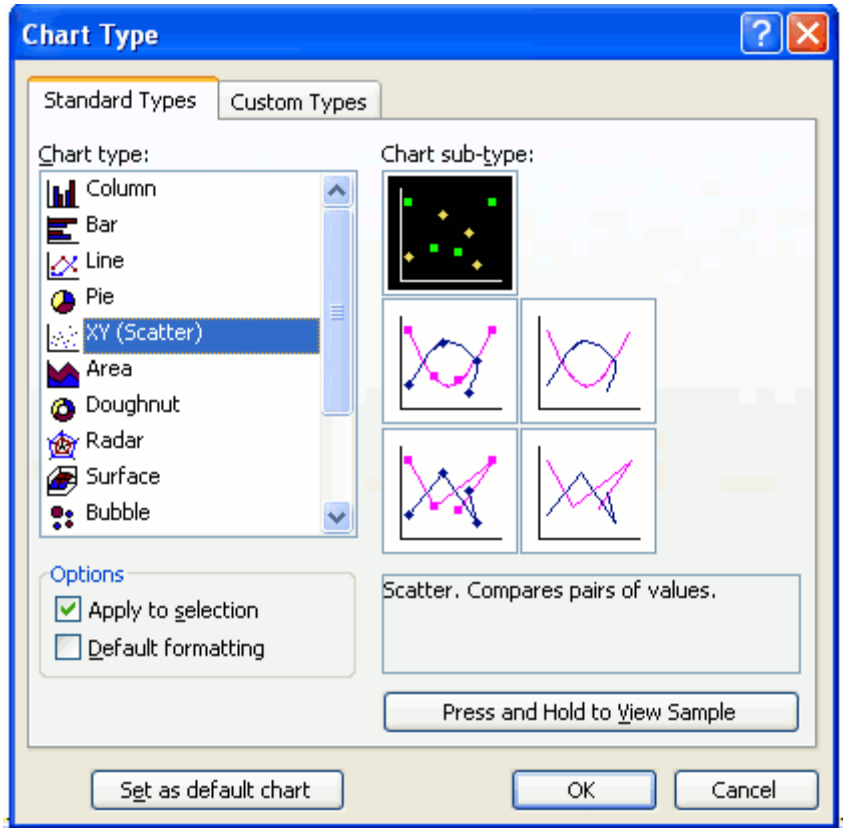
(optionally) click on the chart background, Format Chart Area and select None in the Area Fill box. Your chart will now look something like this:



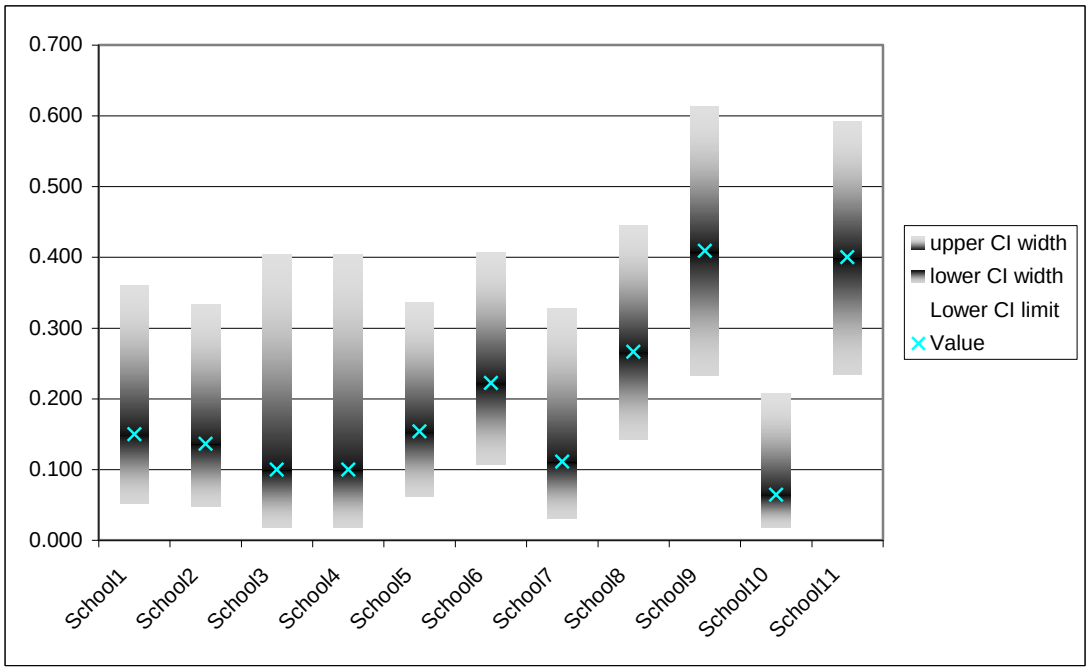
To show the Value also, click on the chart to highlight it, select
Chart
Add data from the menu bar
Highlight the Value heading and data, and click OK.
Your chart will now have the Value data added as an extra stack in the column:



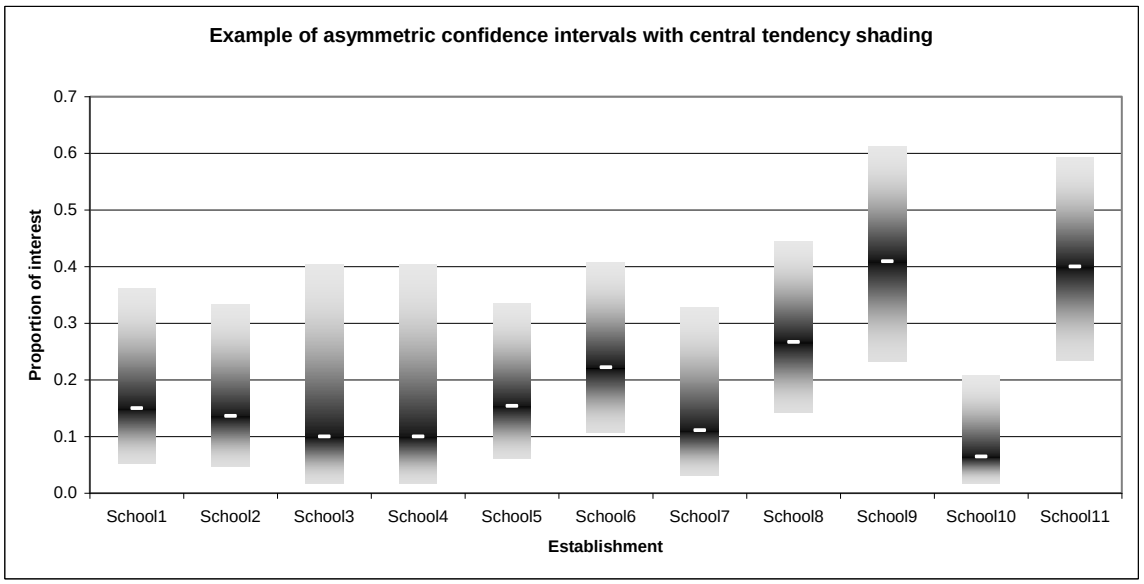
Right-click once on any block in the new series and select Chart Type (or left-click and select Chart Type from the menu bar). In the Standard Types box select Scatter, and the sub-type with no connecting lines:



The value series is now changed to a marker overlaid on the shaded CIs:



Highlight the series to select a different marker. Add titles in Chart Options. Tidy up the value axis in Format Axis. Edit the legend box to remove the Lower CI limit from the legend, or even get rid of the legend altogether. Final result:

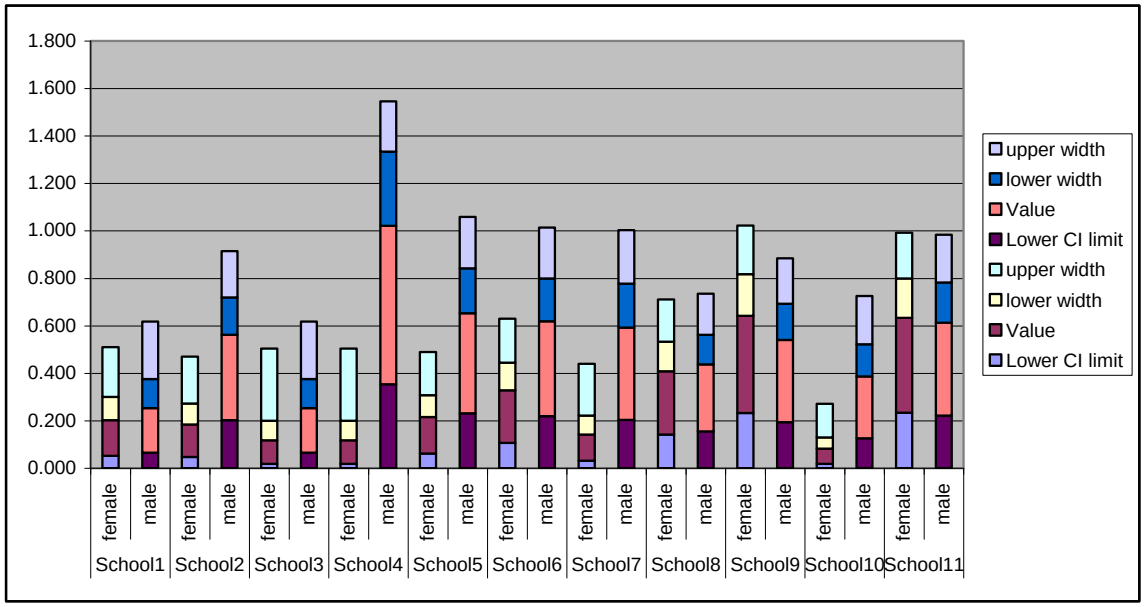


Using this technique it is possible to incorporate more than one series on the chart without loss of clarity. To do this you need to repeat your column headings across the worksheet once for each series, and to intersperse the series using separate lines. The above example related only to girls. Let us compare the indicator for girls and boys by school. Arrange the data thus:

School ID	Sex	Lower CI limit	Value	lower width	upper width	Lower CI limit	Value	lower width	upper width
School1	female	0.052	0.150	0.098	0.210				
	male					0.07	0.19	0.122	0.243
School2	female	0.047	0.136	0.089	0.197				
	male					0.20	0.36	0.158	0.195
School3	female	0.018	0.100	0.082	0.304				
	male					0.07	0.19	0.122	0.243
School4	female	0.018	0.100	0.082	0.304				
	male					0.35	0.67	0.312	0.213
School5	female	0.061	0.154	0.092	0.181				
	male					0.23	0.42	0.190	0.216
School6	female	0.106	0.222	0.116	0.185				
	male					0.22	0.40	0.181	0.213
School7	female	0.031	0.111	0.080	0.217				
	male					0.20	0.39	0.186	0.225
School8	female	0.142	0.267	0.125	0.178				
	male					0.16	0.28	0.126	0.172
School9	female	0.233	0.409	0.177	0.204				
	male					0.19	0.35	0.152	0.192
School10	female	0.018	0.065	0.047	0.143				
	male					0.13	0.26	0.135	0.204
School11	female	0.234	0.400	0.166	0.193				
	male					0.22	0.39	0.170	0.201

As you can see, there are two category heading columns, School ID and Sex, the series names are present once for each of the sexes, and the female series values have blanks in columns where the male series has values, and vice versa.

Highlight the whole table, and click the Chart icon, and select stacked columns as before. The result will look like this:



Each series has a different colour. Because of the blank values in the table, the two sexes have completely different colour schemes.
Highlight the two series labelled Lower CI limit, and set the Border and Fill Area to None, as previously. You will need to do this separately for males and females. Format the upper width and lower width series for each sex as before, choosing appropriate colours.
Highlight the Value series for each sex, Change the Chart Type to Scatter with no connecting lines. Tidy up the format. Using the pink for girls, blue for boys cliché, we end up with something like this: