Box and Whisker Diagrams:
Getting Microsoft Excel to plot them for you

Introduction

This page takes us a stage further down the road of analyzing the dispersion or variability of data sets. When we discussed the standard deviation, we worked with various data sets and found that some of them contained disperse data and others were hardly disperse at all.

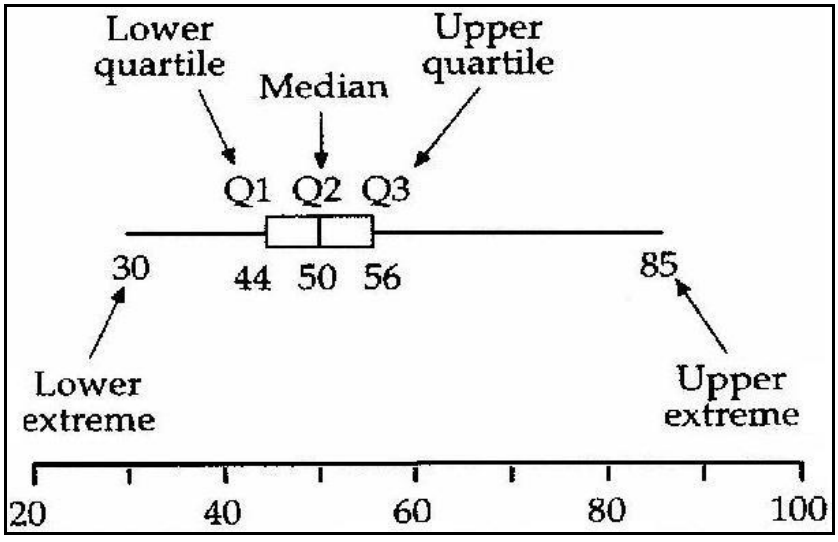
We used the standard deviation as a measure that allowed us to compare data sets and say that if one data set had a standard deviation of, say, 3 and the other data set had a standard deviation of, say, 1.2, then we could conclude that the first data set was more disperse than the second data set.

The problem with the standard deviation, however, is that many people find it an abstract idea and because of that they find it difficult both to calculate and interpret. This page concerns box and whisker diagrams that are not only relatively straightforward to interpret but they are visually representations of the dispersion of data sets so that drawing them and applying them is not such an abstract process.

In order to make this series of pages as complete as possible, we have included coefficient of variation values for all of the examples we use here: we discussed this value when we looked at the means and standard deviations of grouped data.

A box and whisker diagram, or boxplot, provides a graphical summary of a set of data based on the quartiles of that data set: quartiles are used to split the data into four groups, each containing 25% of the measurements. We use the terms box and whisker and boxplots interchangeably.

The 'box', or rectangle, we can see in the diagram below contains 50% of the data, and the extremes of that box are the Q1 and Q3 quartiles: the median value of the data set is the Q2, second quartile, value. Each 'whisker' represents 25% of the data and the extremities of these whiskers are the minimum and maximum values of the data.



Quartiles are defined as being the 25th, 50th and 75th percentiles that contain 25%, 50% and 75% of the data of a data set respectively. More formally, for a percentile:

> *Let $x_1, x_2, \ldots x_n$ be a set of n measurements … arranged in increasing (or decreasing) order. The **p^{th} percentile** is a number x such that p% of the measurements fall below the p^{th} percentile and (100 - p)% fall above it.*
> McClave and Benson p84

Following an email discussion with Sajjad, here is a little more discussion on Quartiles and their derivation:

The solution to your problem has, unfortunately, several answers. Let me give you just two, though, and mention a third.

1 I used MS Excel's in built QUARTILE function to find the values for me:

=QUARTILE(ARRAY,QUART)

Meaning to find the lower (25%) quartile of a data set, select the entire range of the data set that has been put into a spreadsheet in ascending order and then tell Excel to find the Q1, the lower quartile.

In your case, I entered all of your data starting in cell A1 so my quartile functions for Set 1 are:

=QUARTILE(ARRAY,QUART)
=QUARTILE(A2:A11,1) = 14.75
=QUARTILE(A2:A11,3) = 30.50

Exactly as you found and note that the QUARTILE function has been part of Excel since at least Excel XP but i can't guarantee that it was included in earlier versions.

2 Defining the lower quartile as the median of lower half of a data set, I used MS Excel as follows for your Set 1 data:

=MEDIAN(A2:A6) = 13
=MEDIAN(A7:A11) = 31

Try these calculations for yourself for your set 2 data Sajjad.

A third method is to include the median of the entire data set in the Q1 and Q3 calculations if the total number of data points is an odd number; and you can read about that here:

An explanation from Dr Math
An explanation from The Shodor Education Foundation

I know it's still a bit complicated and unclear but that's statisticians for you. Let me know if I can help any more.

Here is the data set that Sajjad sent me to work on:

Set 1	Set 2
11	21
13	45
13	52
20	55
27	57
29	67
29	69
31	78
47	88
48	92

We will see with the boxplots we work on here that some will look as if they have been squeezed into a tiny area and others look very long: the length of the whiskers and the boxes tell us the dispersion of the data set: much more clearly than even the best standard deviation value!

Boxplots are especially useful when comparing two or more sets of data.

Unfortunately, there is no boxplot utility in Microsoft Excel 5, 95, 97 or XP. However, boxplots can be drawn quite easily in Excel and this page will show you how!

The method that is demonstrated here works with Excel 97 and XP and we have left the method for Excel 95 at the end of the page: full credit is due to Neville Hunt who provides these solutions on his web site.

Box and Whisker Diagrams in Excel

What follows is largely based on the method given by Neville Hunt but the final two elements have been devised for this page … remember, you read it here first!

Suppose we have data from three groups, A, B and C for which we want to plot a box and whisker diagram.

Calculate the statistical functions quartile 1, min, median, max and quartile 3 **in that order** for each data set. Arrange the results on an Excel worksheet as shown below.

Statistic	**Group A**	**Group B**	**Group C**
Q1	20	22	30
min	10	15	18
median	40	45	50

max	100	110	90
Q3	70	75	57

In Excel 97:

Highlight the whole table, including figures and series labels then click on the Chart Wizard … let me give you a **Top Tip** at this point: rather than clicking on the Wizard, just press the F11 key … your chart will appear instantly on its own sheet … then click on the Wizard and follow Neville Hunt's method. The reason for this is that if you use the Wizard straight away your graph will appear on your worksheet whereas by pressing the F11 key, your chart will open up in a screen by itself and will be much easier to work with.

select Line Chart

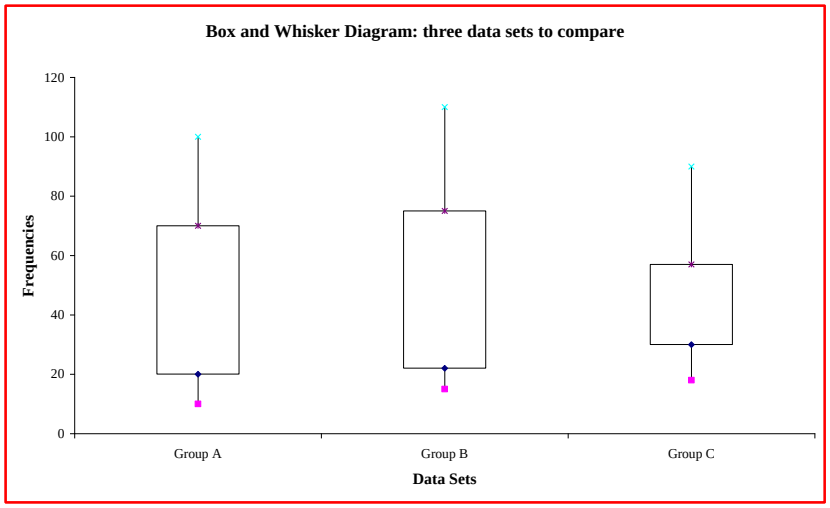
at step 2 make sure you select Plot by Rows, (otherwise Excel will use Plot by Columns and that's no good in this case), click on Finish now

right click on each line on the graph in turn and use Format Data Series to remove the connecting line

right click on **any** line on the graph and use the Format Data Series; select the Options tab and switch on the checkboxes for *High-Low lines* **and** *Up-Down bars*

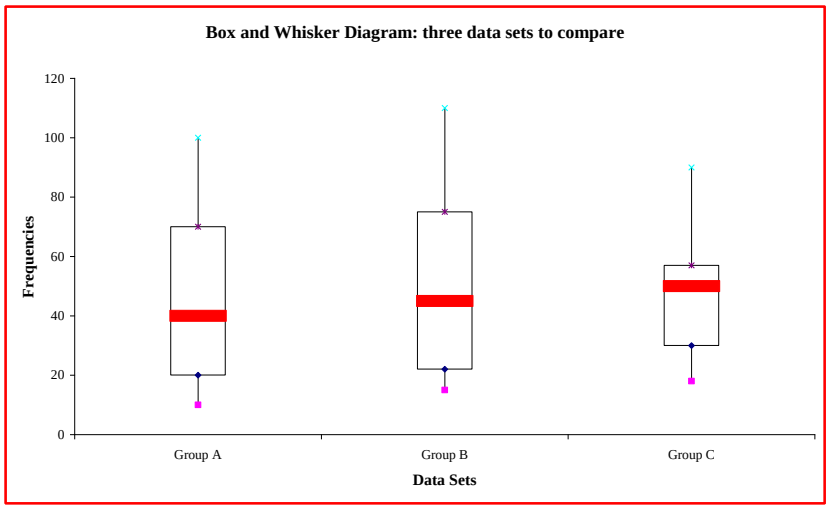
The essential feature of up-down bars is that they connect the first and last series: hence the rather strange ordering of the statistics in the table!

This is what you should have:



We can see immediately how the three data sets are different in terms of the maxima, minima and so on. Group 3's dispersion is the least of all, as we can see from the much smaller box it has when compared with the other two groups' boxes.

What is missing from the box plot we have just seen is the median line: here is a revised version of that box plot with the median lines added.



This is how we added the median lines:

right click on the median data point on any of the three data sets and select the Options tab and then increase the Gap Width to 200

now click on the Patterns tab and select the Custom option for the Marker: choose the cross if it's available in the Style drop down box and 72 pts for the Size

now click Finish and see what you have.

The problem with the choosing the cross as the marker is that it gives a horizontal and vertical line that makes the boxplot look a bit unusual, well, it's not necessarily the end of the world and you can choose the hyphen like I did for the boxplot above.

We have to say that since each chart will be different, the appearance of what you now have will vary so you might need to choose a Gap Width of more than or less than 200 until you are happy with your boxplot. Similarly, sometimes the cross is inexplicably unavailable … chose the Dash, not the minus sign: the minus sign is too awkward, believe it or not!

Your Turn

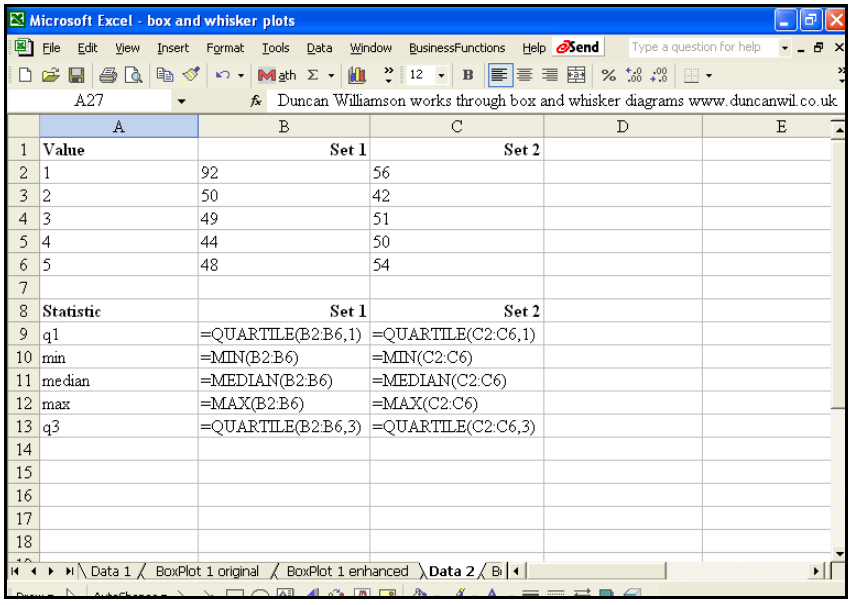
Here is one of tables of data from our standard deviation page, prepare a boxplot of the two series and comment on your findings:

Value	Set 1	Set 2
1	92	56
2	50	42
3	49	51
4	44	50
5	48	54

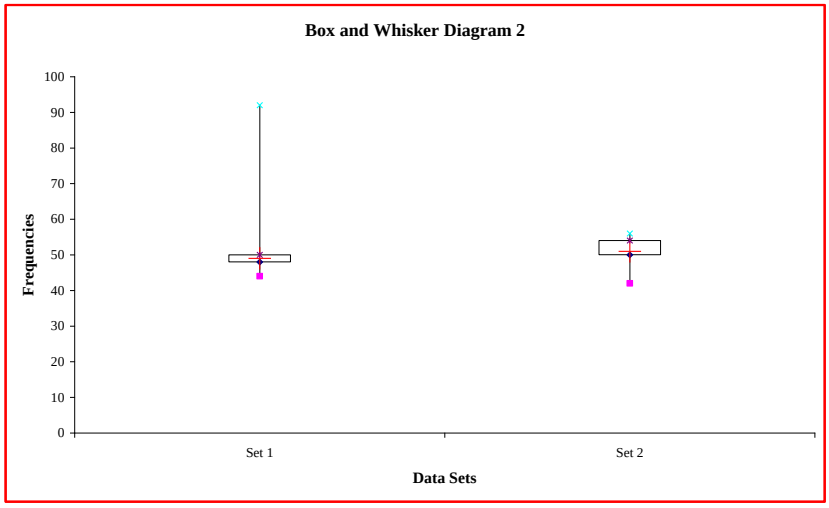
Did you get this?

Here are the variables we need and how Excel calculates the numbers we need for quartiles, minima, maxima and so on:

Statistic	Set 1	Set 2
q1	48	50
min	44	42
median	49	51
max	92	56
q3	50	54
standard deviation	19.92	5.37
inter quartile range	2.00	4.00
coefficient of variation	35.19%	10.61%



	A	B	C	D	E
1	**Value**	**Set 1**	**Set 2**		
2	1	92	56		
3	2	50	42		
4	3	49	51		
5	4	44	50		
6	5	48	54		
7					
8	**Statistic**	**Set 1**	**Set 2**		
9	q1	=QUARTILE(B2:B6,1)	=QUARTILE(C2:C6,1)		
10	min	=MIN(B2:B6)	=MIN(C2:C6)		
11	median	=MEDIAN(B2:B6)	=MEDIAN(C2:C6)		
12	max	=MAX(B2:B6)	=MAX(C2:C6)		
13	q3	=QUARTILE(B2:B6,3)	=QUARTILE(C2:C6,3)		
14					
15					
16					
17					
18					



Can you see that we used the Cross as the marker for the median but we had to restrict its size to 20 points: this is because the upper whisker is very small and the

cross would over write it and make a mess of it otherwise. We made the Gap Width 500 for this graph, too: even so, we did not get the median line to fill the whole box!

As we can see from this boxplot, we have two contrasting data sets: set 1 has a very high maximum value … exactly as we saw on the standard deviation page.

The added advantage of the boxplot is that it gives us the inter quartile range: Q3 - Q1 which is often a good indicator of the dispersion of a data set rather than the standard deviation. The standard deviations for sets 1 and 2 respectively are 19.92 and 5.37 respectively yet their inter quartile ranges are 2 and 4 respectively.

The plot also illuminates the shape of the distribution. The location of the median line and the relative length of the whiskers help indicate how symmetrical the data are. When the median lies far from the centre of the box or if one whisker is much longer than the other, you know that the distribution is skewed to some extent.

The results we now have tell us that although set 1 has a higher standard deviation, that value has been unduly influenced by an extreme value or outlier. We can see that set 1 does have an extreme value, 92, and this has provided us with an unfair view of the data set when we just take the mean and standard deviation into account when analysing the sets.

Further Questions

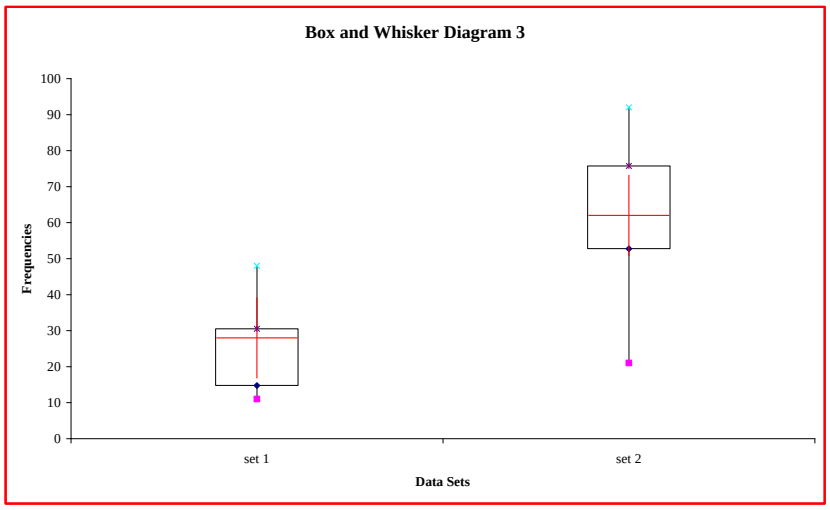
Both sets of data in this section are taken from the standard deviation page to be found on this site: draw a boxplot in both cases. We have added the standard deviation and inter quartile range and coefficient of variation values for you … now isn't that just kindness itself?

Value	**set 1**	**set 2**
1	20	21
2	31	57
3	29	67
4	48	88
5	47	69
6	11	52
7	29	55
8	27	92
9	13	45
10	13	78

Did you get this?

Statistic	**set 1**	**set 2**
q1	14.75	52.75
min	11.00	21.00
median	28.00	62.00
max	48.00	92.00

q3	30.50	75.75
standard deviation	3.028	13.172
inter quartile range	15.75	23.00
coefficient of variation	11.30%	21.11%



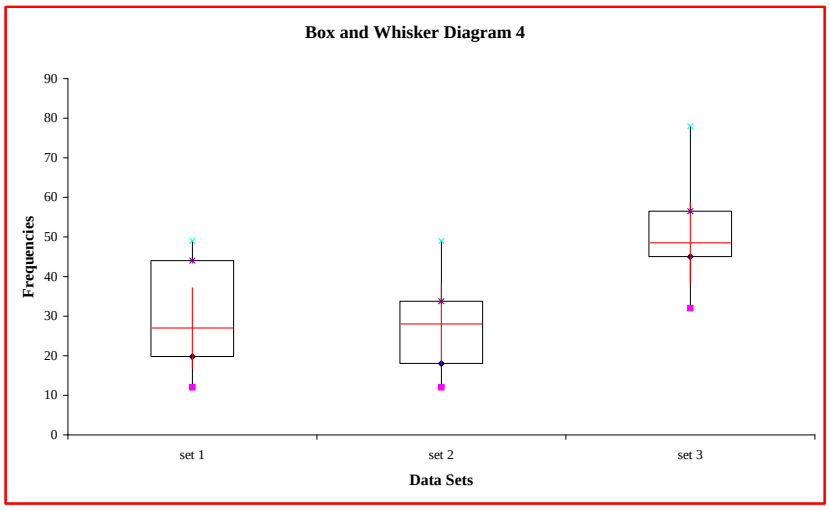
The box and whisker diagram shows us quite clearly, virtually without any effort, what it takes a table filled with 8 variables for each data set confirm: in this case data set 1 is the least disperse data set, it has by far the lowest median standard deviation, inter quartile range and coefficient of variation.

Value	**set 1**	**set 2**	**set 3**
1	47	49	68
2	12	29	48
3	49	12	58
4	19	44	49
5	31	15	48
6	17	17	52
7	23	21	78
8	47	33	44
9	35	34	32
10	22	27	33

Did you get this?

Statistic	**set 1**	**set 2**	**set 3**
Q1	19.75	18.00	45.00
min	12.00	12.00	32.00
median	27.00	28.00	48.50
max	49.00	49.00	78.00

Q3	44.00	33.75	56.50
standard deviation	13.71	12.27	14.24
inter quartile range	24.25	15.75	11.50
coefficient of variation	45.40%	43.66%	27.91%



Conclusions

Box and whisker diagrams are a very useful addition to the statistician's armoury: they provide an instant insight into the dispersion of data in a data set; and they are brilliant at helping us to compare two or more data sets. What's more, preparing box and whisker diagrams in Excel is really easy, once we've tried them a couple of times or so!

17 October 2002 & 26 July 2004

Appendix: Boxplots in Excel 5/95:

Please note*: we have not worked through this method for Excel 95 sowe present it here on an as is basis: we assume it works well enough, though, and recommend you check this with anyone who has access to Excel 95 if you have any difficulties with it. We'll help if we can, of course.*

Highlight the whole table, including figures and series labels.

Use Chart-Wizard - Line - Option 7 - Data in Rows - Finish to produce something like the chart below. Option 7 plots all the series as symbols without connecting lines, but also includes high-low lines which connect the maximum and minimum points for each group.

Now activate the chart and select Format - Chart Type - Options - Options - Up-Down Bars - OK

The outcome should be a set of boxplots, as we have seen already, above.

References

- Banks Tony & Alcorn David (2001) *Mathematics for AQA GCSE Higher Tier* Causeway Press Ltd

- McClave James T & Benson P George (1995) *A First Course In Business Statistics* Prentice Hall International

- Les Oakshott (1994) *Essential Elements of Business Statistics* DP Publications Ltd

Neville Hunt provides the following references that might be useful, although I haven't used them:

- Daly, F, Hand, D J, Jones, M C, Lunn A D and McConway, K J (1995) *Elements of Statistics* Addison Wesley/The Open University
- Devore, J and Peck, R (1990) *Introductory Statistics* West Publishing Co