Quantitative Analysis
Quantitative Analysis for Business

Summarising Data

Part of the reason why we analyse data is to see patterns. It is difficult to see patterns in data without summarising the data in some way. The most common way to summarise data is to convert the data into a summary table, into a graph or picture, or to use summary measures like the average. The benefit of this type of summary is that it gives us an instant picture of what is going on in our data set. The problem is that we often lose the detail of the original data.

Frequency Distributions

Suppose you have collected some data on the number of children in people's families. The variable we are measuring is 'number of children in the family'. The values that this variable can take are numbers like 0, 1, 2 or 3. These are discrete data, in that they can only be measured as whole values. You can't measure children in ever more accurate values like you can time or distance. When you can measure data in continuously more accurate measures (providing you have an appropriate measuring instrument), we call this type of data - continuous data.

Start by turning your data set on the variable, number of children, into a frequency distribution. A frequency distribution is a table that shows the values that a variable can take on the left hand column and then the frequency with which we observe the values in the right hand column. For example, it might look like this.

Frequency Distribution Table

Values of Variable	Frequency
Number of Children	Number of families
0	11
1	12
2	13
3	6
4	3
5	1
6	1

Notice that the right hand column is the frequency with which we find different values of our variable. We found 11 families with no children, 12 families with 1 child, etc. In all we have data from 47 families. You can see the pattern already. Most families have between 0 and 2 children. Having more than 3 children is quite rare. Note that we can still construct the original data from this table. There would be 11 zeros, 12 ones, 13 twos, etc.

Suppose you have a much larger data set with hundreds of values. If you created a table like the one above, it would go on for pages, and you wouldn't see any patterns. So in these cases, we group the values together to give a shorter table. Look at this table of the value of orders received in a company over a sample of 40 orders. It is called a grouped frequency distribution table.

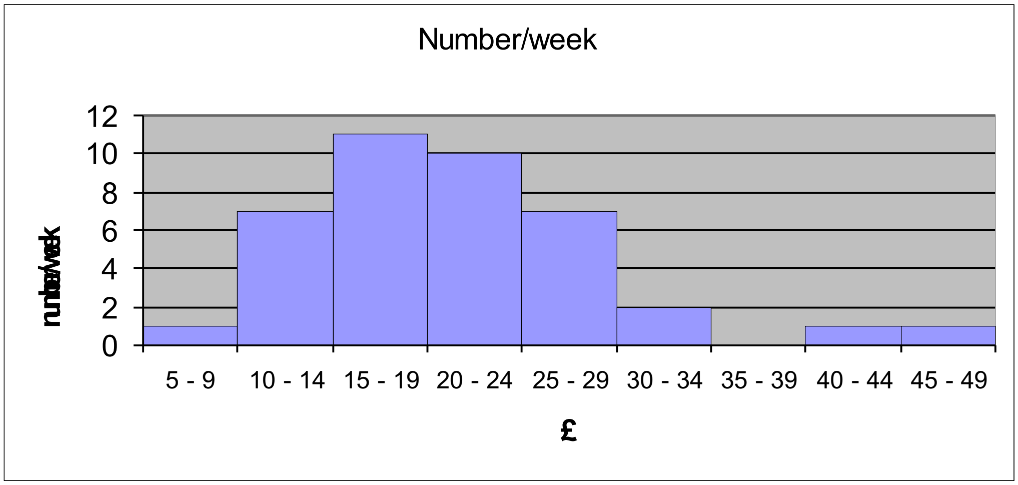
Grouped Frequency Distribution Table
of Value of Orders Received (£)

Values of Variable	Frequency
Value of Order	No. of
£	orders/week
5 < 10	1
10 < 15	7
15 < 20	11
20 < 25	10
25 < 30	7
30 < 35	2
35 < 40	0
40 < 45	1
45 < 50	1

This table shows us that out of a 40 observations (add up the frequency column), most of our orders have a value between £10 and £30. It is rare to get orders with values greater than £30. You can see that there were 7 weeks in which we observed orders with values between £10 and just less than £15, but we cannot tell from this table what the original value of those 7 orders were. We have lost the original data.

Note the groups of values in the left hand column do not overlap. The first group is £5 but less than (<) £10. So £10 goes in the next group. There is no ambiguity here. We know into which group every value in the data set should go. You must not have overlapping groups. If we had £5-£10 and then £10-£15, we wouldn't know whether to put an order worth £10 into the first group or the second group.

Histograms
A histogram is a picture of a grouped frequency distribution. Usually we only draw histograms for continuous data. The picture below shows a histogram of our data. It looks like a bar chart, but it is not a bar chart. Bar charts only have a numerical scale on one axis, the other axis has some sort of category. A histogram has a scale on both axes. The area of the bar in a histogram represents the frequency. If the bar widths are all the same, then the height of each bar is plotted at the frequency for group and you can read the histogram like a bar chart.



You can see that I have drawn this like a bar chart with the values beneath each bar. The height of each bar is set at the value of the frequency. So the first bar tells us that there was only one observations of an order with a value between £5 and £9 inclusive. Really, the scale should go from 0 to 50, but it is difficult to achieve this in Excel. The bar line should fall on a value that is not in the data, so the bar line between the first and second bar would be at the value of £9.5 which is not in the data as all the data have been rounded to whole pounds. However, it is usual now to put the bar lines at the lowest value of the right hand bar. So the left hand bar line of the first bar would be at the value £5, the left hand bar line of the second bar would be at the value of £10, etc.

Problems with histograms.
If the widths of the bars (the groups in your frequency distribution table) are not all the same, you cannot read the histogram like a bar chart. You cannot plot all the heights of every bar at the frequency because it is really the area that represents the frequency and not the height. Look at the example below on the value of orders.

Value of Variable Value of Orders £	Frequency Number of orders /week	Width	Plot height	
0 < 5	1	£5	2	(x 2)
5 < 10	3	£5	6	(x 2)
10 < 20	11	£10	11	
20 < 30	25	£10	25	
30 < 40	41	£10	41	
40 < 50	36	£10	36	
50 < 60	16	£10	16	
60 < 90	15	£30	5	(x 1/3)
90 < 150	9	£60	1.5	(x 1/6)

The first 2 bars are just £5 wide. The middle bars are £10 wide, the penultimate bar is £30 wide and the final bar is £60 wide. If we plotted all the bar heights at the given frequency, then some bars would look dominating and out of proportion. Just think about the areas. If one bar has its height plotted at its frequency and the bar next to it is twice as wide, to keep the areas consistent, the height of the bar that is twice as wide should be plotted at half its frequency. If you double the width of a bar, you must halve its height to keep the area the same. If you halve the width of a bar, you must double its height to keep the area the same. So you can now see why the height of the bar has been adjusted. You need to decide on one set of bars to plot with the height equal to the frequency. Choose the width that occurs most frequently to save time on adjusting other bars. So we plot the height of the middle bars with a width of £10 at their frequency values. All the other bars must be adjusted. For example, the first bar is half the width of the middle bars, so we must double its height (multiply the frequency by 2). The last bar is 6 times wider than the middle bars and so we must divide its height by 6 (or multiply by 1/6). Try drawing this one for yourself.

Ogives or Cumulative Frequency Distributions

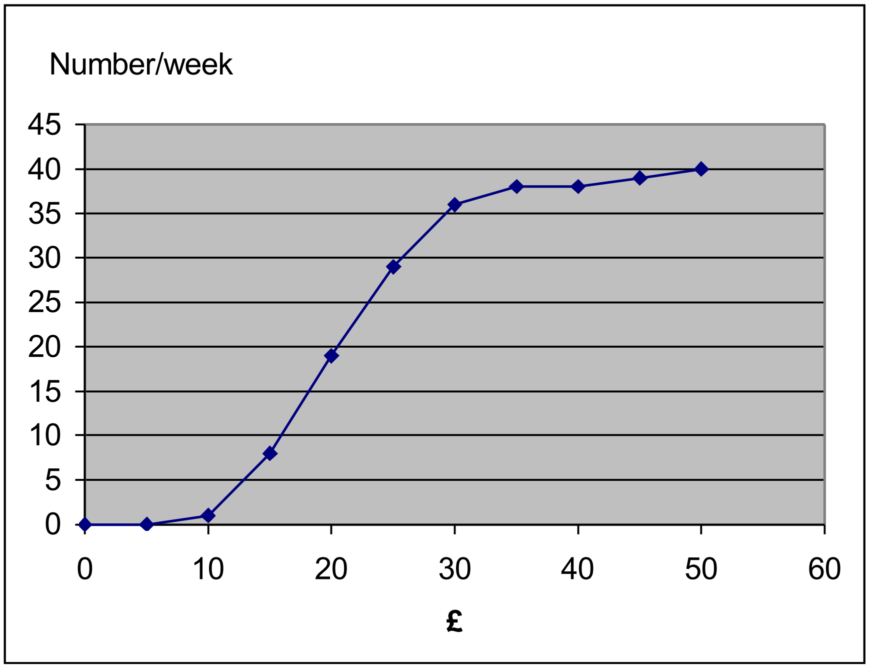
Lets go back to our first frequency distribution of the value of orders. Suppose we now want to ask different questions, like how many orders had values less than £20. We could look at the original distribution below and add up the all the values less than £20. There is 1 less than £10, 7 between £10 less than £15, and 11 of £15 but less than £20. So that’s 19 in all. If we add up all the values in this cumulative way, we create a cumulative distribution.

Frequency Distribution		Cumulative Distribution	
Value of Variable	Frequency	Cumulative frequency	
Value of Order	No. of	No. of orders/week	%
£	orders/week		
5 < 10	1	1	3
10 < 15	7	8	20
15 < 20	11	19	48
20 < 25	10	29	73
25 < 30	7	36	90
30 < 35	2	38	95
35 < 40	0	38	95
40 < 45	1	39	98
45 < 50	1	40	100

If you plot the cumulative distribution as a graph it is called an ‘ogive’. You must put a proper scale on each axes. Plot the cumulative frequency on the vertical axis and plot it against the upper group value of each group in question. You can plot the actual cumulative values or you can plot these values as percentages. So in the case of the first point to plot, you would plot 1 (or 3%) on the vertical axis against £10 on the horizontal axis, then 8 (or 20%) against £15, etc.

You can see the ogive below.

You can use the ogive to answer questions like how many orders do we find below a value of £32. You can’t tell this from the table because £32 is in between two points, but you can read along to £32 on the horizontal axis on the ogive, then read up to the line and read across to the vertical axis to find out how many. You are estimating the number from your graph because we no longer have the original data to find out precisely. Now answer this question. 50% of our orders are more than what value? We have 40 orders in total, so 50% is 20 orders. Read up to 20 orders per week on the vertical axis and then read along and down. It appears to be that half our orders are more than £21.



Measures of Location

Instead of giving visual pictures or tables of our data, we can quote single numbers that say a lot about the data. We all know the arithmetic mean, more commonly known as the average. It gives us an idea about the central value of the data. But it is not the only measure of the centre. We can also use the median and the mode.

The Arithmetic Mean

For Raw Data
When you have all the original data, you put the data into Excel, go to the formula bar and find the mean. It will calculate it for you. The formula if you do it with a calculator is:

Mean for raw datra $\Sigma x/ n$

where Σ means sum, x are all the values of your variable, and n is the number of values you have.

So if you have the age of 10 members of staff,
43, 75, 50, 51, 51, 47, 50, 47, 40, and 48
n = 10, Σx = 502,
Mean age = 502/10 = 50.2 years (not 50.2)

But what happens if you do not have all the original data and you only have a grouped frequency distribution created by someone else. Maybe you have got the data from a report or a government web page as a table. You will have to estimate the value of the mean as you can't get all the original data and put it into Excel.

Value of Variable	Frequency (f)	(x)	(fx)
Value of Order	No. of	mid-point	
£	orders/week	£	
5 < 10	1	7	7
10 < 15	7	12	84
15 < 20	11	17	187
20 < 25	10	22	220
25 < 30	7	27	189
30 < 35	2	32	64
35 < 40	0	37	0
40 < 45	1	42	42
45 < 50	1	47	47
	40		840

So let's look at our data on the value of orders. We know we have 1 value between £5 and less than £10, but we don't know what it was. We don't know any of the values in the table accurately. So to estimate the mean, we have to guess these values. The standard way is to assume that the values we have are equally distributed within each group (or class) and so on average they take the value in the middle of each group. So our 1 value lies at a value between £5 and £9, say £7, and our 7 values all take a value in the middle of £10 to £14, say £12. So you need to find the mid-point for each group as shown above. Now we can add up all our values. We have 1 at £7, 7 at £12, 11 at £17, etc. This requires multiplying all the frequencies (f) by their mid-points (x) in the column shown as fx. Add these up to get the sum of all the values and then divide by how many values we have, 40.

Mean for grouped data $\Sigma fx/\Sigma f = \Sigma fx/n$

Mean order value = £840/40 = £21.0

So the mean tells us that the average value of our orders is £21. Is this a good measure of the centre of our data? Well it probably will be providing that we don't have too many out of the ordinary values in our data. For example, imagine that we got one value worth £1,000. This is called an 'outlier' – an extreme value that is not typical in our data set. Adding this into our data would really bring the average value up. It might come up so high that it no longer represented what was typical any more. The average is dragged up by high outliers, and dragged down by low outliers. One way to get round this problem is to use another measure of centre of the distribution, called the median.

The Median

The median is simply the value of the middle observation in our data, when the data are ordered from smallest to largest.
Let's look at our data on the age of 10 members of staff.

43, 75, 50, 51, 51, 47, 50 , 47, 40, and 48

Now order them from smallest to largest

40, 43, 47, 47, 48, 50, 50, 51, 51, and 75

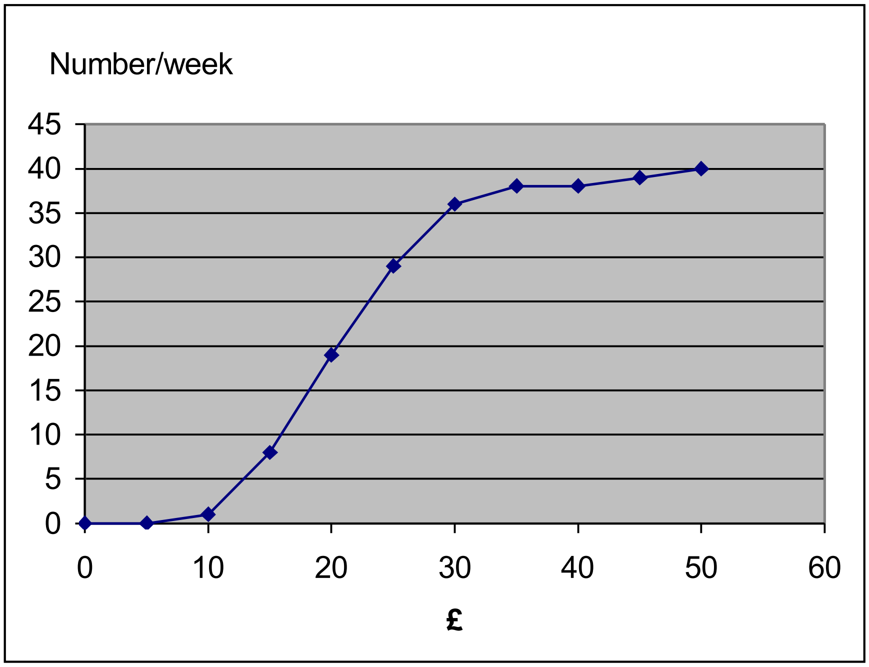
As we have an even number, there isn't a middle observation, so we assume its between observation 5 and 6 and average these two values. Observation 5 is 48 and observation 6 is 50, so the average of these two is 49 years. We take this as the median value.

Compare the median value of 49 years to the mean value of 50.2 years. The mean has been dragged up by the one high value of 75 years. The mean at overstates the typical age for a lecturer. Whilst the difference does not seem overwhelming here, imagine if these values were in millions. Then it would make a difference!

So the median value gives us a good impression of the typical value. However, in some ways it is not very representative as it is only 1 value out of many. It is a good idea to quote both the mean and the median if you think you have outliers in your data.

The median for grouped data

If you have grouped data, you have the same problem estimating the median, as you no longer have the original data. So again you must estimate. You can do this using an ogive. Draw an ogive from your frequency distribution and go half way into your data and read of the value of the observation half way into the data. Look at the ogive we drew above for our order data.



We have 40 observations and so half way into the data is the 20th observation. We have an even number of observations, but at this level of accuracy it's a bit pointless saying we want the value of the 20.5th observation as it won't make much difference.

So read up to 20 and read across and down. The value is about £21. Compare this to the mean we calculated at £21. They are the same.

You can also use a calculation instead. You don't need this for the exam or the assessments but you need it for a class question.

Median = L + [(n/2 – fm-1)/fm] C
L = lowest value in the median group or class
n = number of observations
fm-1 = cumulative frequency in the group (class) before the median
fm = frequency of the median group (class)
C = width of the median group (class)

So for us n/2 = 40/2 = 20 th observation. Look at the cumulative frequency below and you can see that the 20th observation (the median) comes just into the 4th group (class) which is £20 < £25. This is because the 3rd group (£15 < £20) takes us to the 19 th observation. So L = £20. fm-1 is the cumulative frequency in the class before the median class. This class is £15 < £20 and the cumulative frequency of this class is 19. So fm-1 = 19. fm is the frequency (not the cumulative frequency) of the median class which is 10, so fm = 10. C = £5 from £20 to £25 approximately.

Median = L + [(n/2 – fm-1)/fm] C

Median = £20 + [(20 – 19)/10]£5 = £20 + [1/10]£5 = £20 + £0.5 = £20.50
So this is more accurate in the sense that it is calculated and my estimate from my ogive was crude as my I didn't draw it carefully on graph paper. If I had a more accurate graph, I would get a similar estimate of the median from the graph. They are both estimates.

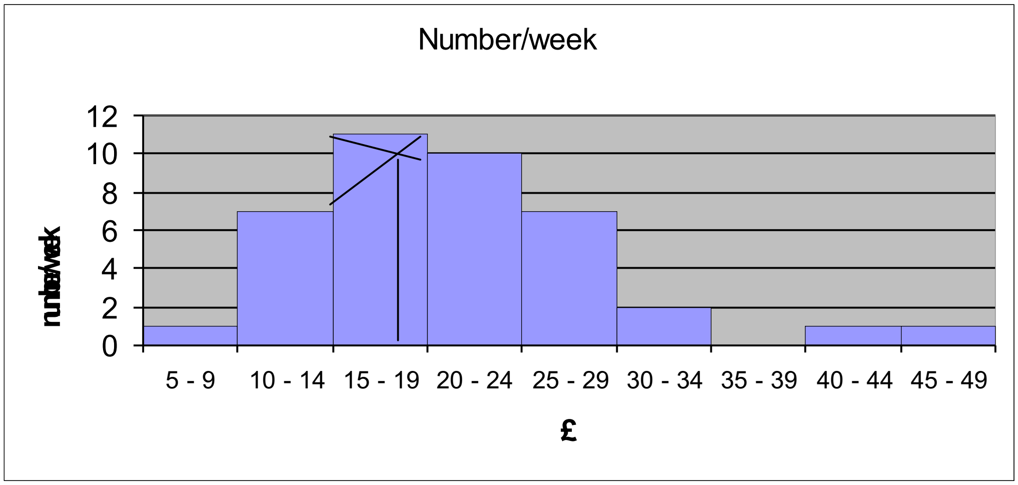
Frequency Distribution		Cumulative Distribution	
Value of Variable	Frequency	Cumulative frequency	
Value of Order	No. of	No. of orders/week	%
£	orders/week		
5 < 10	1	1	3
10 < 15	7	8	20
15 < 20	11	19	48
20 < 25	10	29	73
25 < 30	7	36	90
30 < 35	2	38	95
35 < 40	0	38	95
40 < 45	1	39	98
45 < 50	1	40	100

The Mode

The mode is simply the most frequently occurring value in the data set. Its not used that often, but you can imagine it being used in marketing reports. The most frequently purchased brand is x!

To find the mode in raw data, simply sort your data in Excel and look for the most frequently occurring value.

To find the mode from a grouped frequency distribution you need to estimate the value. Draw a histogram and find the heighest bar. This is the modal bar, but we need one value, not a group of values. So to find the value you draw diagonal lines from the top corners of the heighest bar to the adjacent bar. At the point where the diagonal lines cross, read down to the horizontal axis to estimate the mode. You can see its about £19.



Alternatively you can use a calculation method. Again you don’t need this for the exam or the assessments but you will need it for practice in the class questions.

Mode = L + [D1/(D1 + D2)] C
L is the lowest value in the modal group (class). The modal class is £15 < £20, so L = £15. D1 is the difference between the frequency of the modal group (class) and the frequency of the group(class) before the modal group. D1 = 11 -7 = 4
D2 is the difference between the frequency of the modal group (class) and the frequency of the group (class) after the modal group. D2 = 11 –10 = 1
C is the width of the modal group (class). C = £5.

Mode = L + [D1/(D1 + D2)] C
Mode = £15 + [4/(4 + 1)]£5 = £15 + £4 = £19

When to use the mean and the median

If I were to tell you that the average score on last year’s exam was 50% and the pass mark was 40%, would you feel happy to take this module? Should you switch to another module with a better chance of passing?

The answer depends on whether you are an optimist or a pessimist. The optimists among you will believe that the typical grade was over 50% but a few people really messed up and got zero, dragging the average below the typical value. The pessimists among you will believe that the typical grade was less than 50% but a few swots got 100% dragging the average above what is typical.
Now if I tell you that the median mark was 55%, you now know that the optimists were right. But if I tell you that the median mark was 45%, you know the pessimists were right.

What I am trying to tell you is that you can't really make decisions based on the mean value unless you know something about the median. If the mean and the median are very similar, then the mean tells you what is typical. If the mean and the median are very different, the mean may mislead you unless you have knowledge of the median.

So, when you next read something where the mean is quoted, or someone tells you what the average value is, ask for the median and explain why you need it.