

Measures of Centre and Spread: Ungrouped data

Different averages are used for different sets of data according to whether the data is nominal, ordinal or quantitative.

	nominal data	ordinal data	Quantitative data
characteristics	labels for categories	underlying sequence and order	has sequence and units of measurement. data can be discrete (integers) or continuous.
some examples	male/female occupation favourite sport	rankings ratings	age, salary, number of children, height, weight etc.
graphical summary	pie and bar charts	pie and bar charts	histogram for continuous data bar-chart for discrete data
tabular format	frequency counts by category	frequency counts by ratings/ rankings	frequency counts by numerical value or ranges of numerical values (classes)
summary statistics	percentages mode	percentages mode, median range, inter-quartile range	percentages mode, median and mean range, inter-quartile range and standard deviation

For nominal data, we can only use the mode

For ordinal data we can use the mode, median and inter-quartile range

For quantitative data, we can use the mode, median, mean, inter-quartile range and standard deviation.

A: Ungrouped data

Example: Kidburn Question 2: “How many times a month would you use the cafes/restaurants in the Market Square?”

When the descriptive statistics were generated from the ungrouped data in excel, the following table was produced:

Question 2	
<i>number of times cafes are used each month</i>	
Mean	6.4
Standard Error	0.6020374
Median	5
Mode	4
Standard Deviation	4.257047
Sample Variance	18.122449
Kurtosis	1.6097306
Skewness	1.2700854
Range	19
Minimum	1
Maximum	20
Sum	320
Count	50
lower quartile	4
upper quartile	8

How can we make sense of this?

Averages

Averages represent ways of measuring the **centre** of a set of data; the most commonly used averages are the **mean, median and mode**

The Mode

The mode of a set of data is the value that occurs most frequently.

Example: The mode of the data set:

2	3	5	5	7	7	7	8	9
---	---	---	---	---	---	---	---	---

is the value 7, as 7 occurs most frequently.

It is possible for a data set to have more than one mode or to have no mode at all.

The mode can also be found by determining which value has the highest frequency from the frequency table:

Example

Value, x	3	4	5 **	6	7
Frequency, f	1	2	3	2	1

The value with the highest frequency is 5, thus 5 is the mode.

For the Kidburn data , we can say, "The most common number of times that residents would use the cafes in the Market Square is 4 times a month."

The Median

The median of a data set is the value that divides the data set into 2 halves. The values must be in numerical order before the 'middle' value can be found.

As nominal data has no underlying order, the median cannot be used for nominal data.

The data set:

13	15	17	15	18	19	17	17	12
----	----	----	----	----	----	----	----	----

can be ordered to give

12	13	15	15	17 **	17	17	18	19
----	----	----	----	-------	----	----	----	----

The middle value is 17, thus the median is 17.

The formula $\frac{n+1}{2}$ Gives the **position** of the median in an ordered set.

Example: The data set has 13 items, thus $n = 13$.

11	11	12	13	14	15	15 **	17	18	19	19	20	20
----	----	----	----	----	----	-------	----	----	----	----	----	----

The position of the median is $\frac{n+1}{2} = \frac{13+1}{2} = \frac{14}{2} = 7$ th item

The median can be found at the 7th position; the median is 15.

If a data set has an even number of values then the median will be half way between the two middle values.

12, 13, 14, **	15, 16, 18
----------------	------------

$n = 6$

The position of the median is $\frac{n+1}{2} = \frac{6+1}{2} = \frac{7}{2} = 3.5$

The median can now be found midway between the 3rd and the 4th values.

$$\text{Median} = \frac{14+15}{2} = \frac{29}{2} = 14.5$$

For the Kidburn data, we can say, "The median number of times that residents would use the cafes in the Market Square is 5 times a month."

The Mean

Notation: The symbol $\bar{x}$ is used to denote the mean of a sample [x bar]
 The symbol Σ is used to mean 'the sum of' [sigma]
 The symbol n is used to denote the number of values / sample size

$$\text{Mean} = \frac{\text{sum of all the values}}{\text{number of values}} = \frac{\Sigma x}{n}$$

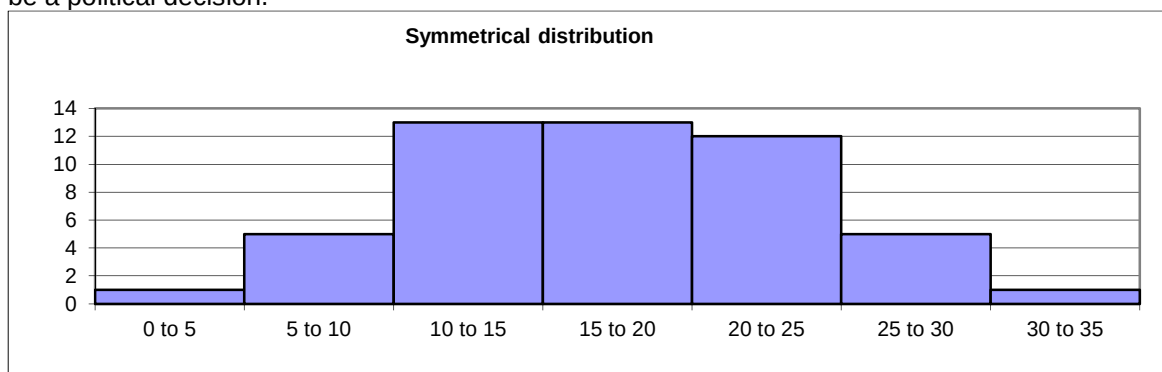
Example: The data set 3, 6, 7, 8, 9, 3, 5, 4 has 8 values, therefore $n = 8$

$$\text{Mean, } \bar{x} = \frac{3 + 6 + 7 + 8 + 9 + 3 + 5 + 4}{8} = \frac{45}{8} = 5.625$$

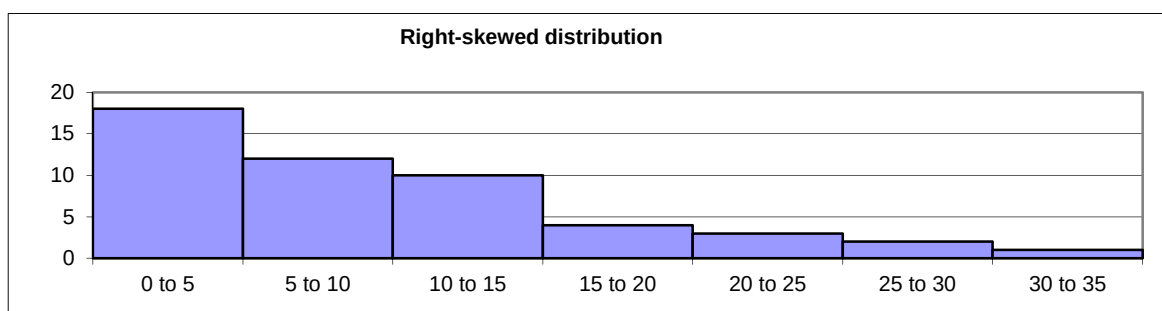
For the Kidburn data, we can say, "The mean number of times that residents would use the cafes in the Market Square is 6.4 times a month".

Comparison of the mean and the median.

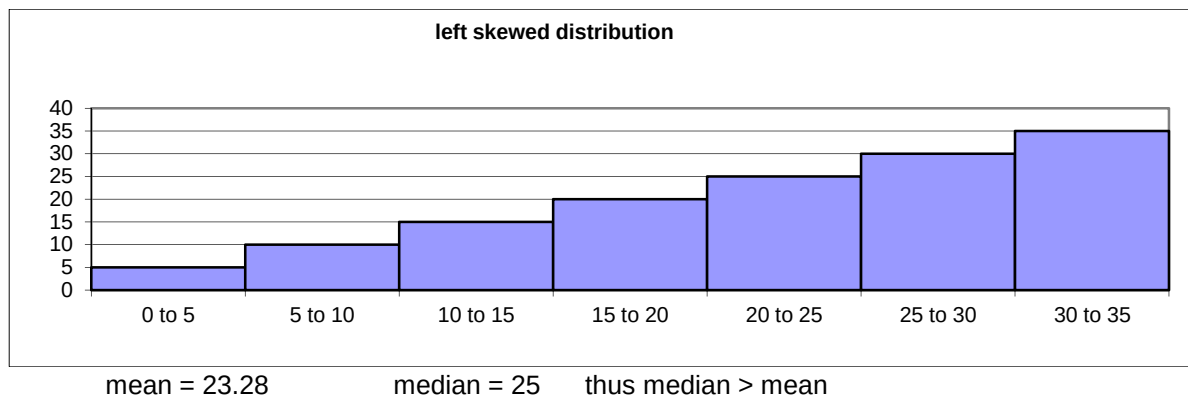
If the distribution of a data set is symmetrical, the mean and the median will be close in value.
 If the distribution is skewed this will not be the case and the choice of which 'average' to use can often be a political decision.



mean = median = 16

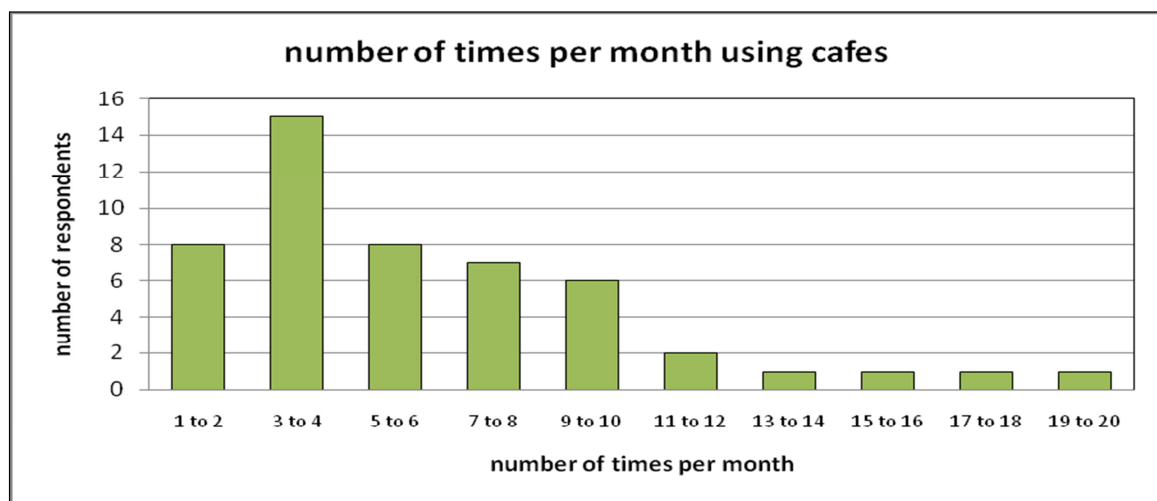


mean = 9.72 median = 8 thus mean > median



	Symmetrical distribution	Right skewed or positively skewed	Left skewed or negatively skewed
Mean	16	9.72	23.28
Median	16	8	25
Notes	Mean = median [or close in value]	Median < mean	Median > mean

For the Kidburn data, we can say, "The mean number of times a month that residents would use the cafes in the Market Square is 6.4 times a month, whilst the median is 5 times a month. As the mean is larger than the median, this indicates that the data is positively skewed. This skewness is a result of a few residents stating that they would use the cafes quite often. This skewness can be seen in the graph ".



Tip : The value of "skewness" in the table of descriptive statistics is positive at 1.2700854., thus the distribution is "positively skewed" as the value of skewness is a positive number.

Spreadsheet Commands for Averages (Excel)

To find the mean of a block of data:	=AVERAGE(block)
To find the median of a block of data:	=MEDIAN(block)
To find the mode of a block of data:	=MODE(block)

A (block) is the section of the spreadsheet that contains the data. It can be selected by highlighting the "block".

Seminar Discussion Questions on Averages

Ten workers from the same company have the following weekly salaries

£112 £125 £115 £164 £177 £125 £128 £125 £135 £350

- Calculate the mean, mode and median salary among these 10 workers.
- Which single average would you consider to be the most representative of these salaries and why?

These workers receive a flat pay increase of £10 each.

- Explain why the three averages found in (a) will increase by the same amount of £10.

Ask your tutor to show you how to use your calculator for the mean

Measures of Spread

Measures of spread seek to measure how variable the data is; i.e. are all the values "close together" or are they very "spread out".

The Range The range of a set of data is: **range = largest value - smallest value**

Data set: 13, 14, 14, 15, 15, 16, 16, 17, 18, 19 19

has range $19 - 13 = 6$

The range can easily be distorted by one extreme value.

For the Kidburn data , the minimum is 1 and the maximum is 20, giving a value for the range of 19.

The Inter-Quartile Range

A data set can be divided into 4 quartiles.

25% of data		25% of data		25% of data		25% of data
Q ₁		Q ₂		Q ₃		
Q ₁ is called the lower quartile		it is found at position $(n+1)/4$				
Q ₂ is called the median		it is found at position $2(n+1)/4$				
Q ₃ is called the upper quartile		it is found at position $3(n+1)/4$				

The inter-quartile range is $Q_3 - Q_1$. It gives the range of the middle half of the data set.

Data set: 13, 14, 14, 15, 15, 16, 16 17, 18, 19, 19

The data set has 11 items, thus $n = 11$

Q₁ is a position $(11+1)/4 = 3\text{rd}$ Q₁ = 14 Q₃ is at position $3(n+1)/4 = 9\text{th}$ Q₃ = 18

Thus the inter-quartile range is $18 - 14 = 4$

For the Kidburn data is the lower quartile is 4 and the upper quartile is 8. Thus , we can say ,
 " The middle half of times that the residents expect to use the cafes is between 4 and 8 times a month" OR " 25% of the respondents use the cafes less than 4 times a month and 25% use the cafes more than 8 times a month."

The Standard Deviation

The formula for the standard deviation of a sample is:

$$s = \sqrt{\frac{\sum (x - \text{mean})^2}{n - 1}}$$

The standard deviation is the square root of the variance.

The value, s, is used as the best unbiased estimate of the population standard deviation and it is used when the population standard deviation is unknown as we only have access to the sample.

Example: Find the standard deviation from the sample data:

13, 14, 14, 15, 15, 16, 16, 17, 18, 19, 19
 $\sum x = 176$, $n = 11$ thus mean, $\bar{x} = 16$

Value, x	x - mean = x - 16	(x - mean) ²
13	-3	9
14	-2	4
14	-2	4
15	-1	1
15	-1	1
16	0	0
16	0	0
17	1	1
18	2	4
19	3	9
19	3	9
totals		42

$$s = \sqrt{\frac{\sum (x - \text{mean})^2}{n - 1}} = \sqrt{\frac{42}{10}} = \sqrt{4.2} = 2.0494$$

standard deviation from sample = 2.0494

Ask your tutor to show you how to use your calculator to find a standard deviation.

For the Kidburn data, the standard deviation is 4.257.

It is hard to put this in a meaningful sentence. The main use of the standard deviation is when comparing the spread of one set of data with another. As we only have one set of data here, there is not much we can say beyond quoting the statistic.

Coefficient of Variation

Often it is helpful to calculate the coefficient of variation. This is the value of the standard deviation divided by the mean (expressed as a percentage).

For the Kidburn data, this gives $4.257/6.4 = 0.665$. Thus the coefficient of variation is 66.5%, indicating the data has a fair degree of spread.

Spreadsheet Commands (Excel) :

To find the range of a data block:	=max(block)-min(block)
To find the lower quartile	=quartile(block, 1)
To find the upper quartile	=quartile(block, 3)
To find the sample standard deviation:	=stdev(block)

Seminar Discussion Questions for Spread

- Work out the mean and standard deviation of the following data samples by drawing up a table like the one above.
 - 16 14 17 21 19 19
 - 7 13 21 19 22 24 31 58
 - 17 23 31 29 32 34 41 68
 - What do you notice between the statistics of samples b) and c)?
- Use your calculator to work out the mean and standard deviation of the following data samples:
 - 12 14 13 17 19 19 12 16 15 17
 - 31.5 18.7 22.4 27.5 20.6

Reporting on Descriptive Statistics.

Report:

Number of times the residents of Kidburn would use the cafes and restaurants in the Market Square.

The most common number of times that respondents said they expected to use cafes is 3 or 4 times a month, with 30% of respondents giving this answer. A total of 60% of respondents indicated that they expected to use the cafes between 3 and 8 times a month. Only 12% of respondents thought that they would use the cafes more than 10 times a month. These results seem to suggest that most respondents would use the cafes about once or twice a week.

The mean number of times that residents would use the cafes in the Market Square is 6.4 times a month, whilst the median is 5 times a month. As the mean is larger than the median, this indicates that the data is positively skewed. This skewness is the result of a few residents stating that they would use the cafes quite often.

The minimum number of times that residents would use the cafes is once a month and the maximum is 20 times a month. Whilst this is a large range of answers, the middle half of the number of times residents expect to use the cafes is between 4 and 8 times a month. Thus most of the residents would use the cafes once or twice a week with a few residents using them more often.

Answers for averages

a)	Mean	Mode	Median
	£155.60	£125	£126.50

b) The median because the sample has one extreme value.

c) For the mean, $\sum f \times x$ will increase by 100. This increase divided by 10, will give an increase of 10. For the mode, the most frequent salary (i.e. £125) will simply increase by 10, since all salaries increase by 10. For the median, the middle salary will remain in the middle but will increase by £10.

Answers to questions on spread

- Mean = 17.6667 Standard deviation = 2.50333
 - Mean = 24.375 Standard deviation = 15.3617
 - Mean = 34.375 Standard deviation = 15.3617
 - The mean in c) is greater than the mean in b) by 10. The standard deviations are the same.
- Mean = 15.4 Standard deviation = 2.6331

b) Mean = 24.14

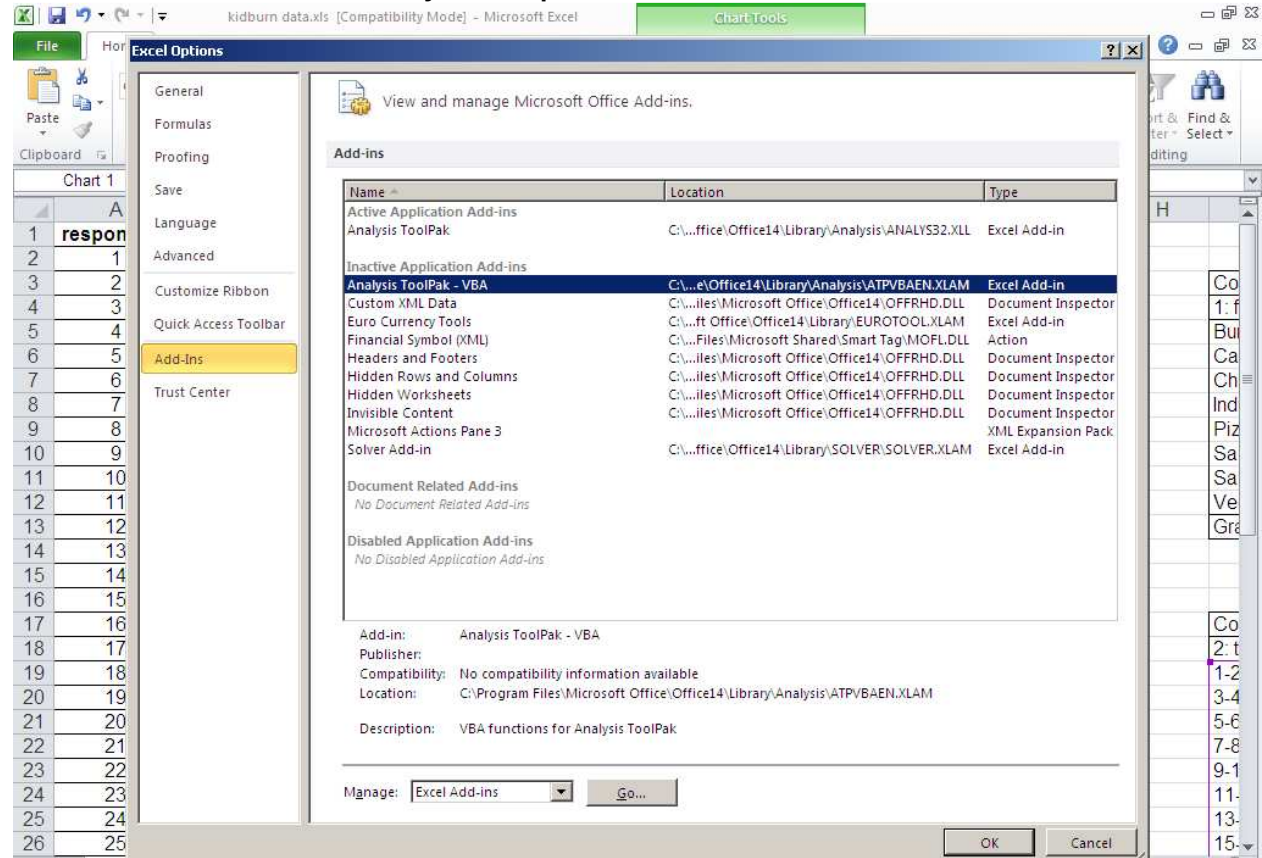
Standard deviation = 5.258611984

IT worksheet: using Descriptive Statistics in Excel for Ungrouped data

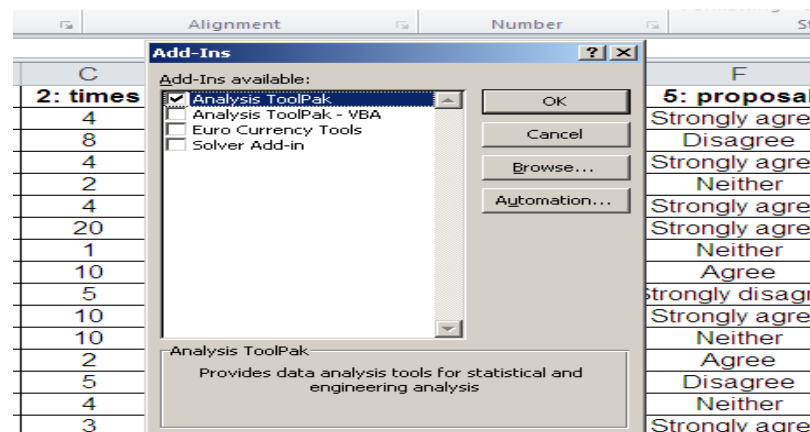
As an example , we will use Question 2 from the Kidburn data file.

From the main “file “ tab, choose **excel options**:

Then choose **add-ins** and **Analysis Toolpak** then **GO**



Tick the option **Analysis ToolPak** and **OK**



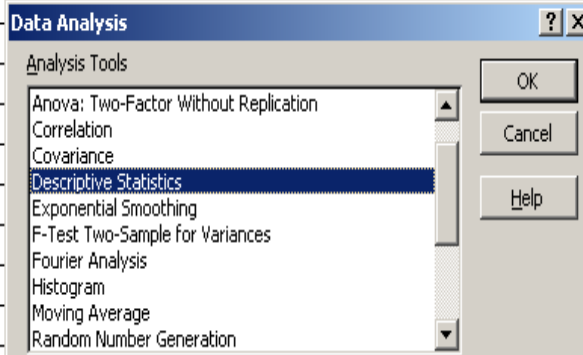
This has loaded up the software in order to use the Data Analysis facility .

From the **Data** tab,

choose **Data Analysis**

and then **Descriptive Statistics** and **OK**

	A	B	C	D	E	F
1	respondent	1: food	2: times	3: spend	4: shop	5: proposal
2	1	Cakes	4	under £1	Clothes:children	Strongly agree
3	2	Vegetarian	6	£2 up to £4	books:video/music	Disagree
4	3	Pizzas				Strongly agree
5	4	Chinese				Other
6	5	Burgers				Strongly agree
7	6	Sandwiches				Strongly agree
8	7	Indian				Other
9	8	Cakes				Strongly agree
10	9	Pizzas				Disagree
11	10	Vegetarian				Strongly agree
12	11	Sandwiches				Other

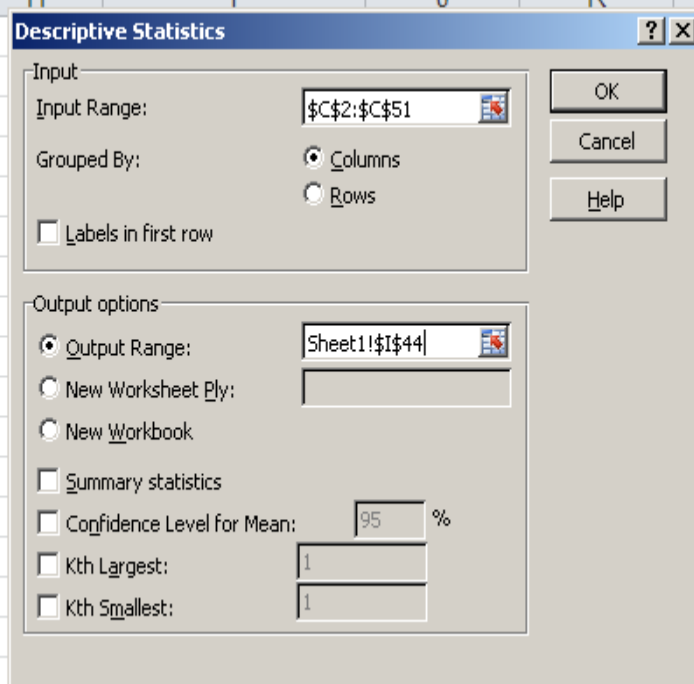


For the **Input range** highlight the values under the heading “2:times”

For the **output range** : click on a blank cell where you want the answers to appear.

Tick the **Summary statistics** box.

	G	H	I	J	K
	Specialist Food				
	Children's Activities				
ee	Specialist Food				
	Live Music				
	Children's Activities				
	Street Performers				
	Specialist Food				
ee	Children's Activities				
ee	Live Music				
	Live Music				
	Specialist Food				
ee	Specialist Food				
ee	Children's Activities				
	Craft Stalls				
ee	Art Exhibitions				
	Children's Activities				
	Children's Activities				
ee	Children's Activities				



The table of statistics will now appear:

Column1	
Mean	6.4
Standard Error	0.6
Median	5
Mode	4

Standard Deviation	4.26
Sample Variance	18.1
Kurtosis	1.61
Skewness	1.27
Range	19
Minimum	1
Maximum	20
Sum	320
Count	50

We can add the values of the quartiles by using the commands “ =quartile(block,1) “ and =quartile(block,3)