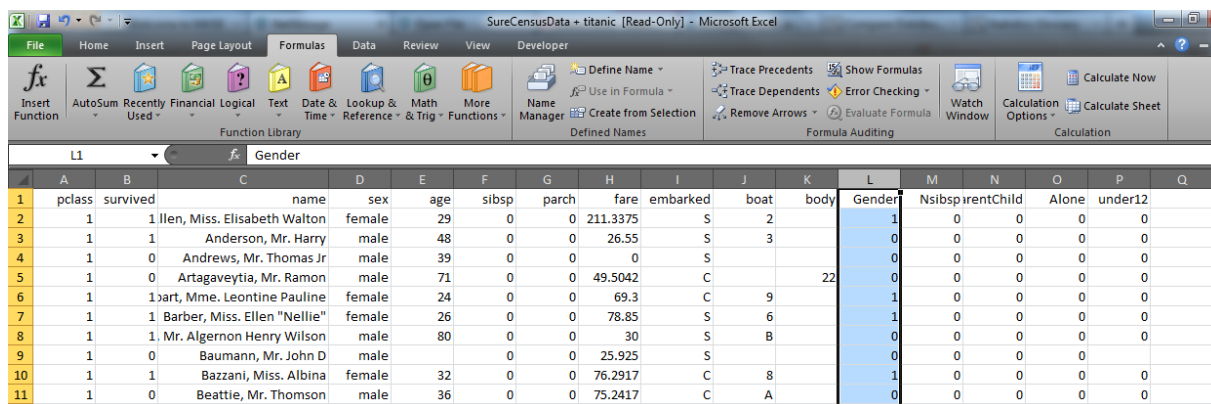


Data

After collection of data, most people will enter their data into a spreadsheet using the following layout:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	pclass	survived	name	sex	age	sibsp	parch	fare	embarked	boat	body	Gender	Nsibsp	parentChild	Alone	under12	
1	1	1	Ellen, Miss. Elisabeth Walton	female	29	0	0	211.3375	S	2		1	0	0	0	0	
2	1	1	Anderson, Mr. Harry	male	48	0	0	26.55	S	3		0	0	0	0	0	
3	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	0	S			0	0	0	0	0	
4	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	49.5042	C		22	0	0	0	0	0	
5	1	0	Barber, Mme. Leontine Pauline	female	24	0	0	69.3	C	9		1	0	0	0	0	
6	1	1	Barber, Miss. Ellen "Nellie"	female	26	0	0	78.85	S	6		1	0	0	0	0	
7	1	1	Mr. Algernon Henry Wilson	male	80	0	0	30	S	8		0	0	0	0	0	
8	1	0	Baumann, Mr. John D	male		0	0	25.925	S			0	0	0	0	0	
9	1	1	Bazzani, Miss. Albina	female	32	0	0	76.2917	C	8		1	0	0	0	0	
10	1	0	Beattie, Mr. Thomson	male	36	0	0	75.2417	C	A		0	0	0	0	0	

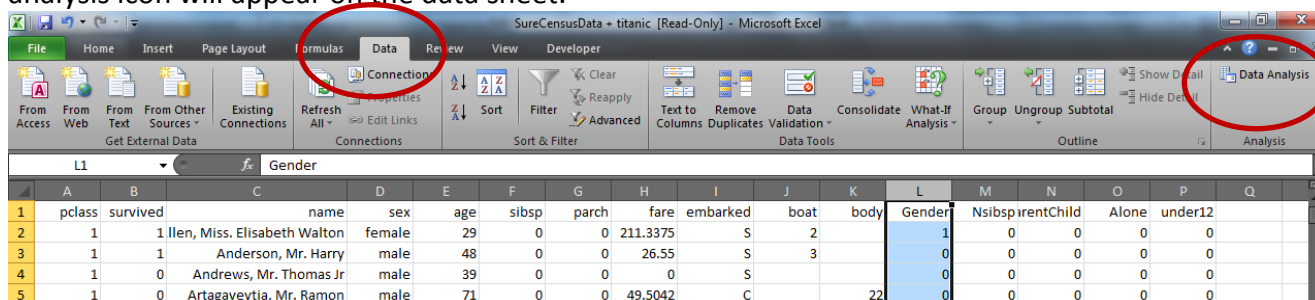
Each row represents an individual and each measurement (variable) for that individual is in a separate column. This example shows data available about the passengers aboard the ship Titanic when it sank in 1912. This dataset will be used to demonstrate summary statistics and charts and is available with this handout so that everything can be reproduced by the reader.

Summarising continuous data

Continuous variables are those measured on a numerical scale such as height, blood pressure and urine flow. The continuous variables for the Titanic data are age and fare.

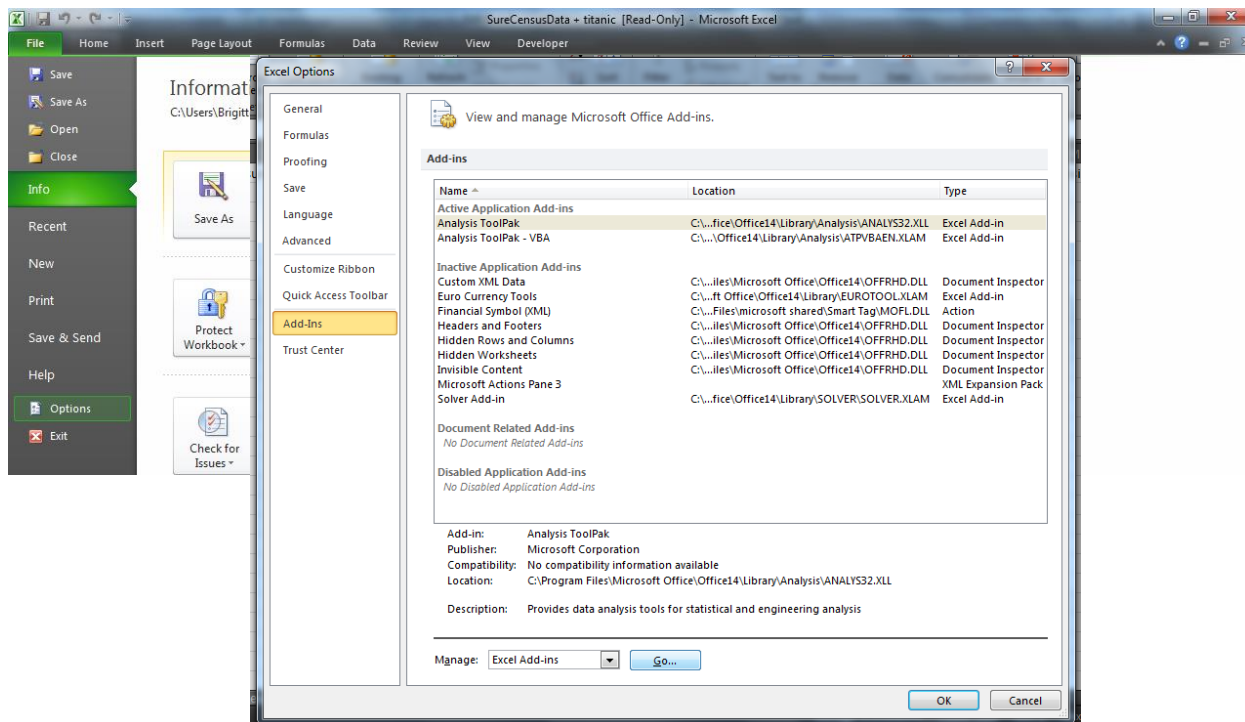
Summary Statistics for continuous variables	Charts
Measures of location: Mean and median	Histogram
Measures of spread: Standard deviation and quartiles	Scatterplot
Minimum and maximum values	Boxplot
	Line plot

The data analysis toolpak has an option for Descriptive Statistics. If the toolpak is loaded up, the data analysis icon will appear on the data sheet.

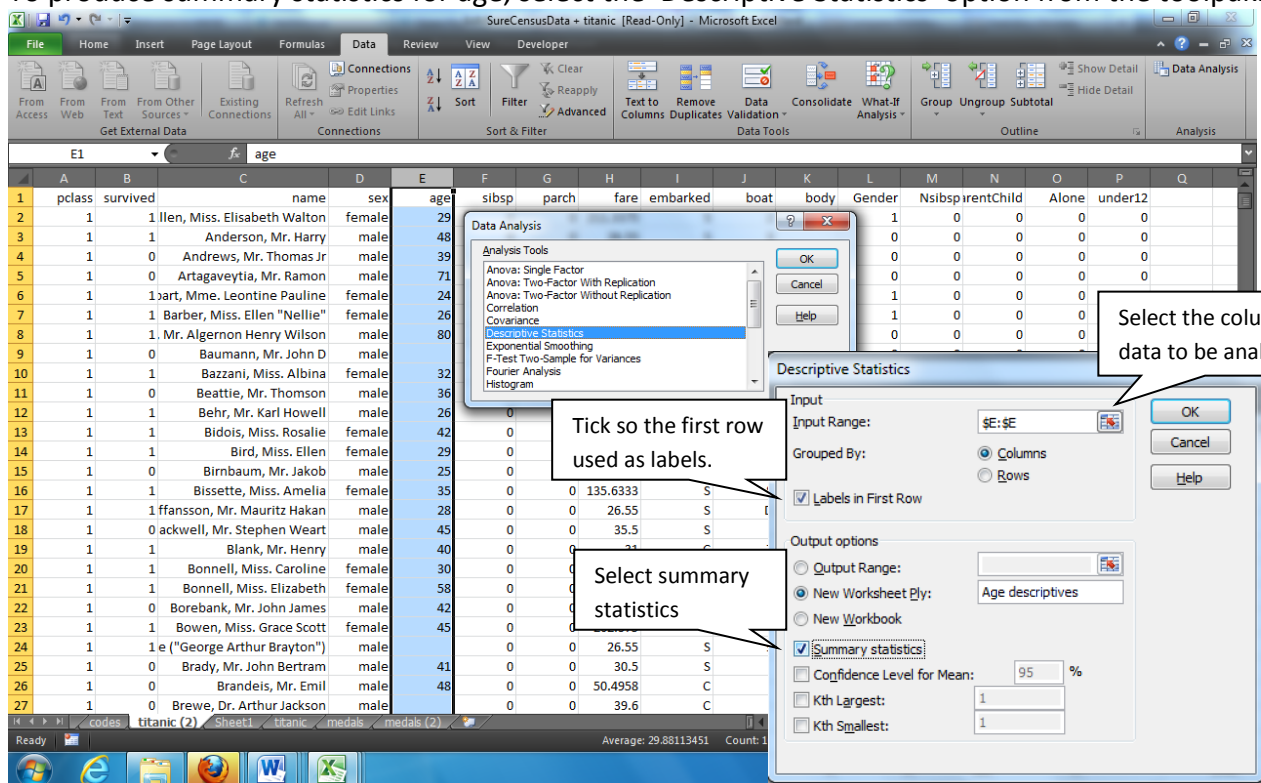


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	pclass	survived	name	sex	age	sibsp	parch	fare	embarked	boat	body	Gender	Nsibsp	parentChild	Alone	under12	
1	1	1	Ellen, Miss. Elisabeth Walton	female	29	0	0	211.3375	S	2		1	0	0	0	0	
2	1	1	Anderson, Mr. Harry	male	48	0	0	26.55	S	3		0	0	0	0	0	
3	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	0	S			0	0	0	0	0	
4	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	49.5042	C		22	0	0	0	0	0	

If it is not there, go to file → options → add ins, select the Analysis toolpak option and then 'Go'



To produce summary statistics for age, select the 'Descriptive Statistics' option from the toolpak.



This gives the following output:

<i>age</i>	
Mean	29.88
Standard Error	0.45
Median	28
Mode	24
Standard Deviation	14.41
Sample Variance	207.75
Kurtosis	0.15
Skewness	0.41
Range	79.83
Minimum	0.17
Maximum	80
Sum	31255.67
Count	1046

1. Measures of location

Mean: $\text{sum} / \text{count} = 29.88$

Median: 28 is the middle age if all the values are ordered from the youngest to the eldest person.

Mode: The age occurring most often is 24.

2. Measures of location:

Sample variance:

$$\frac{\text{sum of the deviations from the mean}}{\text{count} - 1} = 207.75$$

Standard deviation $\sqrt{\text{variance}} = 14.41$

Range: Maximum age – minimum age = 79.83

3. Measuring skewness of the distribution

If the distribution of scores is perfectly normally distributed

- Mean = median
- Skewness = 0
- Kurtosis = 0

The mean and median are fairly similar indicating roughly normally distributed

data. As a guide, compare the skewness statistic with $\pm \left(2 \times \sqrt{\frac{6}{N}} \right)$ where N

is the sample size. If $N > 50$, use 50. Here compare with ± 0.69

Conclusion: Data is approximately normally distributed.

Exercise 1: Calculate the summary statistics for fare. Is the data normally distributed?

Histograms

Another way to assess whether the data is normally distributed is to produce a histogram. Excel will only produce most charts if you supply it with a frequency table rather than the raw data with one row per subject. There is an option for doing this via the histogram option in the data analysis toolpak.

What is a frequency table?

BMI	Frequency
17.5	1
20	12
22.5	36
25	27
27.5	13
30	5
32.5	0
35	0
37.5	1

A frequency table is a summary of the raw data, where each individual is categorised into an appropriate group ('bin'). The data here categorises 95 people according to their BMI. Here the bin width is 2.5, starting at a BMI of 17.5. There is only 1 person with a Body Mass Index (BMI) of less than 17.5 and 12 people with a BMI of between 17.5 and 20.

Excel allocates random bin values so it's best to create your own. Thinking about the age data, the age range was 0 – 80 having a 10 year bin width starting with 0 - 10 and ending with 70 – 80 seems sensible.

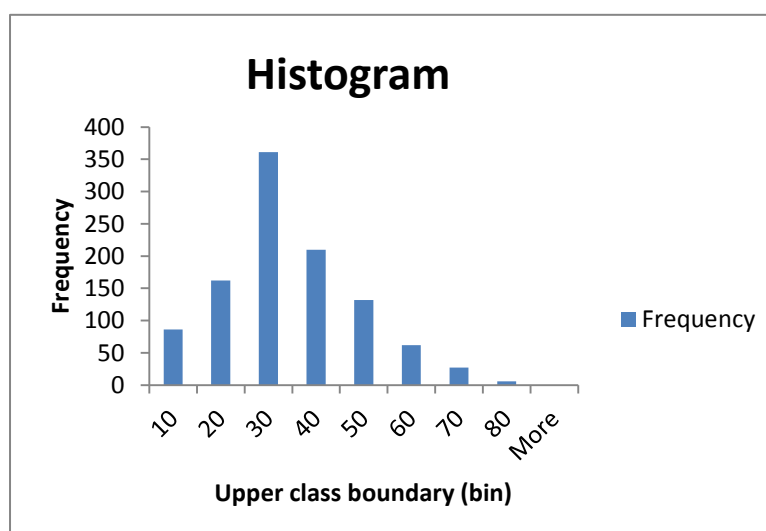
Input Range: \$A:\$A
Bin Range: \$D\$1:\$D\$9
Labels: ☒
Output options:
Output Range: \$G\$1
New Worksheet Ply: ☒
New Workbook: ☐
Pareto (sorted histogram): ☐
Cumulative Percentage: ☐
Chart Output: ☒

If you just need a frequency table do not select the chart output

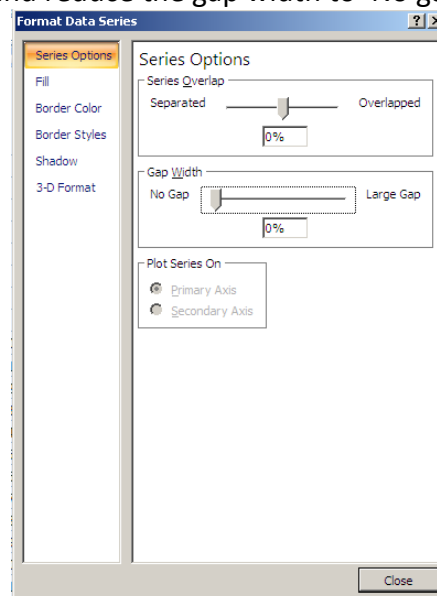
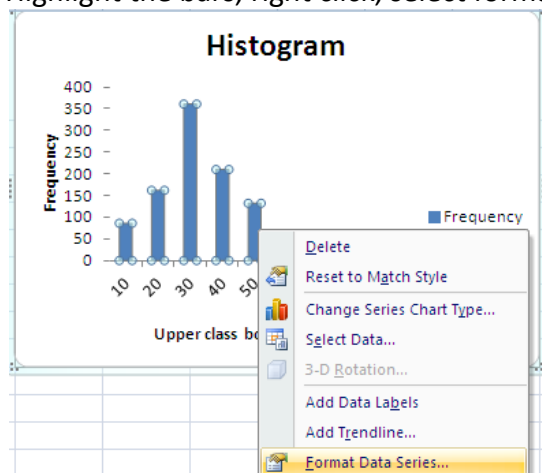
This is the output but the chart is not actually a histogram, it is more like a bar chart and therefore some adjustments are needed.

There should be no gap between the bars and the x-axis labels should be right aligned to be closer to the tick marks. The tick marks represent the boundaries where two classes meet.

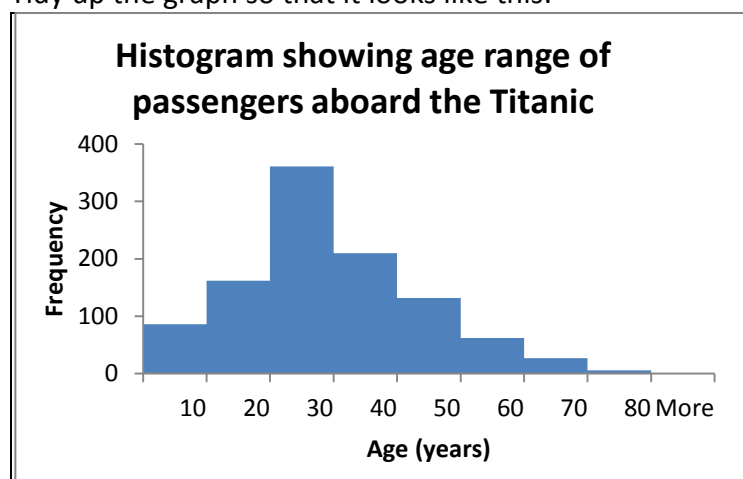
Upper class boundary (bin)	Frequency
10	86
20	162
30	361
40	210
50	132
60	62
70	27
80	6
More	0



Highlight the bars, right click, select format data series and reduce the gap width to 'No gap'.

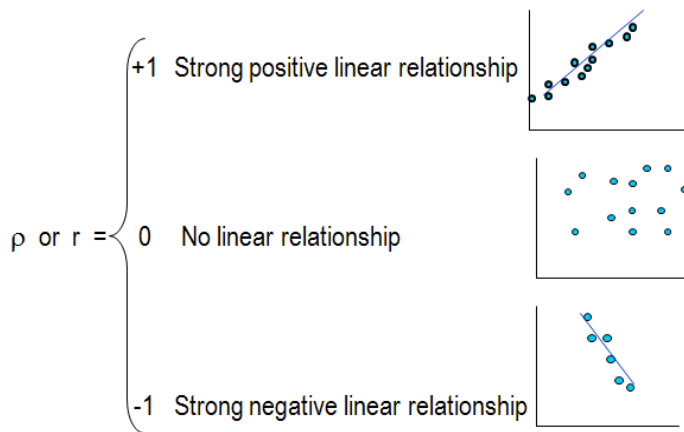


Tidy up the graph so that it looks like this:

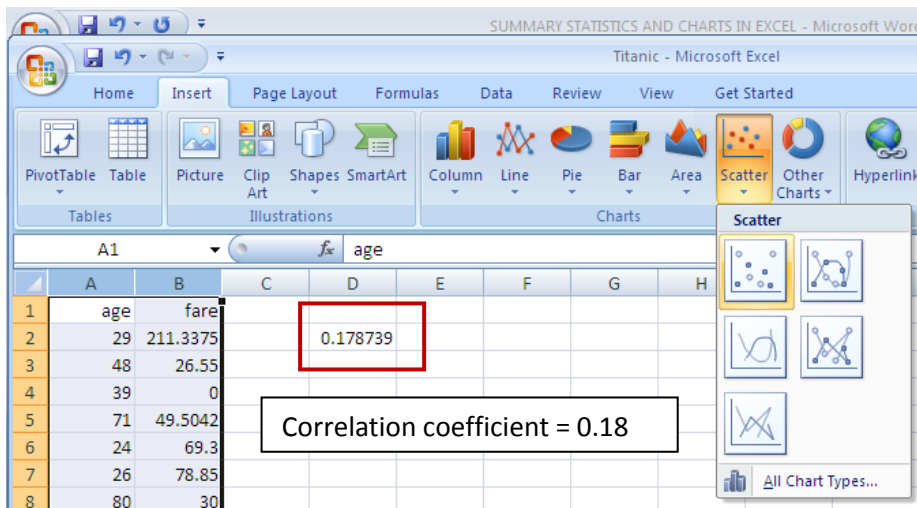


Scatterplots

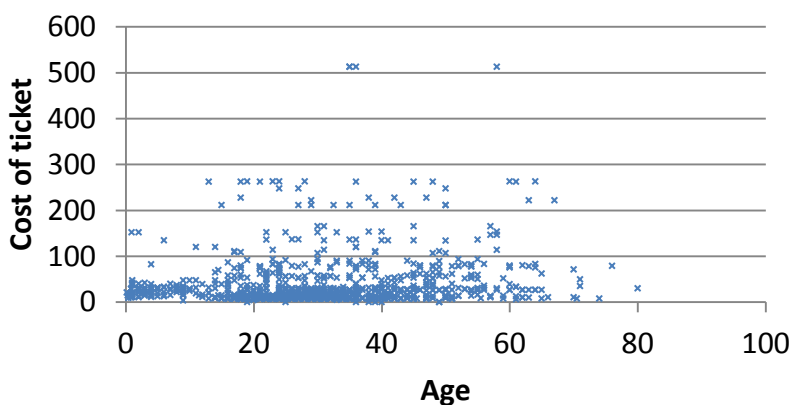
A scatter plot is a useful tool when assessing the strength of a relationship between two continuous variables and correlation helps determine whether relationships between these variables exist. Pearson's correlation co-efficient (r) measures the strength of a relationship between two continuous variables. It is a number between -1 and 1 where -1 is perfect negative correlation and +1 is perfect positive correlation. The command for calculating the correlation coefficient is `CORREL(array 1, array 2)` where array 1 is the 1st variable and array 2 the 2nd.



There are only two continuous variables in the Titanic data set, age and cost of ticket.



Scatterplot showing relationship between age and cost of ticket



The scatterplot shows no real relationship between age and cost of ticket and the very small correlation co-efficient of 0.18 confirms this. It's clear from the chart that there are some outliers with very expensive tickets.

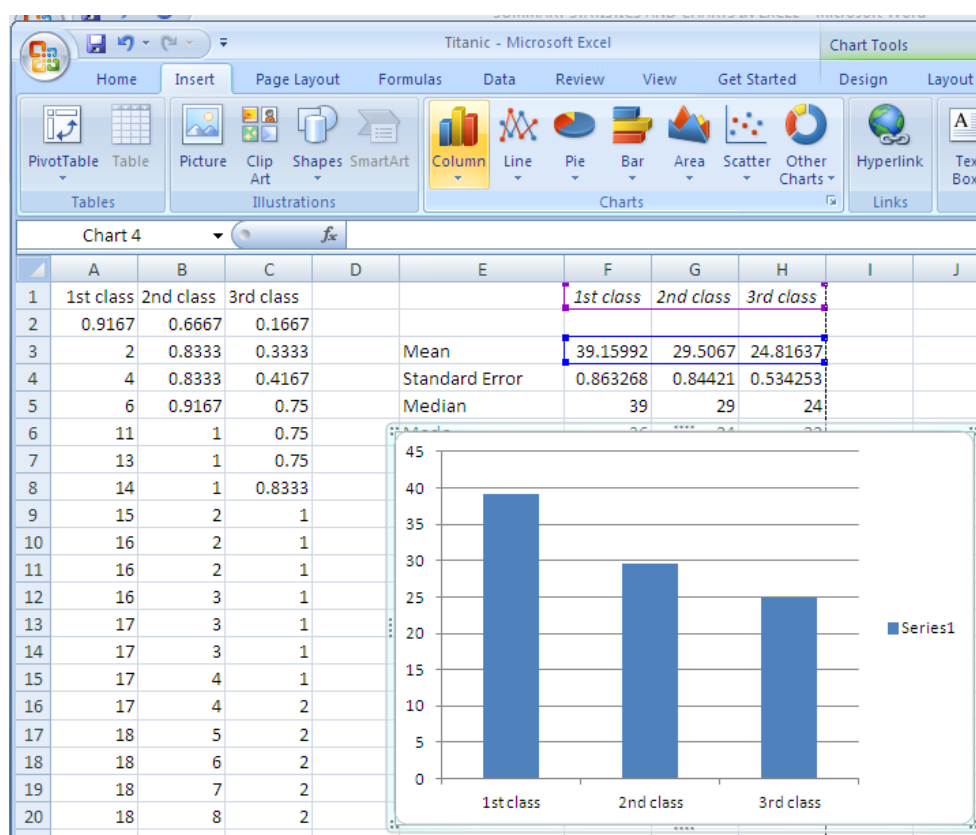
Using bar charts to display means and standard deviations

Sometimes, bar charts are used to compare means and spread for different groups. If this is required, the means and standard deviations for each group must be contained within a summary table. If the data for all groups is contained within one column, the first step is to re-organise the data into separate columns one for each group. Age within each class will be used to demonstrate here.

Sort the class and age variables by class, then move 2nd and 3rd class into separate columns.

Use the descriptive option in the data analysis toolpak to summarise by class.

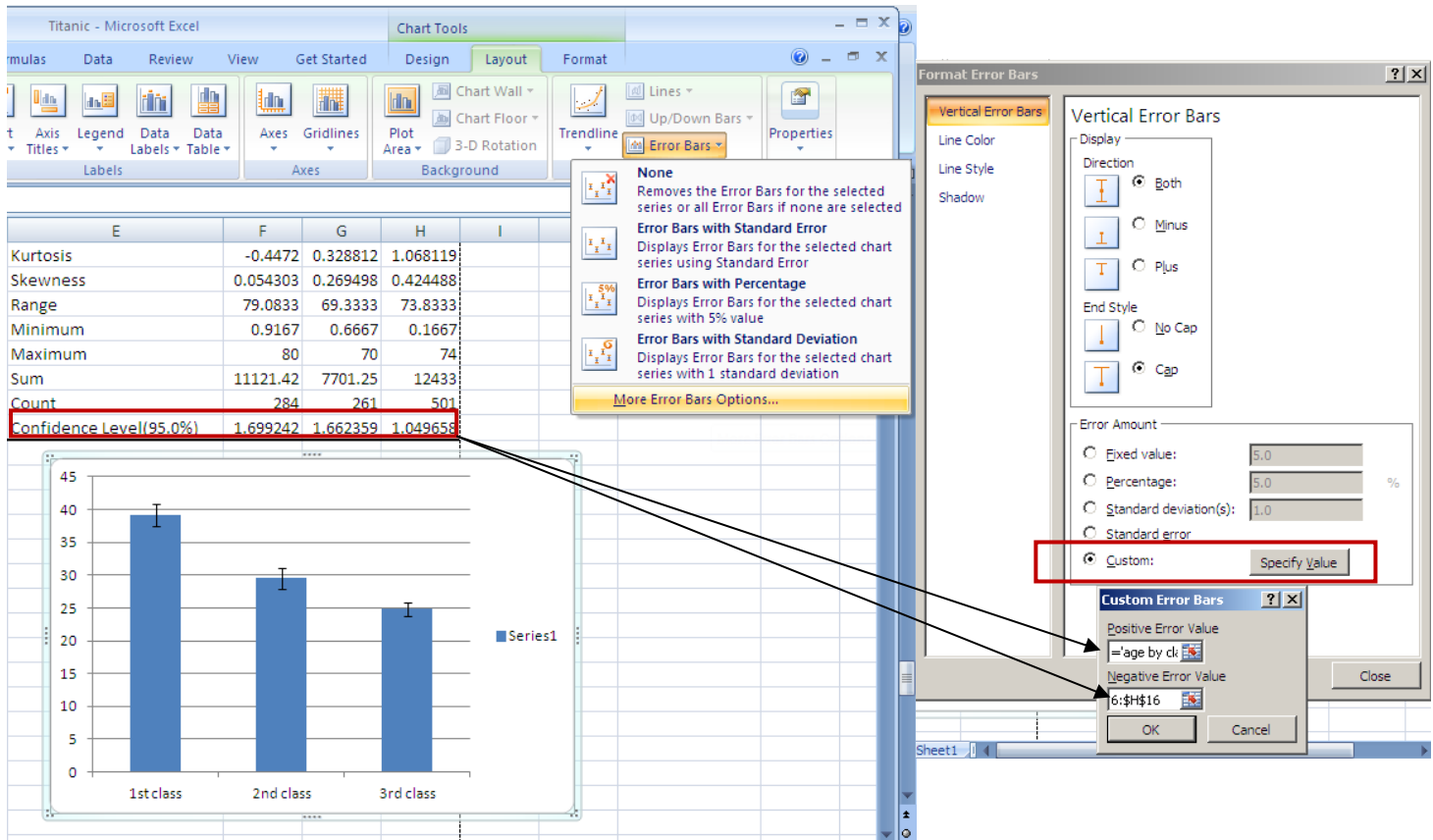
Highlight the class names and the row containing the means before selecting column chart.



It is clear from the resulting chart that the mean age decreases with class i.e. the average age in first class was higher than the average age in 2nd and 3rd class. The addition of confidence interval bars can be useful especially if you are looking for evidence of a difference between groups.

Summarising data in Excel

Maths and Statistics Help Centre



As every sample of people results in a different mean, confidence intervals are often quoted in studies alongside the mean. The 95% confidence interval states the two values between which we would expect to find 95% of sample means if the experiment was repeated numerous times. The wider the interval, the less reliable the sample mean.

The confidence level given in the descriptive statistics output is NOT the confidence interval.

Confidence interval = mean \pm confidence level. For example, for 1st class the confidence interval is 39.16 \pm 1.7 = (37.46, 40.86). So if there were 100 other ships and samples of 284 1st class passengers were taken from each, 95 of those samples would be expected to have a mean age between 37.46 and 40.86.

To add this information to the bar chart go to layout \rightarrow error bars \rightarrow more error bar options \rightarrow custom and select the row with the confidence level values for both the positive and negative error values.

The confidence intervals are very narrow suggesting that the mean age by class is a good estimate of the general population. The sample sizes are quite large which leads to narrower confidence intervals. Confidence intervals are also a quick way of looking for significant differences between groups. If the confidence intervals for two groups do not overlap, there is evidence of a significant difference between the means of the groups. If you just want to display the confidence intervals, choose 'no fill' and 'no line' for the bars so that they disappear.

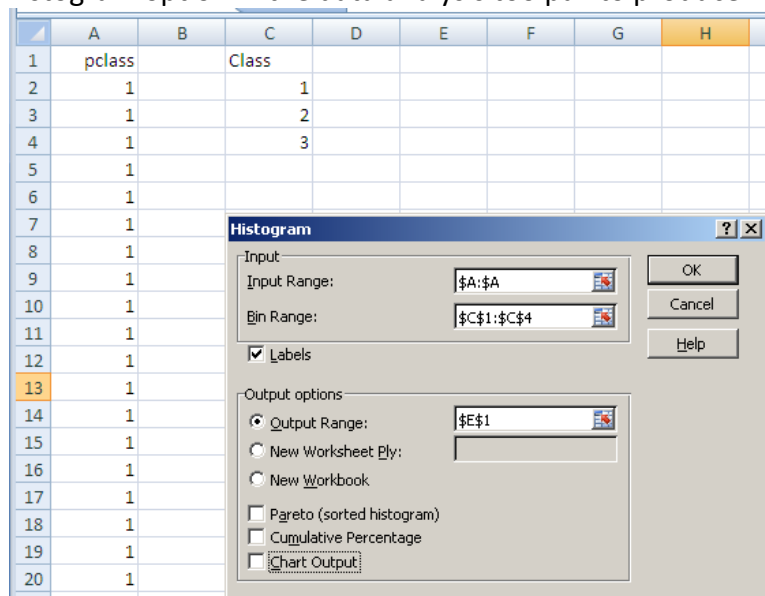


There is no overlap between the confidence intervals for the three classes indicating that there is a significant difference between the mean ages of the three classes.

Exercise 2: Create a chart displaying the confidence interval bars for the mean cost of ticket by survival (i.e. compare those who died and survived).

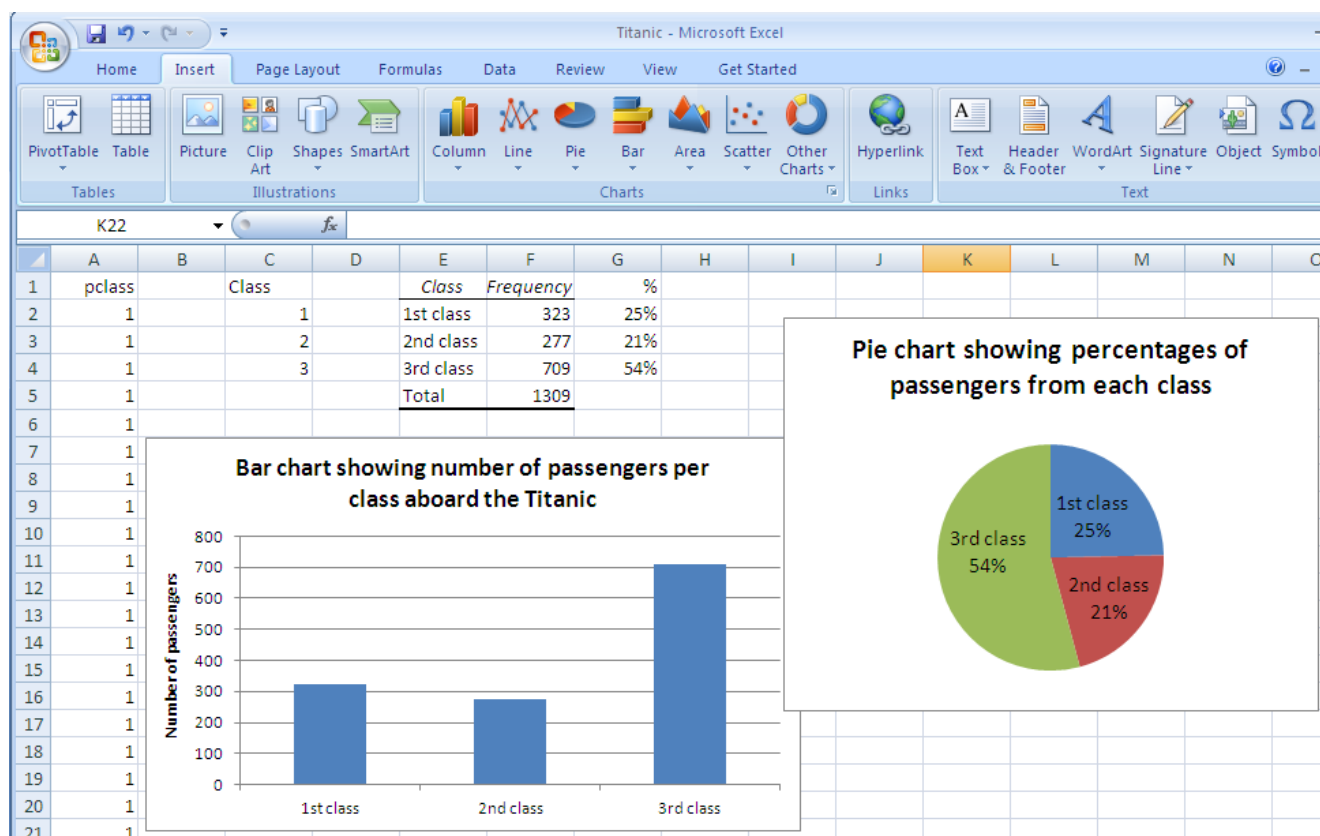
Summarising categorical data

Categorical variables such as gender and likert scales are usually summarised using percentages, bar charts or pie charts. In order to produce all of these, frequency tables need to be produced first. Use the histogram option in the data analysis toolpak to produce frequency tables for the variables of interest.



Class	Frequency
1	323
2	277
3	709
More	0

Remove the more category, calculate a total using sum() and calculate percentages. Highlight the class and frequency columns and select 'Column chart' from the Insert menu to create a bar chart or pie chart to create a pie chart. After producing the basic chart, make adjustments such as adding titles and data labels.



Contingency tables

Stacked or multiple bar charts require count data for two categorical variables displayed in a contingency table. A contingency table or cross-tabulation table displays the frequencies for each combination of the two variables.

For example, the following table shows the numbers of passengers who died within each class.

	1st class	2nd class	3rd class	Total
Died	123	158	528	809
Survived	200	119	181	500
Total	323	277	709	1309

To create the table above, put the variable 'Survived' into three columns, (one for each class) and then use the *countif(data, criteria)* command. The commands for calculating row and column percentages are also included below.

Summarising data in Excel

Maths and Statistics Help Centre

	A	B	C	D	E	F	G	H	I	J
	survived in 1st class	survived in 2nd class	survived in 3rd class				1st class	2nd class	3rd class	Total
1										
2	0	0	0		0	Died	=COUNTIF(A:A,\$E2)	=COUNTIF(B:B,\$E2)	=COUNTIF(C:C,\$E2)	=SUM(G2:I2)
3	0	0	0		1	Survived	=COUNTIF(A:A,\$E3)	=COUNTIF(B:B,\$E3)	=COUNTIF(C:C,\$E3)	=SUM(G3:I3)
4	0	0	0			Total	=SUM(G2:G3)	=SUM(H2:H3)	=SUM(I2:I3)	=SUM(J2:I4)
5	0	0	0							
6	0	0	0		Column percentages					
7	0	0	0		% within class		1st class	2nd class	3rd class	
8	0	0	0			Died	=G2/G\$4	=H2/H\$4	=I2/I\$4	
9	0	0	0			Survived	=G3/G\$4	=H3/H\$4	=I3/I\$4	
10	0	0	0			Total	=SUM(G8:G9)	=SUM(H8:H9)	=SUM(I8:I9)	
11	0	0	0							
12	0	0	0		Row percentages					
13	0	0	0							
14	0	0	0		% within survived		1st class	2nd class	3rd class	Total
15	0	0	0			Died	=G2/\$J2	=H2/\$J2	=I2/\$J2	=SUM(G15:I15)
16	0	0	0			Survived	=G3/\$J3	=H3/\$J3	=I3/\$J3	=SUM(G16:I16)
17	0	0	0							
18	0	0	0							
19	0	0	0							

Exercise 3: Which percentages are better for answering the question ‘Was chance of survival equal in every class?’.

Column %'s	1st class	2nd class	3rd class
Died	38%	57%	74%
Survived	62%	43%	26%
Total	100%	100%	100%

Row %'s	1st class	2nd class	3rd class	Total
Died	15%	20%	65%	100%
Survived	40%	24%	36%	100%

Solutions to exercises.

Exercise 1: Calculate the summary statistics for ticket fare paid. Is the data normally distributed?

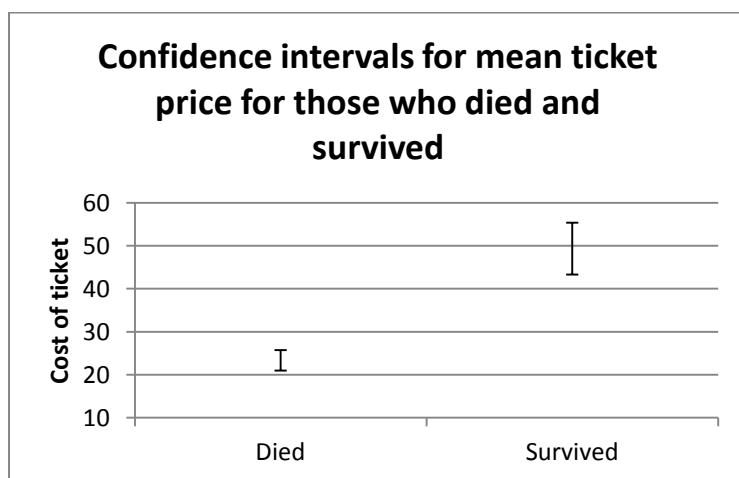
fare	
Mean	33.30
Median	14.45
Standard Deviation	51.76
Sample Variance	2678.96
Kurtosis	27.03
Skewness	4.37
Range	512.33
Minimum	0
Maximum	512.33
Sum	43550.49
Count	1308

The mean amount paid for a ticket on the Titanic was £33.30 and the median £14.45 suggesting that the data is skewed. This is confirmed by the measure of skewness which is 4.37.

The range of the data is £512.33 and standard deviation is £51.76, which is bigger than the mean so the data is very spread out.

Some people did not pay for a ticket at all and the largest amount paid for a ticket was £512.33.

Exercise 2: Create a chart displaying the confidence interval bars for the mean cost of ticket by survival (i.e. compare those who died and survived).



Exercise 3: Which percentages are better for answering the question 'Was chance of survival equal in every class?'

Column %'s	1st class	2nd class	3rd class
Died	38%	57%	74%
Survived	62%	43%	26%
Total	100%	100%	100%

Row %'s	1st class	2nd class	3rd class	Total
Died	15%	20%	65%	100%
Survived	40%	24%	36%	100%

The column %'s are better as there were different numbers of passengers within each class and therefore it is unfair to use the row percentages. 65% of those who died were from 3rd class but 54% of passengers were in 3rd class so we would expect more to have died.

Using the column percentages, (% within class), is fairer. 74% of passengers in 3rd class died but only 15% of 1st class passengers died.