



#### Data

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

X	<b>- 9 - </b> (°	-   <del>-</del>	and the second second	I maintaine		1.000	Sure	ensusData +	titanic [Rea	ad-Only] - N	licrosoft Excel	ALC: N	-	-	-	-		- 0
File	e Hon	ne Inserf	t Page Layout	Formulas	Data	Review	View	Developer										^ (?) -
fi Inse Funct	rt Auto	Sum Recently Used	Financial Logical	* Time	k Lookup & Reference	Hath ™ath ™& Trig ▼ F	More Functions *	Manager		mula • n Selection	S = Trace Prec S = Trace Dep Remove A	endents 🥠 rrows 🔹 🔗	Error Checki Evaluate For	ng - 🔐	h Calcu	lation 🛅 C ons -	alculate Now alculate Shee	
			_	n Library				D	efined Name	s		Formul	a Auditing			Calcula	tion	
	11	•		Gender														
	А	В	С		D			G	Н		J	К	L	М	N	0	Р	Q
1	pclass	survived		name	sex	age	sibsp	parch	fare	embarked	d boat	body	Gender	Nsibspare	entChild	Alone	under12	
2	1	1.	len, Miss. Elisab	eth Walton	female	29	0	0	211.3375	9	S 2		1	0	0	0	0	
3	1	1	Andersor	n, Mr. Harry	male	48	0	0	26.55	5	5 3		0	0	0	0	0	
4	1	0	Andrews, Mr	r. Thomas Jr	male	39	0	0	0	5	5		0	0	0	0	0	
5	1	0	Artagaveytia,	Mr. Ramon	male	71	0	0	49.5042	(	2	22	0	0	0	0	0	
6	1	1)	art, Mme. Leont	ine Pauline	female	24	0	0	69.3	(	9		1	0	0	0	0	
7	1	1	Barber, Miss. Ell	en "Nellie"	female	26	0	0	78.85	5	5 6		1	0	0	0	0	
8	1	1.	Mr. Algernon He	enry Wilson	male	80	0	0	30	5	S B		0	0	0	0	0	
9	1	0	Baumann,	, Mr. John D	male		0	0	25.925	5	S		0	0	0	0		
10	1	1	Bazzani, N	viss. Albina	female	32	0	0	76.2917	(	8		1	0	0	0	0	
11	1	0	Beattie, M	lr. 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

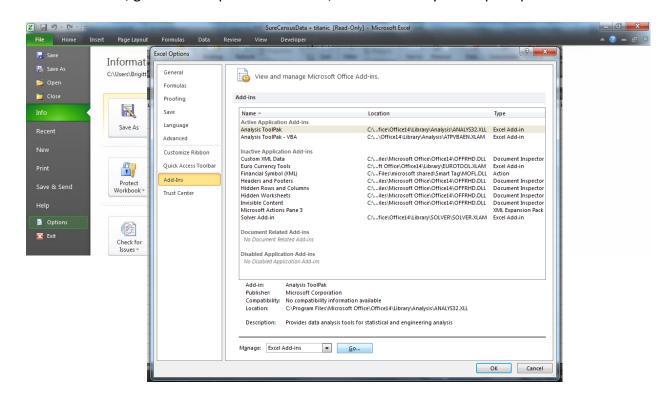
The data analysis toolpak has an option for Decriptive Statistics. If the toolpak is loaded up, the data analysis icon will appear on the data sheet.

<b>X</b>	<b>- 19 • (</b>	¥ -  ∓	_				SureC	ensusData +	titanic [Rea	ad-Only] - Mi	crosoft Excel	-	_	_	_	-	_	- 0 ×
Fi	e Ho	me Insert	Page Layout	formulas	Data	Review	View	Developer										^ ?) = ⊡ ×
Fro	m From	From From		Refresh	Connection Connection	2↓ Z↓	AZA Sort Filt	er Clea	oply Tex	tto Remove	Data	Consolida		Group L	Jngroup Subt	Ш - Т ні	iow Distail de Detsil	💾 Data Analysis
Acce	ss Web	Text Sour			🖘 Edit Links	AV	C	🖌 🏒 Adva	anced Colu	umns Duplicate			Analysis *	*	· ·			Anataria
		Get External I	Jata	Co	nnections		Sort 8	Fliter			Data Too	DIS			Outlin	ie	la l	Analysis
	L1	-	f <sub>≭</sub> Ger	nder														~
												К	L					Q 🗖
1	pclass	survived		name	sex	age	e sibsp	parch	fare	embarked	boat	body	Gender	Nsibsp	arentChild	Alone	under12	
2	1	1.1	en, Miss. Elisabeth	Walton	female	29	ə 0	0	211.3375	S	2		1	0	0	0	0	
3	1	1	Anderson, N	/Ir. Harry	male	48	3 0	0	26.55	S	3		0	0	0	0	0	
4	1	0	Andrews, Mr. Th	nomas Jr	male	39	9 0	0	0	S			0	0	0	0	0	
5	1	0	Artagaveytia, Mr	. Ramon	male	71	L 0	0	49.5042	С		22	0	0	0	0	0	





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



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

File	Hor	ne Insert Page Layout Fo	ormulas Data		eveloper	Contry - Whereson Excer			
From	From	From From Other Existing R	Refresh All - Connections	tions $2 \downarrow \overline{Z}$	Clear Reapply Advanced	o Remove Data ns Duplicates Validation Data Tor		Group Ungroup Subtotal	w Detail 🛄 Data Analysis
_	E1	✓ fx age	Connections	3011 01	inter	Data To	015	Outline	isi Analysis
	E1	-8-		-	0 11				
1	A	B C survived	D name sex	age sibsp	G H parch fare	embarked boat	k L body Gender	M N O NsibsparentChild Alone	P Q under12
2	pciass 1	1 llen, Miss. Elisabeth W		20		embarked boat		0 0 0	
3	1	1 Anderson, Mr.		48 Data Anal	/sis			0 0 0	0
4	1	0 Andrews, Mr. Thor		39 <u>Analysis</u>	Tools		ОКО	0 0 0	0
5	1	0 Artagaveytia, Mr. R		Anova:	Single Factor			0 0 0	0
6	1	1)art, Mme. Leontine Pa		Ariova.	Two-Factor With Replication Two-Factor Without Replica		Cancel 1	0 0 0	
7	1	1 Barber, Miss. Ellen "N		26 Correlat Covaria		=	Help 1	0 0 0	Select the column of
8	1	1. Mr. Algernon Henry W	Vilson male	80 Descript	ive Statistics		0	0 0 0	
9	1	0 Baumann, Mr. J	lohn D male		itial Smoothing wo-Sample for Variances	-			data to be analysed.
10	1	1 Bazzani, Miss. A	Albina female	32 Fourier	Analysis	-	Descriptive Statistics		
11	1	0 Beattie, Mr. Tho	omson male	36 Histogra	m		Input		
12	1	1 Behr, Mr. Karl H	lowell male	26 0	The second	h - C	Input Range:	\$E:\$E	ОК
13	1	1 Bidois, Miss. R	osalie female	42 0	TICK SO T	he first row	Input Kange.	\$C:\$C	Cancel
14	1	1 Bird, Miss.	. Ellen female	29 0	used as	lahels	Grouped By:	Olumns	Cancer
15	1	0 Birnbaum, Mr.	Jakob male	25 0				Rows	Help
16	1	1 Bissette, Miss. A		35 0	0 135.6333	S	Labels in First Ro	w	
17	1	1 ffansson, Mr. Mauritz I		28 0	0 26.55	S [			
18	1	0 ackwell, Mr. Stephen		45 0	0 35.5	S	Output options		
19	1	1 Blank, Mr. I		40 0			<b>-</b> · ·		
20	1	1 Bonnell, Miss. Ca		30 0	Select	summary	Output Range:		
21	1	1 Bonnell, Miss. Eliz		58 0	0		New Worksheet	Ply: Age descriptives	
22	1	0 Borebank, Mr. John J		42 0	statist	ICS	New Workbook		
23	1	1 Bowen, Miss. Grace		45 0	0 26 55				
24 25	1	1 e ("George Arthur Bray 0 Brady, Mr. John Be		41 0	0 26.55	s s	Summary statisti		
25	1	,/		41 0	0 30.5	C	Confidence Leve	for Mean: 95 %	
27	1	0 Brandeis, Mr 0 Brewe, Dr. Arthur Ja		48 0	0 39.6	c	Kth Largest:	1	
14 4		des titanic (2) Sheet1 / tita		nedals (2) / 🐄 /	0 35.0		Kth Smallest:	1	
Ready	1 📶	~ · · · · ·			Average: 2	29.88113451 Count: 1	in Kur Jindiest.		
7	) (	š 📋 🙆 👿							





This gives the follow	ing output:	1. Measures of location
age		Mean: $sum / count = 29.88$
		Median: 28 is the middle age if all the values are ordered from the youngest
Mean	29.88	to the eldest person.
Standard Error	0.45	Mode: The age occurring most often is 24.
Median	28	2. Measures of location:
Mode	24	Sample variance:
Standard Deviation	14.41	$\frac{\text{sum of the deviations from the mean}}{\text{= 207.75}}$
Sample Variance	207.75	count - 1
Kurtosis	0.15	Standard deviation $\sqrt{variance} = 14.41$
Skewness	0.41	Range: Maximum age – minimum age = 79.83
Range	79.83	3. Measuring skewness of the distribution
Minimum	0.17	If the distribution of scores is perfectly normally distributed
Maximum	80	• Mean = median
Sum	31255.67	• Skewness = $0$
Count	1046	• Kurtosis = 0 The mean and median are fairly similar indicating roughly normally distributed
		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 $\pm 0.69$
		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?

BM	I Frequency
17.5	5 1
20	) 12
22.5	5 36
25	5 27
27.5	5 13
30	) 5
32.5	5 O
35	5 O
37.5	i 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.



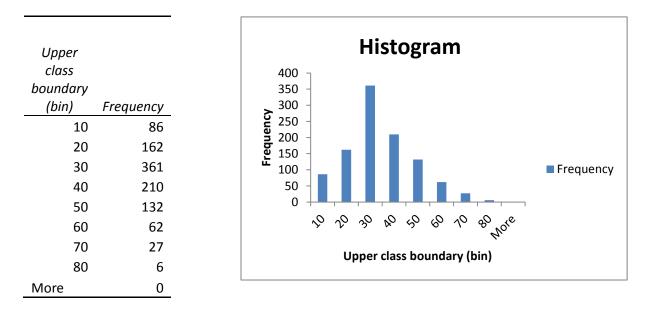
#### Summarising data in Excel Maths and Statistics Help Centre



23/ Home Formulas Data Get Started Insert Page Layout Review View Connections 🕅 Clear 📑 Data Validation 🔿 Group 🕯 0클 Data Analysis L) + a 41 A Z A Properties = E Consolidate 🔄 Ungroup 🂫 Solver Get External Refresh Text to Remove Filter Z↓ Columns Duplicates P What-If Analysis 📾 Edit Links 💯 Advanced Subtotal Data 1 All Connections Sort & Filter Data Tools Outline Analysis A1  $f_{x}$ <del>-</del> () М Α В C D Е F G Н Κ Upper ? × Histogram Lower class Input ОК class boundary Input Range: \$A:\$A 1 1 boundary (bin) Cancel age \$D\$1:\$D\$9 -Bin Range: 2 29 0 10 <u>H</u>elp ✓ Labels 3 10 20 48 4 39 20 30 Output options 5 71 30 40 \$G\$1 Output Range: 6 50 24 40 New Worksheet Ply: 7 26 50 60 New Workbook 60 8 80 70 Pareto (sorted histogram) 9 70 80 Г Cumulative Percentage 10 32 🔽 Chart Output 11 36 12 26 13 42 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.

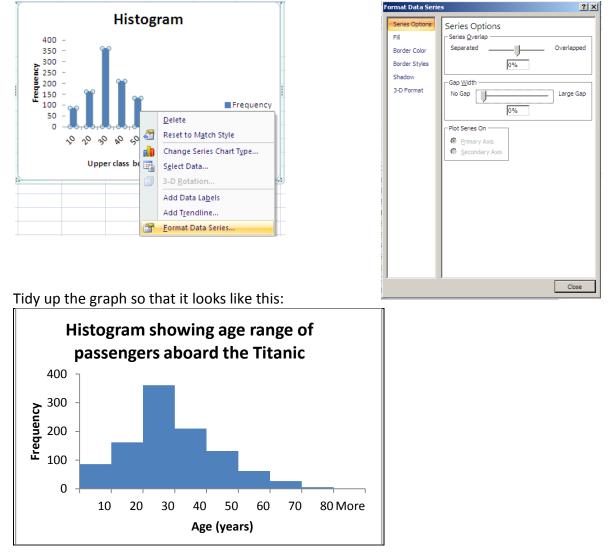
There should be no gap between the bars and the x-axis labels should be right aligned to be closer to the tick marks. The tick marks represent the boundaries where two classes meet.







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

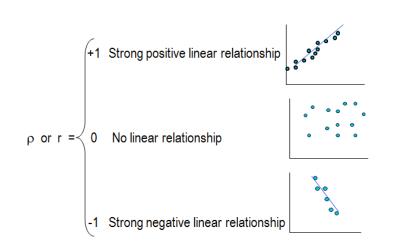


#### **Scatterplots**

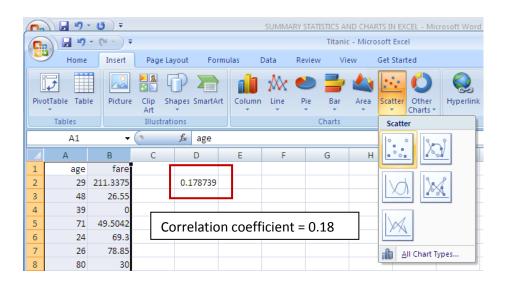
A scatter plot is a useful tool when assessing the strength of a relationship between two continuous variables and correlation helps determine whether relationships between these variables exist. Pearson's correlation co-efficient (r) measures the strength of a relationship between two continuous variables. 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 1<sup>st</sup> variable and array 2 the 2<sup>nd</sup>.

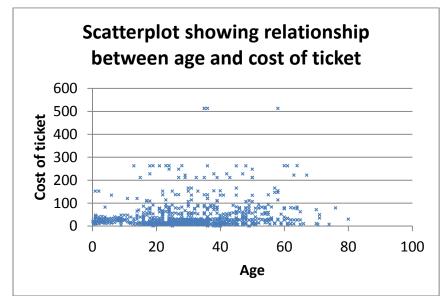






There are only two continuous variables in the Titanic data set, age and cost of ticket.









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 2<sup>nd</sup> and 3<sup>rd</sup> 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.

6	) 🖬 🤊	- ( <u>u</u> - );	;			Titanic	- Micro	sof	t Excel				Chart Too	ls
C	Home		Page Lay	yout Fo	rmula	s D	ata	R	eview	View	Ge	t Started	Design	Layout
Pivo	otTable Table	le Picture	Clip Sh Art Illustrati	apes Smart4	.rt C	olumn	Line		Pie Ba Charts	r A		atter Othe Charts		Box
	Chart 4	↓ <del>•</del>	()	f <sub>x</sub>										
	А	В	С	D		E			F		G	Н	I.	J
1	1st class	2nd class	3rd class						1st class	200	l class	3rd class	]	
2	0.9167	0.6667	0.1667										]	
3	2	0.8333	0.3333		Mea	n			39.1599	2 2	9.5067	24.81637	1	
4	4	0.8333	0.4167		Stan	dard E	rror		0.86326	8 0	.84421	0.534253		
5	6	0.9167	0.75		Med	lian			3	9	29	24	L I	
6	11	1	0.75	ŀ		1_		_	2					1
7	13	1	0.75		45 -									
8	14	1	0.8333		40 -									
9	15	2	1		35 -									
10	16	2	1											
11	16	2	1		30 -									
12	16	3	1		25		-							
13	17	3	1		20 -									Series1
14	17	3	1											
15	17	4	1		15 -									
16	17	4	2		10 -									
17	18	5	2		5.									
18	18	6	2		-									
19	18	7	2		0 ·	1	st class		200	d class		3rd class		
20	18	8	2			1	scelass		200	u class		ord class		
21	10	-	2											

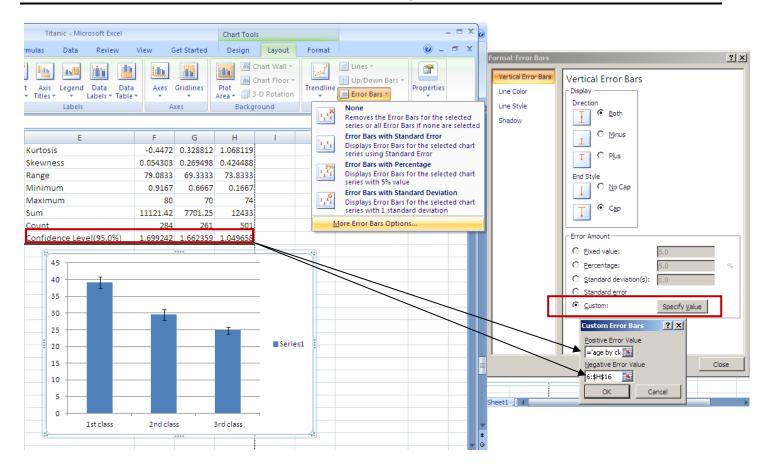
It is clear from the resulting chart that the mean age decreases with class i.e. the average age in first class was higher than the average age in 2<sup>nd</sup> and 3<sup>rd</sup> class. The addition of confidence interval bars can be useful especially if you are looking for evidence of a difference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

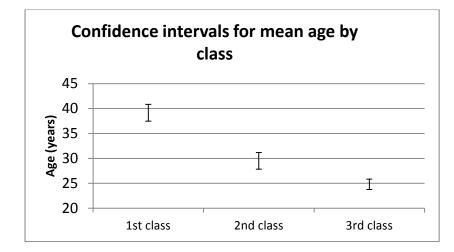
The confidence level given in the descriptive statistics output is NOT the confidence interval. Confidence interval = mean  $\pm$  confidence level. For example, for 1<sup>st</sup> class the confidence interval is 39.16  $39.16 \pm 1.7 = (37.46, 40.86)$ . So if there were 100 other ships and samples of 284 1<sup>st</sup> 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.





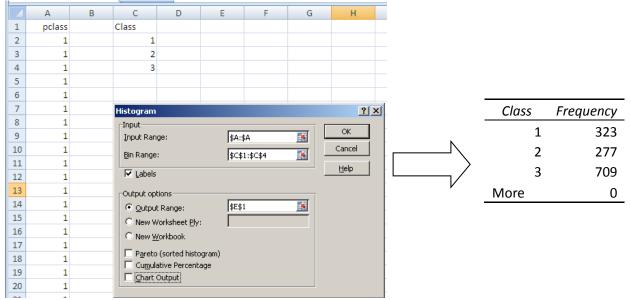


There is no overlap between the confidence intervals for the three classes indicating that there is a significant difference between the mean ages of the three classes.

# **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.



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.



# Summarising data in Excel



Maths and Statistics Help Centre

	) 🔒 🤊 -	(21 - ) =					Titanio	- Microsof	t Excel						-
C	Home	Insert	Page La	yout Fo	mulas	Data Rei	view Vie	ew Get	t Started					(	0 - 1
Pivo	tTable Table	Picture	Clip Sh Art Illustrati	apes SmartA	rt Columr	Line Pi	e Bar Charts	Area Sc	tter Other → Charts →	Kan	A Text Box •	Header Wo & Footer	ordArt Signat Line Text		Ω Symbol
	K22	•	()	$f_{x}$											
	А	В	С	D	E	F	G	Н	I.	J	K	L	М	N	0
1	pclass		Class		Class	Frequency	%								
2	1		1		1st class	323	25%								L
3	1		2		2nd class	277	21%			Pie ch	art sho	owing pe	ercentag	tes of	
4	1		3		3rd class	709	54%					rs from			
5	1				Total	1309				pas	senge	13 11 0111	each cia	133	
6	1														
7	1		Barcha	art showi	ng num	ber of pa	cenger	ner							
8	1		Dai che			the Titan		sper							_
9	1			class	aboard	the Ittan	IC					1st c			
10	1	800									3rd cl	ass 25	»%		
11	1	<sup>700</sup>									549				
12	1	<b>6</b> 00						-					d class		_
13	1	500											21%		_
14	1	Number of passengers 200 400 200 200													_
15	1	300													_
16	1	<b>4</b> 300													
17	1														
18	1	100	<u> </u>												
19	1	0			-		_								
20	1		1	st class	2	2nd class	3	Brd class							
21	1														

### **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.

		2nd	3rd	
	1st class	class	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



	А	В	С	D	E	F	G	Н	l.	J	
	surviv	surviv	surviv								
		ed in	ed in								
	ed in										
4	1st		3rd				1 at al a a	2014	2.4		Tetel
1	class						1st class		3rd class		Total
2	0		0		0		=COUNTIF(A:A,\$E2)	=COUNTIF(B:B,\$E2)	=COUNTIF(C:C,\$E2)	=SUM(G2:12)	
3	0	0	0		1			=COUNTIF(B:B,\$E3)	=COUNTIF(C:C,\$E3)	=SUM(G3:I3)	
4	0	0	0			Total	=SUM(G2:G3)	=SUM(H2:H3)	=SUM(12:13)	=SUM(G4:I4)	
5	0	0	0								
6	0	0	0		Column per	entages					
					% within						
7	0	0	0		class		1st class	2nd class	3rd class		
8	0	0	0			Died	=G2/G\$4	=H2/H\$4	=12/1\$4		
9	0	0	0			Survived	=G3/G\$4	=H3/H\$4	=13/1\$4		
10	0	0	0				=SUM(G8:G9)	=SUM(H8:H9)	=SUM(18:19)		
11		0	0				()				
12	-	0	0		Row percent	2005					
13	-	0	0		Now percent	ages					
13	0	U	0		0/tale too						
	_	_	_		% within						
14			0		survived		1st class				Total
15		0	0				=G2/\$J2	=H2/\$J2	=12/\$J2	=SUM(G15:I15)	
16	-	0	0			Survived	=G3/\$J3	=H3/\$J3	=13/\$J3	=SUM(G16:I16)	
17	0	0	0								
18	0	0	0								
10		history	am example	- / ch	eet2 🧹 titanic (	2) contingency tabl	e class summary	Scatter 1			

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

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

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



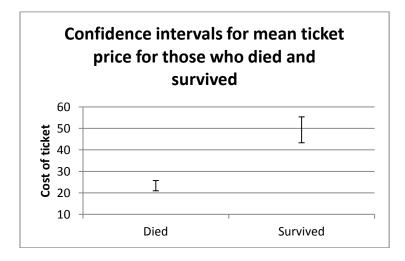


Solutions to exercises.

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

**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			2nd	3rd	
Died	38%	57%	74%	Row %'s	1st class	class	class	Total
Survived	62%	43%	26%	Died	15%	20%	65%	100%
Total	100%	100%	100%	Survived	40%	24%	36%	100%

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

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