

# Session 7.3: Trainee Notes

## Presenting data

*'Graphical excellence demands a blend of statistical rigour and graphical design skills that is unfortunately rare.'* (Tufte, 2001)

### Content

1. Introduction
2. Tables
3. Basics for graphical design
4. Displaying categorical data
5. Displaying numerical data
6. Displaying time series data

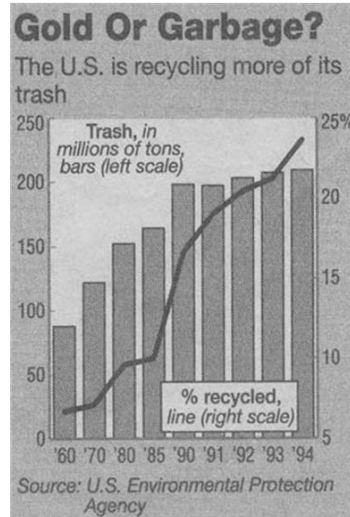
### Learning Objective

By the end of this session you will be able to:

- Assess the best way to present your data
- Choose the best graph for different scenarios and audiences
- Present graphics clearly

# 1. Introduction

Pictures are often appealing. If designed properly, they are easy and quick to understand. Or, as the following chart falls into, they can also be misleading and confusing.



So, some thought is important before delving into a graphing package and finding the prettiest graph! You should think about your **audience** (who are often non-specialists) and **what you want your audience to understand**.

Tables and graphs should also be **simple** and **self-explanatory** – the reader should be able to understand them without making detailed reference to the text. Conversely, if there is no need to describe and discuss the graphs or table in the text then there is generally no need for the table or graph at all.

## Advice

Write an aim for your table or graph before you start building it.

### Example

I want my graph to show relative amounts of mileage travelled by different vehicles each year.

## Choosing graphs or tables

Data can be displayed in many different ways to suit your audience, your data and your message. Graphs and tables should be chosen carefully to emphasize and reveal important facts, comparisons and relationships in the best possible way.

In general **tables** are better than graphs **for giving structured numeric information**, whereas **graphs** are better **for indicating trends and making broad comparisons or showing relationships**.

## 2. Tables

For displaying data use a table if:

- The trend is not important (or at least not the most important feature); or
- The number of values is small.

Firstly consider if you need to show actual data values, or whether a summary of the data would be more informative. Consider Table 1 (a) and (b).

### Example

The aim of these tables is to portray the differences in motorised traffic levels by different vehicle types in UK in 2008.

**Table 1(a): Motorised traffic levels in GB by vehicle type (2008)**

Vehicle type	Traffic (billion veh-km)
Cars & taxis	401.749
LGVs	68.0970
Buses & coaches	5.18446
Motorcycles	5.14152
HGVs: 2 axles rigid	10.7401
HGVs: 3 axles rigid	2.01217
HGVs: 4 or more axles rigid	1.85355
HGVs: 3 and 4 axles artic	1.60562
HGVs: 5 axles artic	6.51870
HGVs: 6 or more axles artic	6.00398
All vehicles	508.906

**Table 1(b): Motorised traffic levels in GB by vehicle type (2008)**

Vehicle type	Traffic (billion veh-km)
Cars & taxis	402
LGVs	68
HGVs	29
Buses & coaches	5
Motorcycles	5
All vehicles	509

Table 1 (b) shows a summary of the information, by **rounding numbers effectively**, **ordering numbers** and **removing unnecessary information**, from Table 1 (a), and it is much easier to see the pattern.

The ordinary human being can discriminate only up to two digits in numbers so the general advice is to **round figures in tables to two significant digits**. In fact data are almost always estimates, not exact figures, so presenting too many significant digits suggests a misleading level of confidence in our estimates.

## Percentages in tables

It is often helpful to present percentages in tables, rather than numbers, especially if you are trying to compare relative patterns rather than absolute numbers. You should always show the sample sizes on which the percentages are based.

If your data include some unknown values then it is also important to mention these, and show whether these values are included in the percentages.

You can:

- Make row percentages add up to 100
- Make column percentages add up to 100
- Make 100 across whole table

Each of these options facilitates different types of comparisons

### Example

The aim of this table is to show the number of different types of traffic offences observed by Police Officers during 1 week on four different roads in London.

**Table 2: Number of traffic offences spotted by Police Officers, by road and type of offence.**

Road	Speed	Mobile phone	Seatbelt	Insurance
Road A	56	127	211	9
Road B	182	126	218	10
Road C	123	179	236	9
Road D	81	76	215	8

### Column distribution

The aim of this table is to show the distribution of each type of traffic offence across the different roads.

**Table 2a: Distribution of each type of traffic offence spotted by Police Officers, by road.**

Road	Speed	Mobile phone	Seatbelt	Insurance
Road A	13%	25%	24%	25%
Road B	41%	25%	25%	28%
Road C	28%	35%	27%	25%
Road D	18%	15%	24%	22%
Sample size	442	508	880	36

Each column adds up to 100% and comparisons can be drawn between columns.

E.g.: The majority of speed offences (41%) were observed on road B, whereas the majority of mobile phone offences (35%) were observed on road C.

**Example ctd.**

*Row distribution*

The aim of this table is to show for each road, the distribution of traffic offences.

**Table 2b: Distribution of traffic offences spotted by Police Officers across each road.**

Road	Speed	Mobile phone	Seatbelt	Insurance	Sample size
Road A	14%	32%	52%	2%	403
Road B	34%	24%	41%	2%	536
Road C	22%	33%	43%	2%	547
Road D	21%	20%	57%	2%	380

Each row adds up to 100% and comparisons can be drawn between rows.

E.g.: There were proportionately more seatbelt offences spotted on road D than any other road  
The majority of offences (between 41% and 57%) on every road were seatbelt offences.

*Total distribution*

The aim of this table is to show the distribution of traffic offences and roads.

**Table 2c: Distribution of different types of traffic offences and roads (sample size = 1,866).**

Road	Speed	Mobile phone	Seatbelt	Insurance
Road A	3%	7%	11%	<1%
Road B	10%	7%	12%	1%
Road C	7%	10%	13%	<1%
Road D	4%	4%	12%	<1%

The whole table adds up to 100% and comparisons can be drawn across the whole table.

E.g.: The most traffic offences were seatbelt offences spotted on road C.

## Principles for tables

1. Round data consistently in summary tables
2. Use consistent units
3. Use captions and row and column headings to show what the units are and what the numbers mean
4. Right justify numbers in columns (or at least make all units, tens, hundreds etc. line up)
5. Make sure that all tables in a publication are in a similar format
6. Show time either from left to right or top to bottom
7. Show row totals to the right and column totals to the bottom
8. Put data to be compared in columns rather than rows
9. Keep tables as simple as possible

## Examples of bad tables

### Example 1:

**Table 3: Citrus prices**

	1995/96	1996/97	1997/98	1998/99	1999/2000
	national currency/kg				
<b>ORANGES AND TANGERINES</b>					
Germany (DM) Spanish navels	1.46	1.43	1.38	1.53	1.27
Spanish clementines	2.25	2.16	1.97	2.09	2.04
United States (cents) California navels	66.48	65.58	90.06	132.19	68.11
Japan (yen) Average	181.44	293.58	269.17	264.08	237.1

This table is presented on the website of an international organisation. We would question the validity of this table at all, given we can only look at the patterns of particular citrus types for each of a small selection of countries.

*Possible improvement*

**Table 3a: Average price of citrus fruit per kg at current USD price, by year**

	Spanish navels (Germany)	Spanish clementines (Germany)	California navels (USA)	All (Japan)
1995/96	0.99	1.52	0.66	2.11
1996/97	0.97	1.42	0.65	3.41
1997/98	0.93	1.33	0.90	3.13
1998/99	1.03	1.41	1.32	3.07
1999/00	0.86	1.38	0.68	2.76

**Example 2:**

**Table 4: Survey of cycling trips in France, compared to pre-existing survey results in London and UK\***

Number of cycling trips	France		UK	
	Survey	%	London data	National survey
3+ trips a week	271	7.4	34	22
1-2 trips a week	443	12.2	10	34
Less than that but at least 1 per month	1849	50.7	2	23
Less than that but at least 1 per year	992	27.2	2	12
Never cycle	89	2.4	51	11
Total valid responses	3644			
<i>No response</i>	125	3.3		

\* For all except the last row, percentages calculated as a proportion of the valid responses.

This table is similar to one presented in a TRL report (anonymised!). The comparison is a little odd – comparing a national survey across cyclists with a city survey, over a sample which clearly includes people who do not cycle at all.

*Possible improvement*

**Table 4a: Distribution of cycling trips in a survey carried out in France, compared with a pre-existing UK survey.**

Number of trips	France	UK
3+ times a week	7%	22%
1-2 times a week	12%	34%
Less than that but at least 1 per month	51%	23%
Less than that but at least 1 per year	27%	12%
No trips made	2%	11%
Total valid responses	3,644	-
<i>No response</i>	125	-

### 3. Basics for graphs

Graphs are often the best way of presenting trend data or large amounts of data. As with tables you still need to think carefully before you start drawing graphs!

Understand:

- The type of data to be presented
- The key feature to be portrayed
- How the information will be used
- Your intended audience

The following sections should help you to design a graph that is easy to understand and difficult to misinterpret!

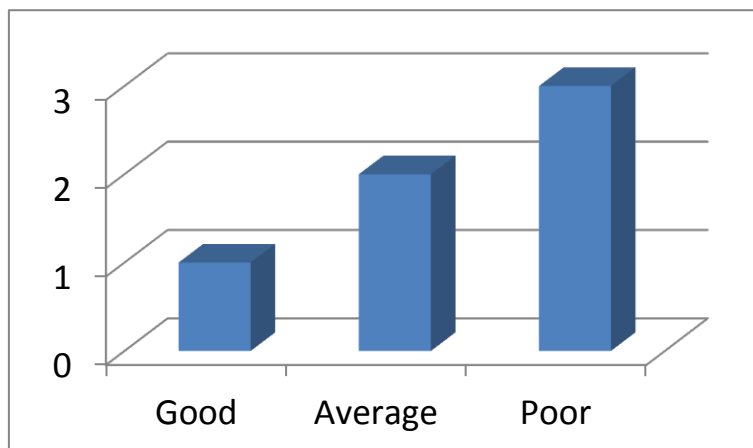
#### 3-D graphs

The general guidance on 3D graphs (which should also become law) is **DO NOT USE THEM** (unless you are presenting three dimensions of data)!

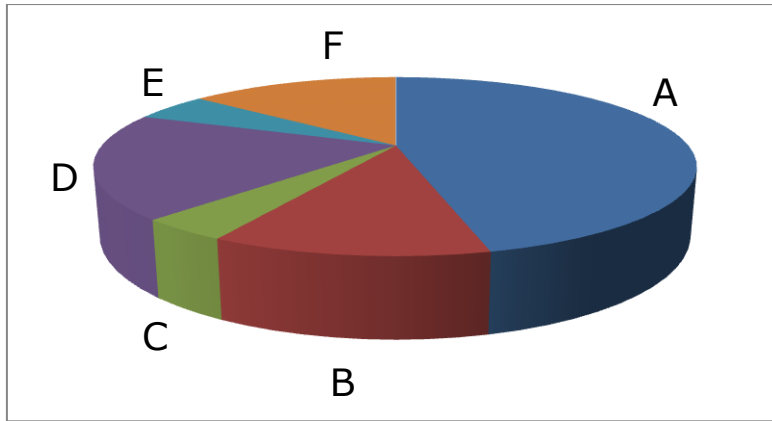
Take the following 3D bar chart for example: the data used to build this graph was Good = 1, Average = 2, Poor = 3.

*All three bars are below the gridlines representing 1, 2 and 3.*

*The front and side areas are proportional, but the top of each bar has the same area. This makes the first block appear relatively bigger.*



If you haven't got it already, here's another hideous example:



Segment B represents a proportion which is smaller than segment F.

The additional depth in the front segments and the angle that the chart is set at makes the apparent area of the segments different to the actual proportions they represent. Table 5 shows the apparent area and actual proportions

**Table 5: Apparent and actual proportions shown by the pie chart above**

	Actual	Apparent
A	46%	43%
B	12%	22%
C	4%	7%
D	19%	17%
E	5%	3%
F	14%	9%

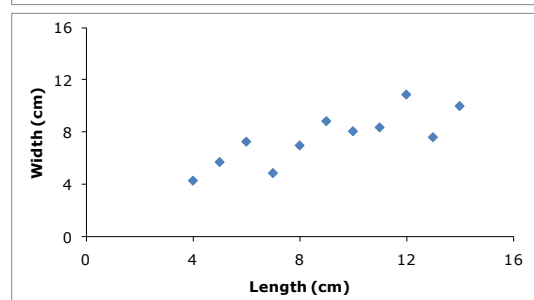
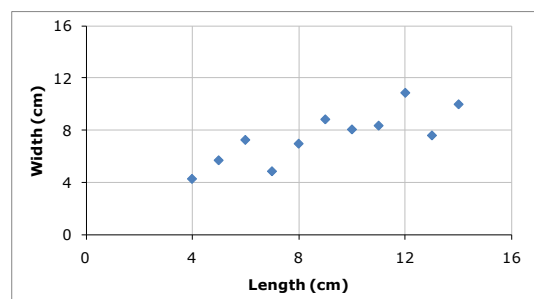
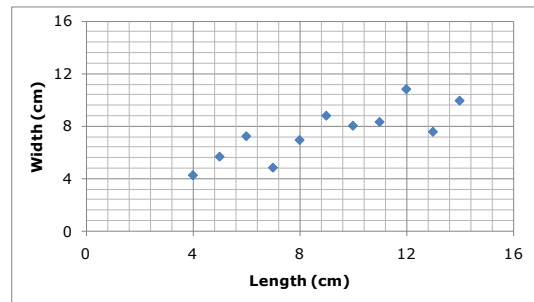
## Gridlines

Gridlines are lines which divide graph area into segments. The idea is that they help to define the location of the data points. The important message in a graph should come from the data itself, not 'graph furniture', so a balance has to be struck between helpful guidance and distracting from the data.

The graphs show some different choices of the use of gridlines. A compromise between too many (top) and no gridlines (bottom) seems to be the best option in this case.

The general guidelines for gridlines are:

- Use only as many that are needed to get an approximate idea of the value of any given data point in chart
- The axes of the graph should be heavier than the grid lines



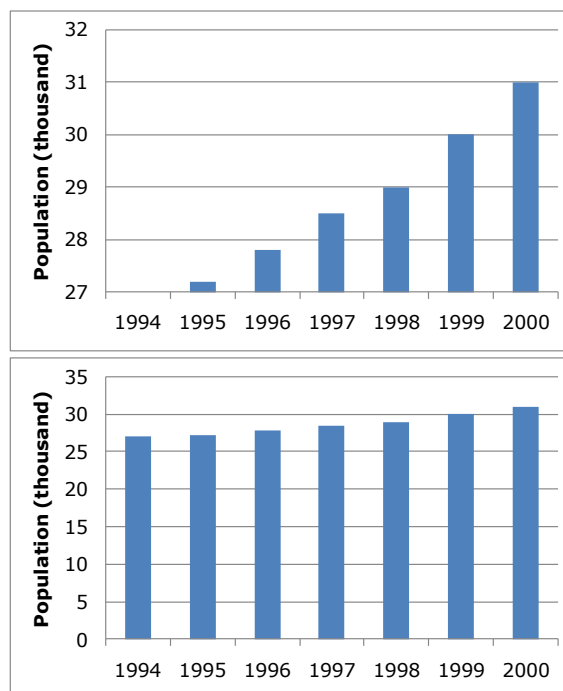
## Axes

The scale on the axes of graphs is usually the part which causes the most confusion and is the part which leads to misinterpretation.

The graphs on the right show the same data – the population of a town in Berkshire from 1994 – 2000. The top graph looks far more interesting, but without taking proper notice of the scale on the y-axis (and most readers won't), you can make the following interpretations:

- The population in 1994 was 0
- The population doubled from 1998 to 2000

When displaying numbers that need to be interpreted as **absolute numbers** you need to **start your axis at 0**.

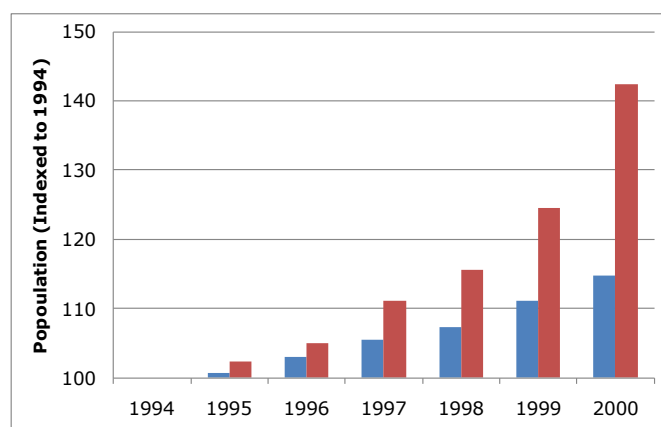


For **relative comparisons** between different series consider indexing your numbers to a baseline. Each index value is computed by dividing the relative population figure by the baseline figure, which, in this case, is the population in 1994.

**Table 6: Population (real in thousands and indexed to 1994) of a town and a city in Berkshire from 1994 to 2000.**

Year	Town		City	
	Population	Index	Population	Index
1994	27.0	100	562	100
1995	27.2	101	575	102
1996	27.8	103	590	105
1997	28.5	106	625	111
1998	29.0	107	650	116
1999	30.0	111	700	125
2000	31.0	115	800	142

**Figure 1: Relative population figures of a town and a city indexed to 1994 values**



## Colour

Colour will obviously enhance your graph, but use it carefully so that it improves the interpretability, rather than just making it look pretty!

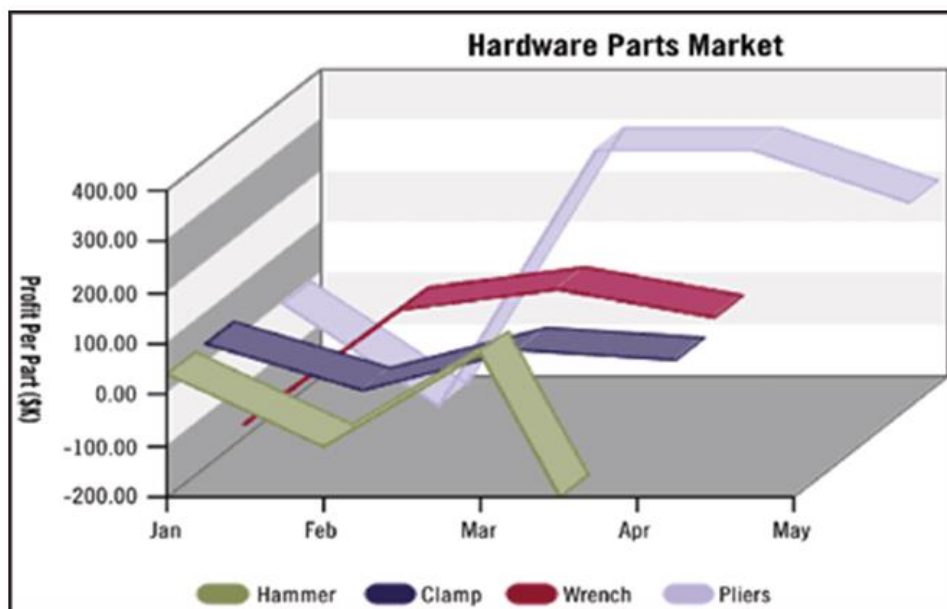
## Principles for graphs

1. Never use 3D (unless you have 3D data)
2. Use colour sparingly – shades of a single colour are often more effective
3. Always put the zero point on the scale when graphing absolute numbers
4. Label diagrams effectively
5. Show gridlines to aid interpretation
6. Round the numbers on the axes appropriately
7. Ensure the picture accurately represents the data
8. Make the diagram simple enough so that the reader can quickly assimilate the message

## Examples of bad graphs

### Example 1:

**Figure 2: Visual mining example**

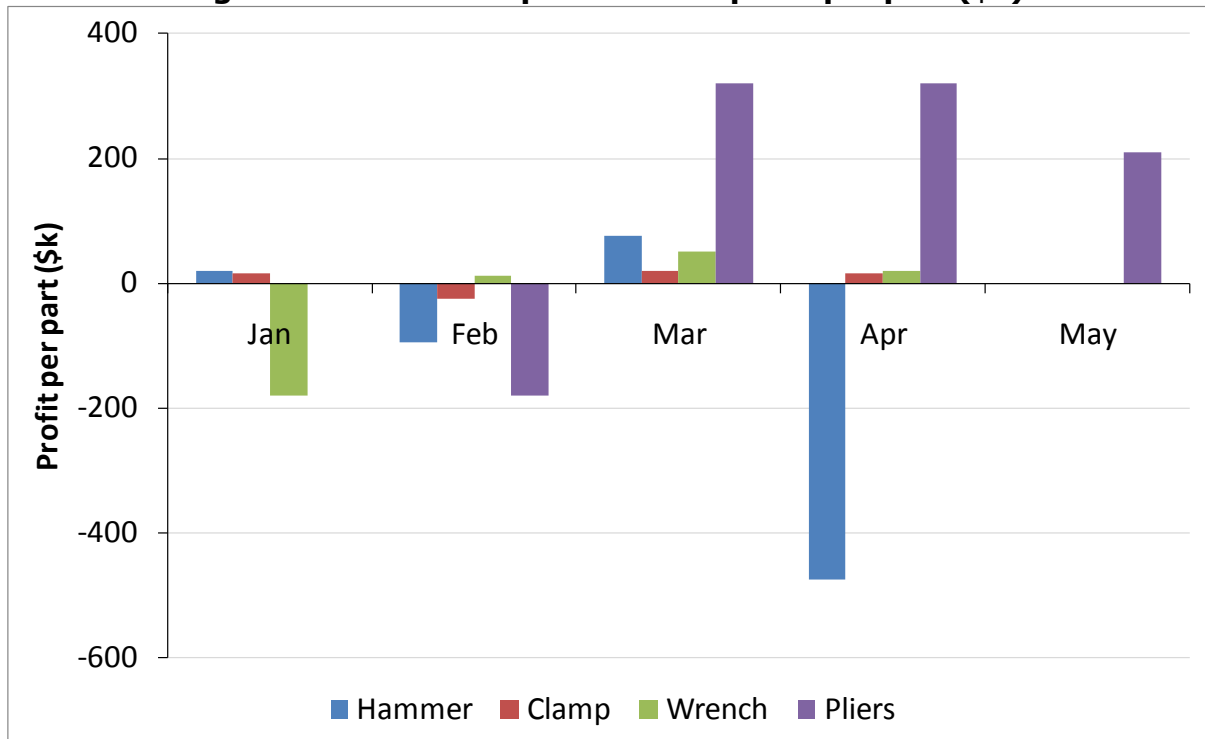


Does this give you a good idea what is happening to the profits of the different hardware parts in relation to each other over time?

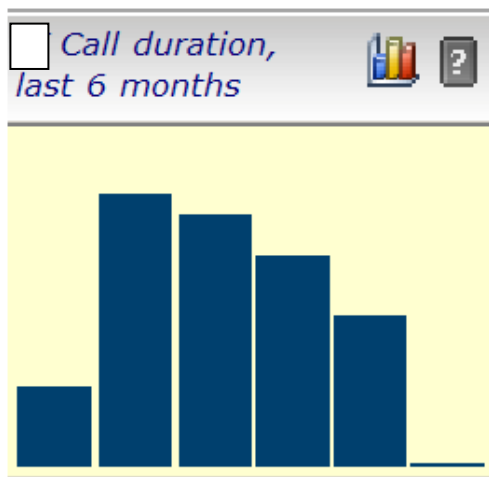
- Firstly, it's in 3D and it's virtually impossible to tell where the lines are in relation to each other and the axes.
- There are some missing values in May for which there is no reason given.
- For hammers, it isn't clear whether the final measurement was taken half way through March instead of April, or whether they chose to shorten the axis.
- Showing 2 decimal places in the y-axis is unnecessary and shows a false sense of accuracy

Possible improvement

Figure 2a: Hardware parts market profit per part (\$k)



Example 2:



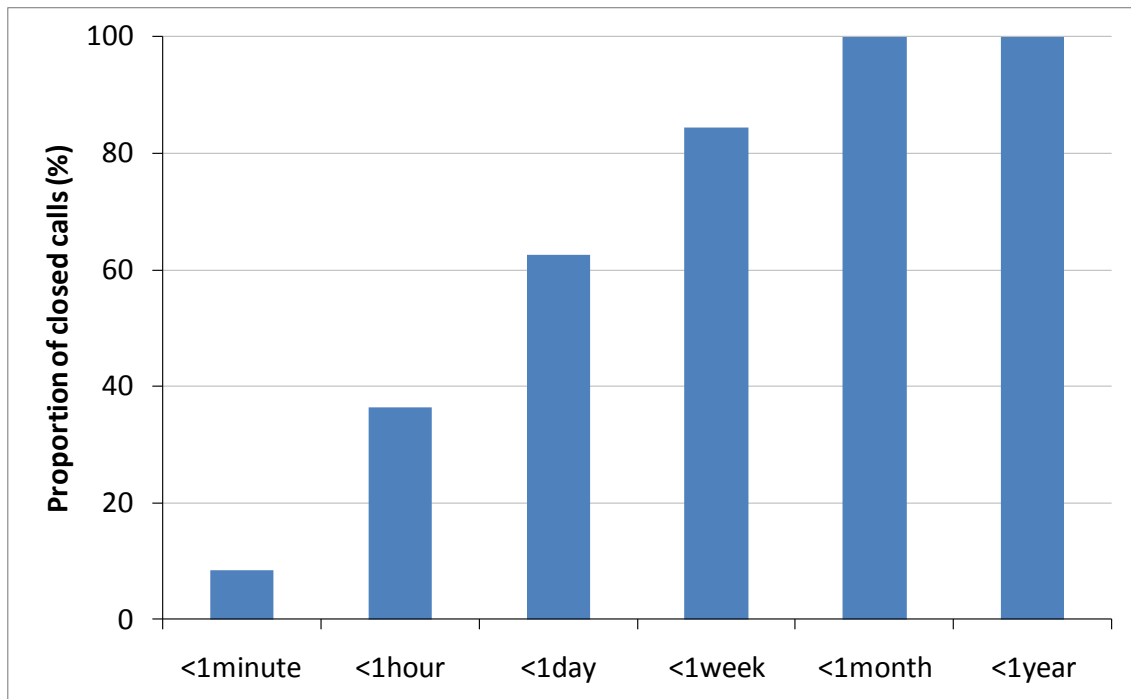
You might recognise this!

Note the lack of axes and also the lack of information about the number of open calls. Once you discover what the axes and labels are, you'll also discover that the percentages add up to 96% and the proportion of calls that are closed in less than 1 month is less than the proportion of calls that are closed in less than 1 week.

### Possible improvement

This graph presents the cumulative distribution of closed calls. Each time category includes the closed calls in the category before. For example, the proportion of closed calls that were closed in less than 1 day includes the calls which were closed in less than one hour.

**Figure2: Cumulative distribution of duration of closed calls over 6 months**



Alternatively to present a graph with a similar pattern to that above (i.e. non-cumulative), the time categories need to be

- $\leq 1$  minute
- $> 1$  minute and  $\leq 1$  hour
- $> 1$  hour and  $\leq 1$  day
- $> 1$  day and  $\leq 1$  week
- $> 1$  week and  $\leq 1$  month
- $> 1$  month

## 4. Displaying Categorical Data

Categorical data consist of names or labels. The values can be ordered (ordinal data) or not (nominal data).

For example:            Motorways where accidents occurred  
                              Type of motorbike;  
                              Accident involvement;  
                              Age and gender.

Commonly used graphs are:

- Bar charts
- Pie charts

### Bar Charts

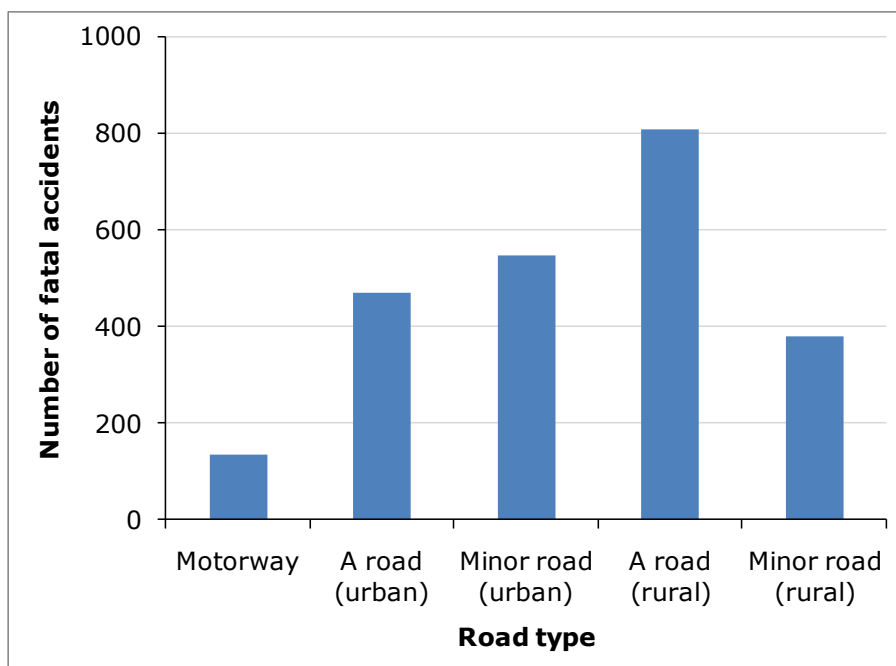
#### Graphs → Bar

The bar chart presents data in the form of bars or columns, arranged either horizontally or vertically on a scale of values. Most often there are two scales, one for the group or class and the other for the percentage or number.

A **simple bar chart** shows how several items differ from each other in a single characteristic. The best arrangements of bars for emphasis and ease of comparison are in order of size, either descending or ascending, although the nature of the data may require some other order. There is no measurement value to the thickness of the bars and, therefore, they should be uniform throughout the chart.

#### Example:

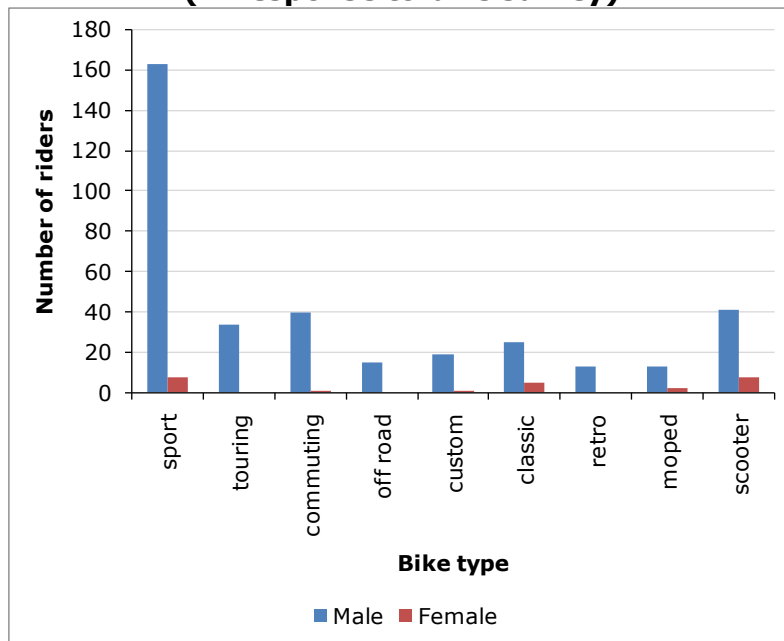
**Figure 4: Number of fatal accidents on different road types in 2008**



A **clustered bar chart** can be used to display a further categorical variable such as sex.

**Example:**

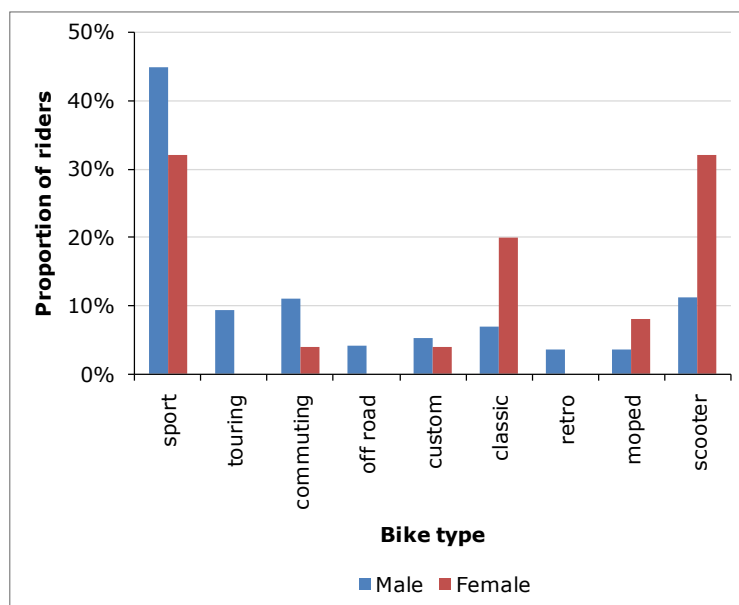
**Figure 5: Type of bike ridden in the last 12 months by sex of rider (in response to bike survey)**



But female numbers are small, making them difficult to read. Rather than displaying counts it may be better to display percentages.

For example, we might be interested in comparing the distribution of bikes across the two sexes.

**Figure 6: Type of bike ridden in the last 12 months by sex of rider (response to bike survey). 365 male responders, 25 female responders**



### General guidance for bar charts:

- Bar charts are useful for discrete, grouped data of ordinal or nominal scale
- When comparing more than one variable, group the bars to be compared (sex is the variable of interest in the example above)
- Include the zero point on a scale, otherwise 40 and 20 are not proportionate e.g.
- Don't overlap bars – the visible area is smaller for overlapped bar
- Stacked bar charts are good at displaying variation in the first category and the total but not other categories.

## Pie Charts

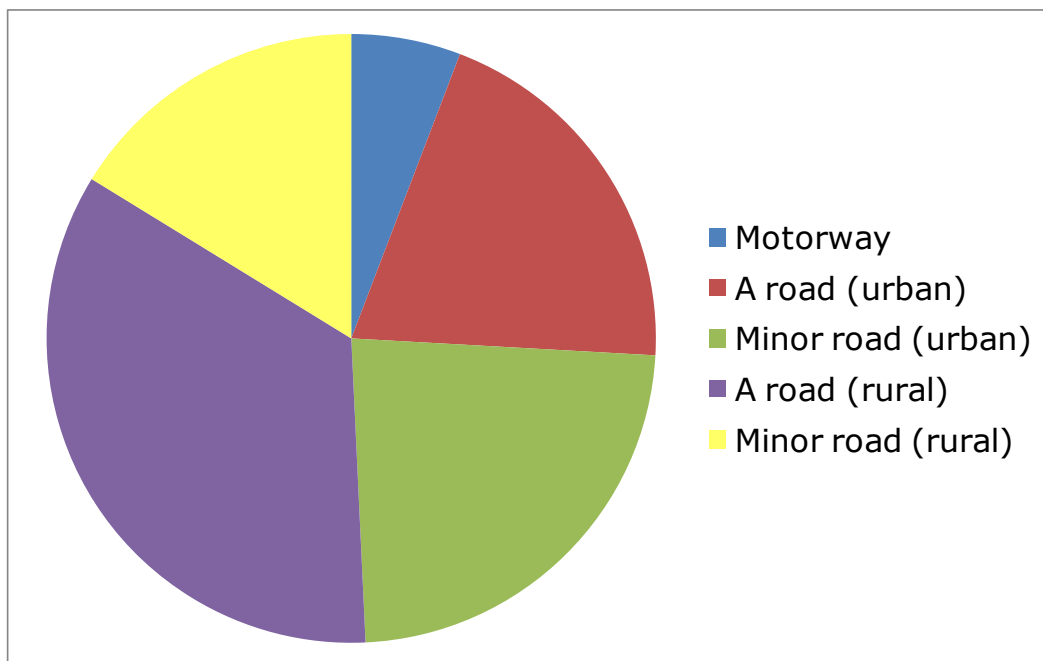
### Graphs → Pie

A **pie chart** is a circular chart divided into sectors and together the sectors equal 100%. A pie chart is useful when comparing the size of a slice with the whole pie but not so useful when comparing one slice with another (a bar chart would be more effective). The optimal number of slices is around 6.

### A PIE CHART SHOULD ALMOST NEVER BE USED

### Example:

**Figure 7: The proportion of fatal accidents by road type in 2008**



### General guidance for pie charts:

- You have to have a really good reason for using a pie chart in your report! Often a bar chart will present your data in a more meaningful way.
- Pie charts are effective for presenting a small number of pieces of data
- Pie charts are useful for different sized segments
- They should only be used when the total sum of all the segments has meaning
- The optimal number of segments is 6 (between 3 and 10 is advised)

## 5. Displaying Numerical Data

For example: Journey time from London to Dover;  
Length of a motorway delay due to an accident;  
Measurements (e.g. Stopping distance, speed, weight, )

Commonly used graphs are:

- Histograms
- Scatter plots

### Histogram

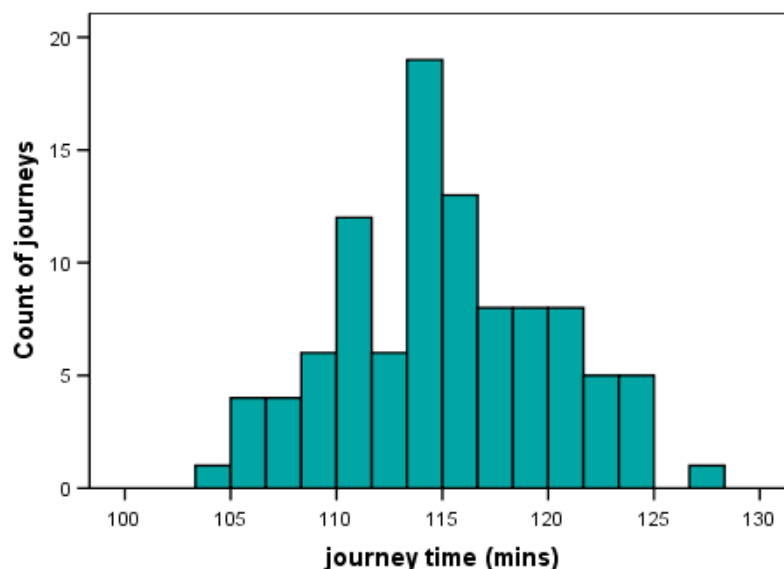
#### Graphs → Histogram

A **histogram** is used to plot continuous data which have been put into a limited number of distinct groups or classes. A histogram differs from a bar chart in that it is the area of the bar that denotes the value, not the height, a crucial distinction when the categories are not of uniform width.

How many groups should you choose for a histogram? If you choose too many, the display will be too fragmented to show an overall shape. But if you choose too few, you will not have a picture of the shape and too much of the information in the data will be lost. SPSS automatically allocates widths to groups or the user can define them.

#### Example:

**Figure 8: 100 journey times from London to Dover**



#### General guidance for histograms:

- Plot continuous data with histograms
- Choose bar widths carefully
- If absolute comparison are necessary, start the axes at zero to avoid misinterpretation

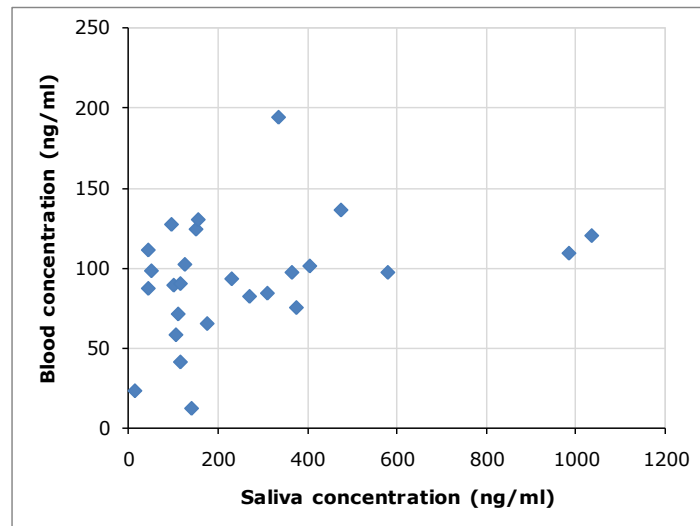
## Scatter plots

### Graphs → Scatter/Dot

In a **scatter plot**, one variable is plotted on the horizontal axis (x-axis) and the other on the vertical axis (y-axis). They are useful for investigating associations between two numerical variables.

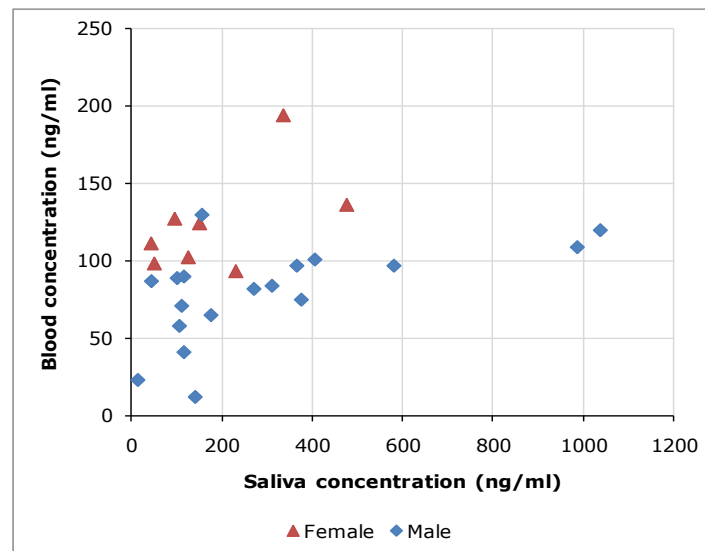
**Example:** Study of cannabis impairment using a simulator:

**Figure 9: Plot of blood concentration against saliva concentration for 26 subjects**



Gender information may also be added to the scatterplot.

**Figure 10: Plot of blood concentration against saliva concentration for 26 subjects, by sex**



### General guidance for scatter plots:

- Compare two continuous variables with a scatter plot
- The point at which the axes cross should almost always be (0,0) (SPSS does not do this well)

## 6. Displaying Time Series Data

A time series is a sequence of data points which are ordered in time.

For example:           Unemployment figures;  
                              Road accidents;  
                              New car registrations.

Commonly used graphs are:

- Line charts

### Line chart

#### Graphs → Line

A line chart is a series of data points connected by a line and should be used for continuous data and time series.

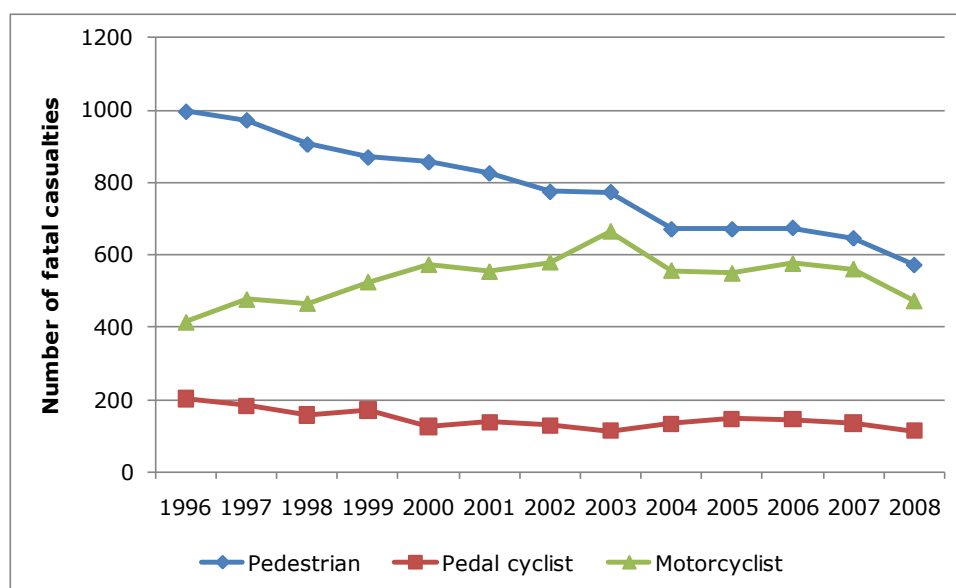
In theory continuous lines should not be used to display non continuous data, such as annual counts of casualties. In practice, there are often many data points to present and a line is the most obvious and efficient solution. We suggest adding points to the line at the points where the data exists, so that it is clear that the graph is based on discrete measurements.

Multiple lines can be plotted to provide comparisons. However, it is not recommended to plot more than four or five lines on a chart, especially if the different data points for the lines are close in value.

The selection of the range of the scale on the Y- axis (vertical axis) is very important. Excessive range minimizes differences in the data. Too small a range may result in an exaggerated picture of the differences.

#### Example:

**Figure 11: Number of casualties killed in road accidents from 1996 to 2008, by road user type**



### *General guidance for line graphs:*

- If the line represents point measurements, add points to the lines
- The optimal maximum number of lines on one chart is 3-4, unless they are well separated
- In a set of line graphs use consistent plotting styles and colours and scales

## References

University of Reading Statistical Services Centre. Informative Presentation of Tables, Graphs and Statistics. [www.reading.ac.uk/ssc/publications/guides/toptgs.html](http://www.reading.ac.uk/ssc/publications/guides/toptgs.html) accessed 04/08/10

Albert Goodman. Graphical Data Presentation. [www.deakin.edu.au/~aggodman/sci101/chap12.php](http://www.deakin.edu.au/~aggodman/sci101/chap12.php) accessed 04/08/10

Local Government Data Unit – Wales (2004). Presenting Data. Performance Management Support Portfolio 5.3. [http://www.dataunitwales.gov.uk/Documents/Publications/adviceandguidance/CPS15010\\_Presenting\\_Data\\_090722\\_FINAL\\_eng.pdf](http://www.dataunitwales.gov.uk/Documents/Publications/adviceandguidance/CPS15010_Presenting_Data_090722_FINAL_eng.pdf) accessed 04/08/10

Department of Chemical and Process Engineering. Presenting Data. <http://lorien.ncl.ac.uk/ming/dept/tips/writing/data.pdf> accessed 04/08/10.

Phillip Good and James Hardin (2009). Common Errors in Statistics (and how to avoid them). 3<sup>rd</sup> edition: Wiley, USA.