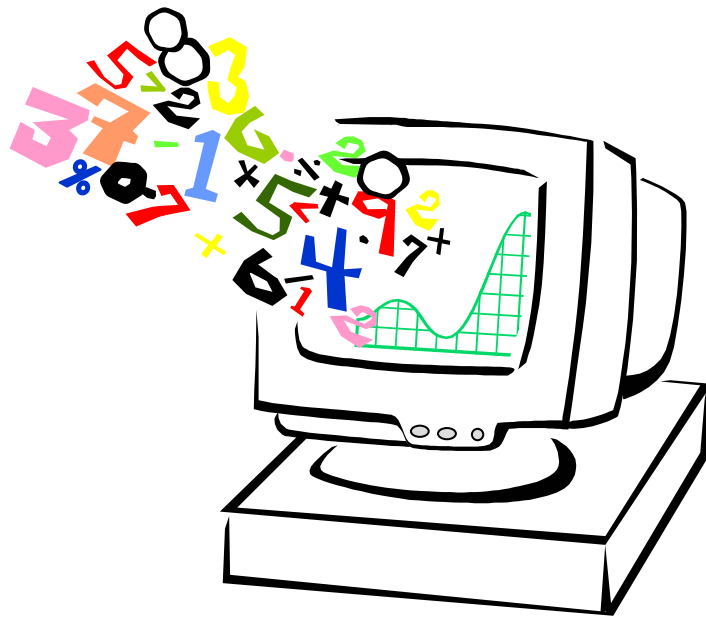




Sheffield Hallam University

Faculty of Health and Wellbeing
Professional Development 1
Quantitative Analysis



Student companion

This booklet is designed as a companion to the Course Document.

It consists of a set of answers and comments on the tasks.

It works better if you read through it *after* doing an activity, and don't use it instead of thinking carefully about the work or discussing it with other students and your tutor.

It may prove useful if you have had to miss some sessions, but is not a good substitute for them.

Comments on the Tasks with Answers.

Task 1

Entering and saving Data.

By the end of the task you should have saved a file you created in SPSS – to check this have a go at closing SPSS then restarting SPSS and opening the file.

Task 2

This task was designed to help you to focus on

- a) measures of level: mean and median
- b) measures of spread: interquartile range, range and standard deviation.

By the end of the task you should have a good idea of what these measures tell us about a set of data and how they help us to compare different sets. You should also understand how a Boxplot gives a quick visual picture of median, range and interquartile range, as well as showing outliers.

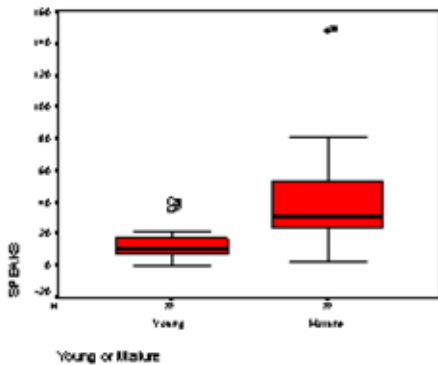
Looking at Data:

- Questions might be: do mature students contribute more in class than younger students? Are second year students different from first year students? Are all mature students similar or do they differ among themselves?
- Each entry in the data is the number of times a particular student spoke in the 12 hour period, so each entry corresponds to a different student. The three columns are *speaks* (the number of time they spoke), *age* (1 = young 2 = mature) and *year*. There are 12 young first years, 11 mature first years, 10 young second years and 11 mature second years.
- From looking at the data it might be clear that the mature students tend to talk more than the young ones.
- Whether general conclusions can be drawn is open to discussion. The sample only came from one university and one course, which may attract mature students of a different kind from other universities, and courses. However conclusions could be

drawn about the likely character of young and mature students in future intakes at this university, provided admissions policy remained similar.

Drawing Boxplots

You need to be clear that a Boxplot shows the median, the quartiles and the extremes, as well as showing outliers separately.



It should look like this.

The extreme outlier corresponds to one of the second year mature students who spoke 148 times. Because this is so different from the others in that group, it has been drawn as an asterisk, otherwise you might think several other students were close to speaking that much.

This plot tells us that, on average, mature students speak more than young students.

The interquartile ranges, (size of the box) shows that the group with most variability are the mature students.

The visual display may have helped you to pick out differences that might not be obvious otherwise. For instance, all the mature students say something: it is only the young students group that includes individuals who did not talk at all.

The range and interquartile range are both higher for the mature students. Note that the extreme outlier among the second year mature students has a huge effect on the range.

Using Descriptive Statistics

You should get this:

Statistics		
SPEAKS		
N	Valid	44
	Missing	0
Mean		26.39
Median		19.50
Mode		7 ^a
Std. Deviation		26.616
Range		148
Minimum		0
Maximum		148
Sum		1161
Percentiles	25	8.25
	50	19.50
	75	35.50

a. Multiple modes exist. The smallest value is shown

It is good practice to use descriptive statistics to get an overall view of the data before further analysis.

N is the number of students in each group. Minimum is the lowest value and Maximum is the largest. Percentile 25 is the lower quartile (Q1) and Percentile 75 is the upper quartile (Q3). The range is the Minimum taken from Maximum for each group. To find the interquartile range, subtract the lower quartile from the upper quartile for each group.

The mean is another measure of central tendency or level. It is often very similar to the median, but is calculated differently. In some circumstances it is preferable to use the mean, in others the median. (You will learn more about this later.) The standard deviation

is a measure of spread, like the interquartile range. If one dataset has a larger interquartile range than another, it will usually also have a larger standard deviation.

The mean and median are very different in the group of second year mature students. This is because outliers can have a strong effect on the mean, and this group has an outlier.

The note at the bottom of the Statistics table says, "Multiple modes exist. The smallest value is shown" The mode is the most frequently occurring data value. In small sets of data it is quite simple to pick out by eye.

It is possible for a data set to have several modes. In the example we have seen in the students data, both Young and Mature student populations have more than one mode, i.e. there isn't one most frequently occurring number. When this is the case SPSS chooses the lowest and adds a footnote.

Young or Mature		Statistic	Std. Error	
SPSS Young	Mean	12.77	2.100	
	95% Confidence Interval for Mean	Lower Bound	8.27	
		Upper Bound	17.28	
	5% Trimmed Mean	11.20		
	Median	11.00		
	Variance	100.202		
	Std. Deviation	10.100		
	Minimum	0		
	Maximum	40		
	Range	40		
	Interquartile Range	10.75		
	Skewness	1.263	.491	
	Kurtosis	2.020	.399	
	Mature	Mean	40.00	0.000
95% Confidence Interval for Mean		Lower Bound	38.27	
		Upper Bound	59.73	
5% Trimmed Mean		38.40		
Median		31.50		
Variance		250.043		
Std. Deviation		15.816		
Minimum		2		
Maximum		143		
Range		141		
Interquartile Range		29.25		
Skewness		2.125	.491	
Kurtosis		8.460	.399	

The "Explore" feature lets you calculate the statistics again but separately for each group (Young and Mature).

Mean (Young) __ 12.77

Mean (Mature) __ 40.00

Median (Young) __ 11.00

Median (Mature) __ 31.50

The Effect of Errors

Young & Mature		Statistic		Std. Error
SPEAKS	Young	Mean	15.23	9.42 [†]
		95% Confidence Interval for Mean	3.11	
		Lower Bound	22.34	
		Upper Bound	19.19	
		5% Trimmed Mean	11.00	
		Median	297.51 [†]	
		Variance	18.047	
		Std. Deviation	0	
		Minimum	11	
		Maximum	11	
		Range	12.00	
		Interquartile Range	2.221 [†]	
		Skewness	0.475	
		Kurtosis	0.000	
	Mature	Mean	40.00	9.800
		95% Confidence Interval for Mean	28.21	
		Lower Bound	50.79	
		Upper Bound	38.49	
		5% Trimmed Mean	31.30	
		Median	250.043	
		Variance	90.803	
		Std. Deviation	2	
		Minimum	143	
		Maximum	148	
		Range	29.25	
		Interquartile Range	2.125	
		Skewness	0.480	
		Kurtosis	.000	

Mean (Young) __ 15.23

Mean (Mature) __ 40.00

Median (Young) __ 11.00

Median (Mature) __ 31.50

You should get this:

You should notice that the median and quartiles are not much affected (so the interquartile range would also not be affected), whereas the error does have an effect on the mean, the standard deviation and the maximum (so the range would be affected.) This tells you that if you suspect there may be errors in your data and you cannot easily

correct them, the median and interquartile range are better measures of level and spread to use than the mean and standard deviation.

Different types of data

The data are ratio. It makes sense to say that one student talks twice as much as another if their entry is twice the other's.

The difference between Mean and Median

This example is intended to show when differences between the mean and the median can be important and what causes these differences.

Presumably, the director earns 100,000.

The mean is 16,300 and the median is 7,000. The manager would quote the mean, to show that employees were 'on average' well paid, and the union negotiator would quote the median. Note that in this case although the mean is an average, it is not representative of most workers' earnings.

The interquartile range is 0, because both the quartiles (25th and 75th percentiles) are 7,000. The standard deviation is 29,409 because the director's salary has affected it. The Boxplot will look very strange: a line at 7,000 which is the box and the whiskers since the minimum, lower quartile, median, upper quartile and maximum apart from the outlier are all 7,000. The director's salary will be shown as an outlier at 100,000.

The use of the mean, median or mode as the average can be used to give a desired impression of the data, sometimes the type of data will dictate the type of average which is best suited.

A very trivial example of Mean, Median and Mode.

Three groups of school children were given the task by their teacher of finding out the average colour of staff cars.

They all chose a simple (but possibly flawed) sampling method; they looked in the car park. For our purpose this doesn't matter, but you may reflect on why I feel the method could be flawed.

Colour	Number
red	2
green	3
black	2
white	2
blue	1

On that day the car park was populated as shown in the table.

Group one decided to find the average using the mean, they couldn't do the maths so they mixed paint in the art room, using one spoonful each coloured powder paint to represent each car. So they had two spoons of red, mixed with three green and so on. They then added water and painted a picture of the average car. It was a sludgy brown. None of the teachers had a brown car - although some were a bit rusty.

Group two used the median, but couldn't do the maths so they got out ten model cars of the same colours as the teachers and lined them up neatly in order of brightness. There was considerable argument about which colours were brightest but eventually the order was agreed as; white, white, red, red, blue, green, green, green, black, black. This caused confusion because there was no middle car. They eventually decided on turquoise. No teacher had a turquoise coloured car.

Group three chose the mode. The most frequently occurring observation was green. The most common car colour was green.

Thoughts on - An example:

The problem with using the mean on the data for this application is that a relatively small number of older children will increase the mean disproportionately.

If we want to convey a general figure for the age of adoption it might be better to either say a more general statement like "well over half the children adopted in 2003 were between the ages of one and four" this succinctly paints a picture of the figures, alternatively we could use the median rather than the mean, this would combat the tendency for the small number of much older children to skew the average higher.

Age at adoption	2003
Under 1	240
1 to 4	2,100
5 to 9	1,000
10 to 15	180
16 and over	10
Average age	4 years 3 months

*Have a look at the available summary of the data in the table, which type of average would the relatively small number of older children have the greatest effect on? (**Mean**)*

*What type of average do you think would be best for this type of data? (**Median**)*

Task 3

Standard Deviation (S.D.) using SPSS.

This task is intended to show what the Standard Deviation is measuring and what this can tell us about our data?

Use SPSS to work out the MEAN, MAX & MIN, remember From the Analyze menu select Descriptive Statistics then Frequencies. There are other ways of getting descriptive statistics in SPSS but this is the easiest for now.

	German (%)	Geography (%)	IT (%)
MEAN	33.3	33.3	33.3
MAX	67	42	67
MIN	11	26	11

- 1 Which set(s) of figures has the largest range? (Answer – German & IT)
- 2 Which set(s) of figures has the largest number in it? (Answer – German & IT)
- 3 Which set(s) of figures contains the smallest number? (Answer – German & IT)
- 4 Which set of figures has the largest minimum? (Answer – Geography)

When you've worked out the values for the standard deviation you are asked to answer the following questions;

The values I got for the data are:

	German	Geography	IT
S.D.	19.044	4.877	13.849

For a variable with the value 33.3 in each of 10 cases, the total should be 333, the mean, median, maximum and minimum should all be 33.3, the standard deviation should be 0, the values don't deviate from the mean at all.

- 1 Which of the three sets of figures, German, Geography, IT, is the least spread out?
- 2 Of the two subjects with the same mean, and the same range, which varies least?
- 3 Which of the three sets of figures, German, Geography or IT, varies most?

I think the answers are:

- 1 Geography is the least spread out.
- 2 Of the two subjects with the same mean, and the same range, IT varies least.
- 3 German varies the most.

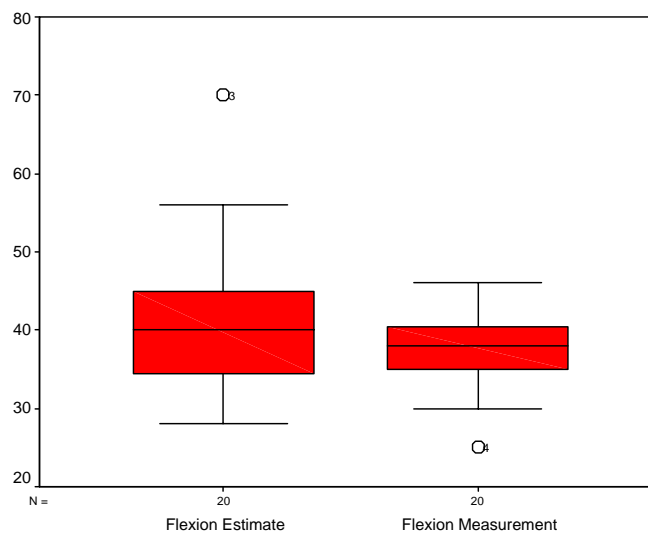
Comparison of Visual Estimations and Mean Goniometric Measurements of wrist flexion and wrist extension.

Statistics

		Flexion Estimate	Flexion Measurement
N	Valid	20	20
	Missing	0	0
Mean		42.25	37.40
Median		40.00	38.00
Std. Deviation		10.172	4.999
Range		42	21

- Which column of flexion results (estimated or measured) appears most varied? The adjacent results would lead me to say that the estimate is more variable.
- Was the tendency to underestimate or overestimate the flexion? The above results show a slight over estimation, but it is quite a small difference and may be due to chance.

The Boxplot for the two variables allows a visual comparison of the level and spread.



Thoughts on - A Simple example: In this rather unlikely example, If the mean heights and Standard deviations were as follows;

Town	mean	Standard deviation	
Youngville	175cm	5.25	<input type="checkbox"/>
Oldton	169cm	15.50	<input checked="" type="checkbox"/>

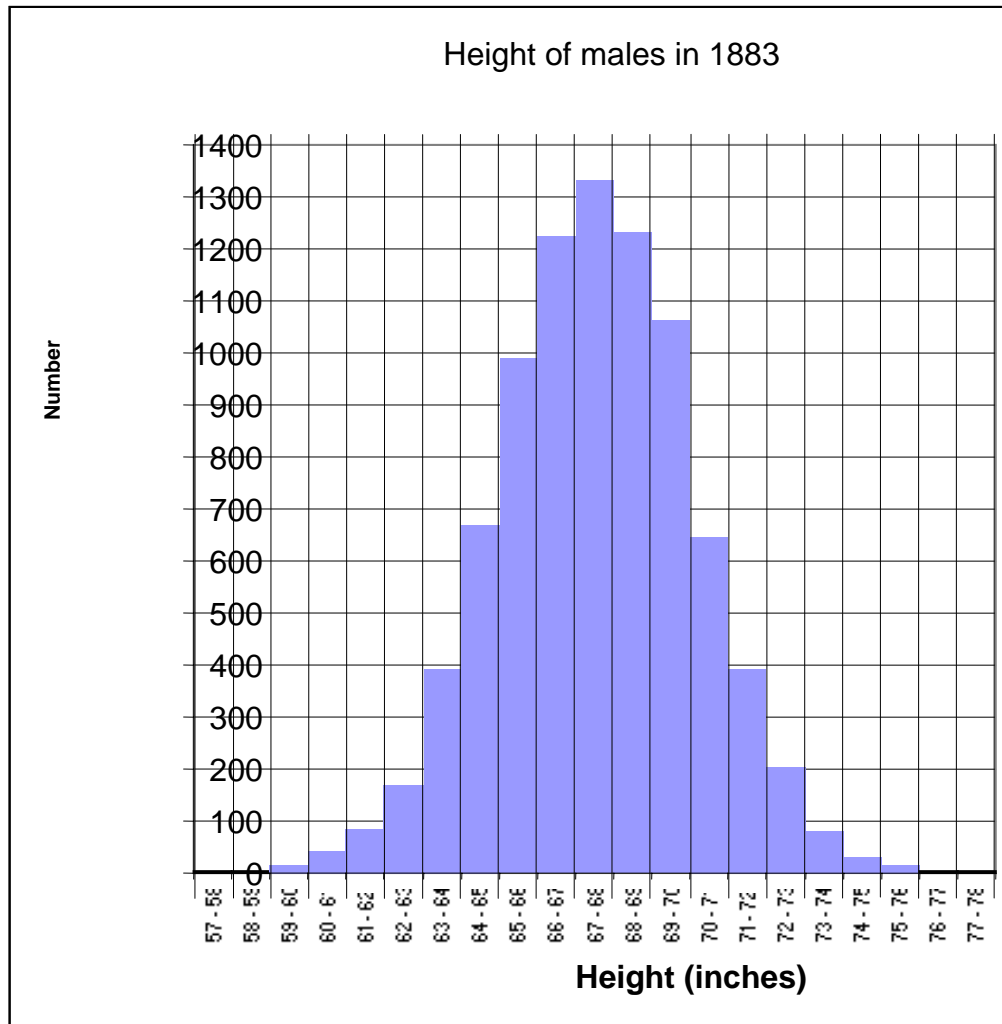
Which sample varies most?

The sample from Oldton seems more varied - it does perhaps lead us to think there are some differences in the samples other than the people in one town being taller.

Task 4

Histograms and the Normal Distribution

How many people were between 64 and 65 inches tall? 669



When you've completed all the bars you should have a reasonably good example of the bell-shaped Normal distribution.

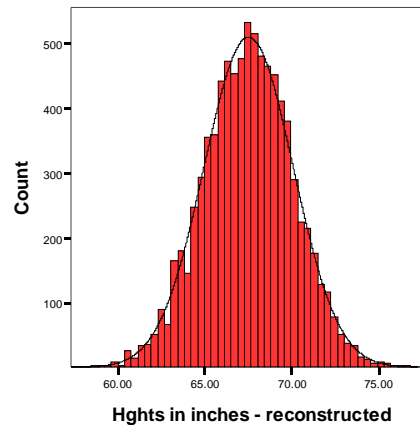
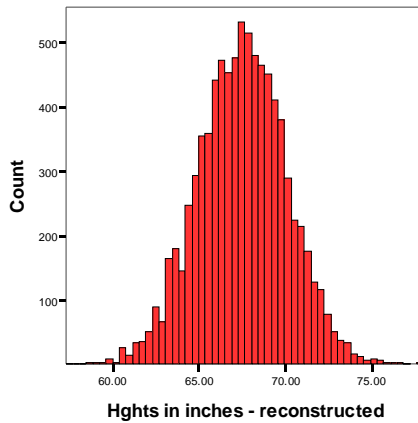
Drawing the same graph in SPSS.

If you get strange results in the output it is worth pressing the **Reset** button in the Create Histogram dialog, to prevent the scales from previous data being used.

You should see a normal (bell shaped) pattern to the distribution of the data. This is typical in many natural distributions. The majority of subjects are clustered round the mean and the numbers of individuals in the categories more distant from the mean is far less, in this example there are less very tall or very short males.

The graphs below show the output you should see if you follow the instructions. The first two are the histograms without and then with the normal curve superimposed.

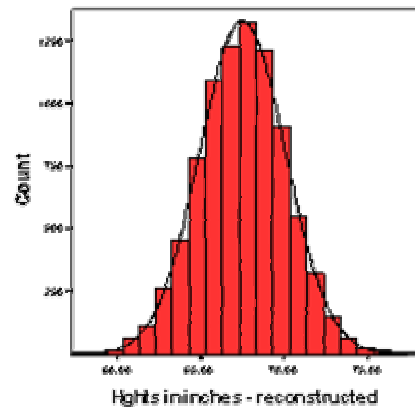
Is the data we are looking at here Discrete or Continuous? It is Continuous.



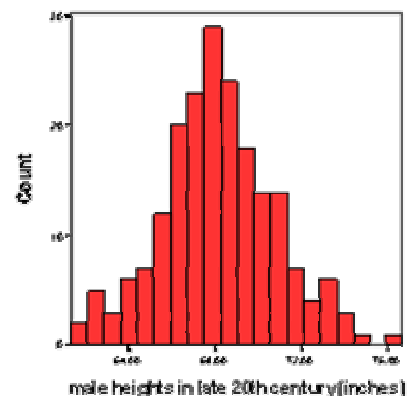
The graph adjacent shows the results of manually setting the interval size. It is similar to the one you created in the first part of this task.

Compare 19th to 20th century heights.

Are people getting bigger? (I really should have said taller!) I could not say from this evidence one way or another, however the second graph, it does appear that there is a skewing of the data showing a larger number of taller men.



Histograms give more information than a Boxplot about how data is distributed, but don't allow as clear a comparison of level and spread. They are particularly useful in looking for a normal (bell-shaped) distribution. Biological data: e.g. heights of 20 year old men, weight of 2 year old girls, foot-lengths of 25 year old women, tend to have this sort of distribution. A normal distribution must be symmetrical: rough bell-shapes with an obviously greater tail on one side than the other indicates a non-normal distribution. The height distributions we have seen are roughly normal.



Examples I would expect to be normally distributed:

Normally distributed?	Yes	No
Ages of people in a town.		✓
Heights of 20 year old men.	✓	
Weights of one-year-old squirrels.	✓	
The price of drinks in a bar.		✓
The life (in total hours switched on) of light bulbs.	✓	

Examples of normal distributions could include:

- the life (in hours switched on) of a type of light bulb – although it would probably only be true for a given bulb type from one manufacturer.
- Weights of one-year-old squirrels, it may be more clear if they were all from one area – different areas may have better nutrition available.
- Heart rates or body temperatures (would we need to specify healthy specimens at rest?)

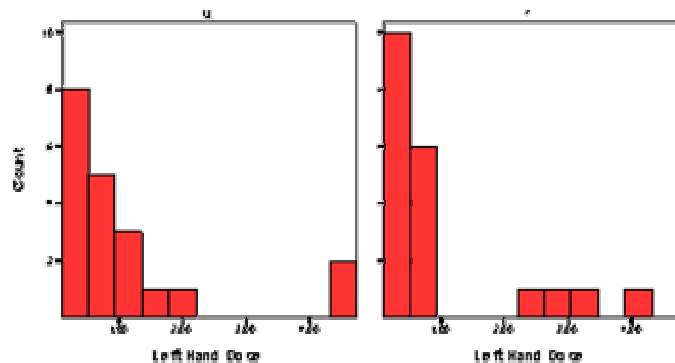
Examples of non-normal distributions could include:

- ages of people in a town (highest frequencies among the youngest, going down gently with a long tail at the very oldest),
- ages of people in a baby clinic (high frequencies at 0-1, low frequencies from 2 to 5, very low frequencies from 5 to 16, moderate frequencies from 16 to 40, low frequencies from 40 upwards)
- Birth-weights of babies (a rough bell-shape with a longer tail at the bottom end of the range, showing very premature babies).

An example of data with a discriminatory variable in:

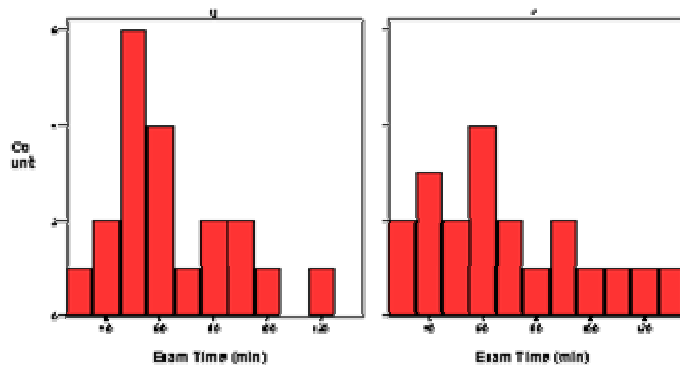
What do the histograms show us about the data?

The small sample size makes it difficult to draw conclusions, however it would appear that the screen has increased the number of radiologists receiving a lower left hand dose.



The examination time also appears to be altered, more examinations appear to be taking longer.

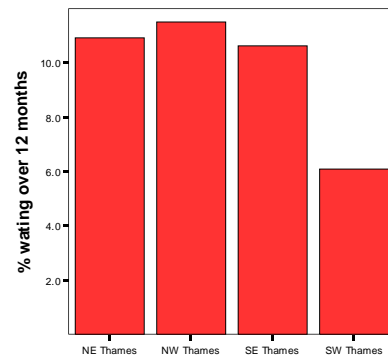
If you want to examine the data more it is worth playing with the Interactive Boxplot feature in SPSS.



Of the following which would best be displayed in a Histogram or Bar chart. Fill in the table below, put H for Histogram or B for Bar chart in the end column.

	H or B
The number of students in the age groups 18-27, 28-37,38-47 etc. This is continuous data, people can have any age in a continuous range - hence use a histogram.	H
The number of people living in each of three towns. This is not continuous data, it is discrete - use a bar chart.	B
The number of patients visiting an Optician with short sight, long sight and no sight defect. This is discrete - the data is giving the number of patients in each of three categories. Use a bar chart.	B
The marks of each individual student in a class.	B
The number of students in each range of marks in 10% intervals.	H
The number of men vs. women in a town. This is certainly discrete not continuous data - you could use a bar chart or in this case a pie chart may also be an option.	B

Example - what does the bar chart show?



Is it a Bar chart or a Histogram? **Bar Chart**

Is this an appropriate way to display the data? **Yes - this is discrete data**

Is it done "well"? **Looks good on first inspection, it is clearly labelled and the units of measure are displayed. My criticism would be that it doesn't tell us when the data was recorded.**

Is it better to live in a region with a tall bar or a short bar? **Short bars are best here, a smaller percentage of people waiting over 12 months is good.**

What does it tell you? **NW Thames was just the highest, but SW Thames does considerably lower, better than the rest.**

Would this graph support the argument that people in London wait longer than those outside London? **Not without similar information from regions outside London.**

Would this graph support the argument that a smaller proportion of patients in the SW Thames region of London wait over 12 months, when compared to other London regions?

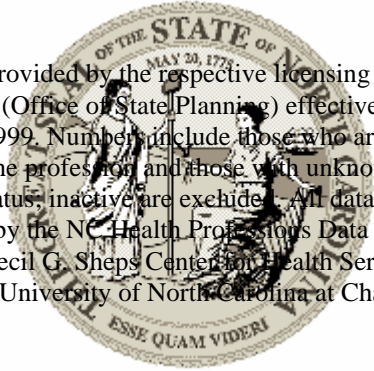
Yes it does exactly that.

Task 5

Totals, averages and Percentages

The file called **North Carolina Nurses** has in it data about the number of nurses in North Carolina, USA.

The data is already aggregated. Sometimes pre-aggregated data is difficult to do any further analyses on, for example it is not usually sensible to average averages.



Data are provided by the respective licensing boards and LINC (Office of State Planning) effective October 1999. Numbers include those who are active in the profession and those with unknown activity status; inactive are excluded. All data are compiled by the NC Health Professions Data System, Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill.

What is the total number of nurses registered in North Carolina?

70620

How many counties are there in North Carolina?

100

Have a go at working the Mean out manually to check SPSS isn't kidding us.

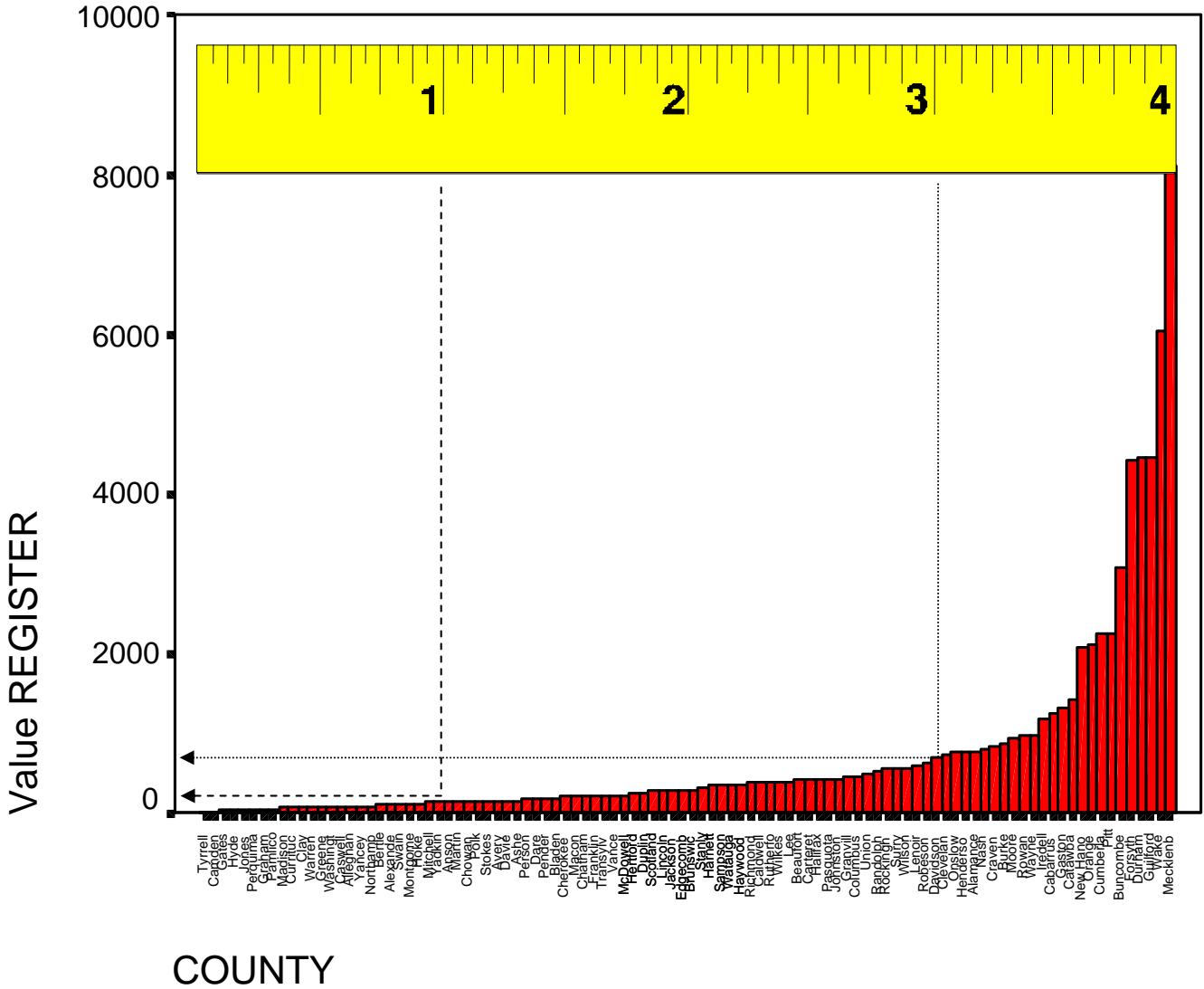
706.2

The mean was worked out by dividing the number of nurses by the number of counties.

$$70620 \div 100 = 706.2$$

Use the graph, then SPSS to fill in the table below.

	Estimate from graph.	Figure from SPSS
Lower Quartile (25 th percentile)		123
Median (50 th percentile)		287
Upper Quartile (75 th percentile)		680.25
Inter Quartile Range (IQR)		557.25 (680.25 - 123)



The spending money allocations.

Name	Spending Money per month	Percentage of total Spending Money	working
Tom	8.00	40%	$100 * 8 \div 20$
Rachel	7.00	35%	$100 * 7 \div 20$
Jodi	5.00	25%	$100 * 5 \div 20$

Cleveland is probably the easiest one to check, 735 as a percentage of 70620 is

$100 * 735 \div 70620 = 1.04078164825828377230246389124894$ which is near enough for me.

Summary: Percentages show proportions, it should be clear what they are percentages of.

Example: In a year an Optician saw 150 (30%) patients with long sight 250 (50%) patients with short sight and the rest had normal sight. How many patients did she see and how many had normal sight. What type of vision was most prevalent?

Working backwards, we know that the percentages should add up to 100% therefore I know that 30% have long sight 50% have short sight, that gives 80% with imperfect sight. $100-80=20$ so I can deduce that 20% have normal sight.

*Finding the total number of patients is harder, we know that 50% represents 250 patients so in this case we could double the 50% to get 100%, and so doubling 250 would give us a **total of 500 patients.***

More generally you would work percentages backwards with the following formula:

$$\text{Total} = 100 \times \text{number} / \text{percentage}$$

e.g. if we knew only that 150 (30%) of patients had long sight $500=100 \times 150 \div 30$

What the formula does is find out what one- percent would be then multiply it by 100 to find 100 percent.

*So we can say that our 20% with normal sight represent $20 \times 150 \div 30$ patients, i.e. **100 patients had normal sight.***

*What type of vision was most prevalent? **short sight - 250 (50%) patients.***

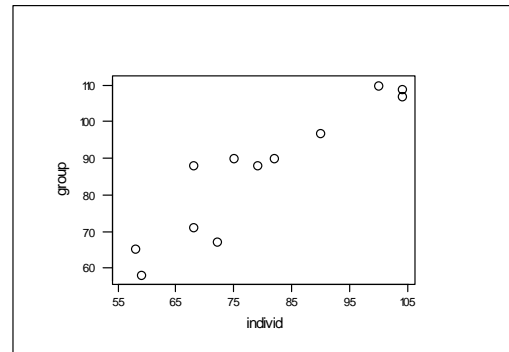
Task 6

This task is intended as an introduction to the process of looking for changes in datasets.

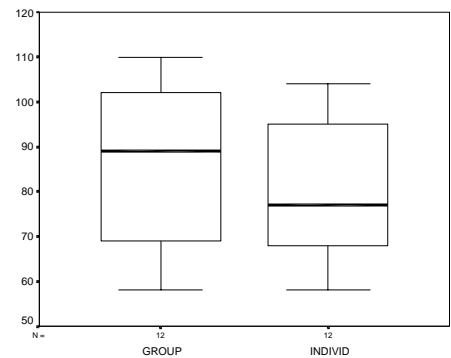
Using Scattergrams to look for Changes

This should look like this:

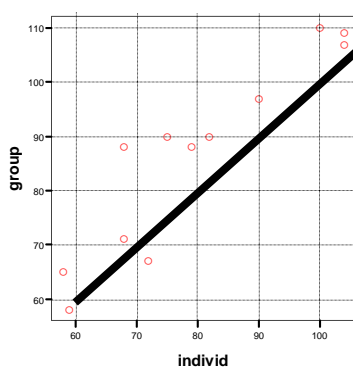
The data appear to show that subjects do more steps under group conditions, although a couple do less.



From the Boxplot we can only see that there seems to be some increase: we cannot tell how many or which subjects do more steps under group conditions. For instance we cannot tell whether the patient who does most steps under individual conditions is the same as the one who does most under group conditions.



Your graph should now look like this:



The line I drew is extended a little beyond the (100, 100) co-ordinate. You should note that most of the points are above the line. This indicates that most people did more steps when working in a group.

The “Point id Tool” lets you see what line of data created a point. The data on line 3 corresponds to the point below the line, 67 steps in the group and 72 individually, this individual has bucked the trend!

Is the data we are looking at here Discrete or Continuous? Discrete.

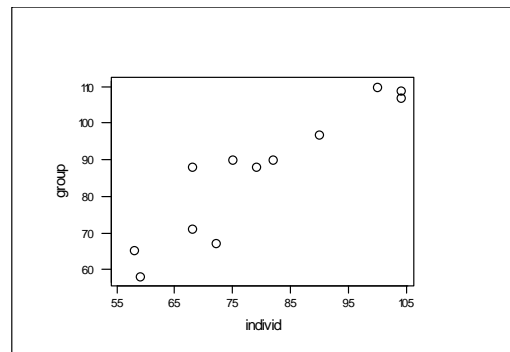
Task 7

This task is designed to introduce you to correlation. It is important at this stage that you distinguish between comparing levels and correlation. For instance, in a study of patients' weights before and after using a new diet there might be positive correlation showing that patients who were heavier before tended to be heavier later, and those who were lighter before tended to be lighter later. This could occur whether the patients as a whole gained weight or lost it. Questions about correlation are different questions from questions about increases or decreases. It is important to always be clear what kind of question you are asking.

Correlation

Looking for Correlation is different from looking for increases or decreases

The diagram shows that subjects who did more steps in individual conditions tended also to do more steps in group conditions and those who did less steps in individual conditions tended also to do less steps in group conditions. This is shown by the points forming a shape which slopes upward from left to right, with very few points near the top left or bottom right. This shows fairly strong positive correlation. For examples of different types of correlation look in the Glossary document.



My guesses on the following correlations are below – but this isn't necessarily the correct answer (*I know it's a cop-out, but that's statistics for you!*):

	Strong/Weak	Positive/Negative
An individuals' height and body weight?	Quite strong?	Positive
The number of new cars sold per year and the price of houses?	Probably some correlation	Positive
A babies age in months and weight in kg?	Strong	Positive
The weight of a dog and its owner?	I have no idea!	

One important point here is not to confuse correlation with cause and effect. It is unlikely that the number of new cars sold per year effects the price of houses, it is more likely that they are both linked to a third factor.

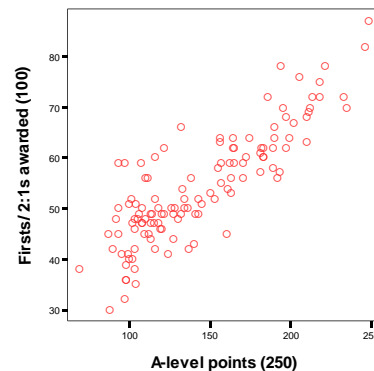
Summary: Scatter plots are used to show paired data, where for example one person is tested under two circumstances, each individual will have a pair of readings. In this example a scatter plot can be used to indicate changes between the performance in different circumstances. Scatter plots are also typically used to show correlation. Scatter plots should be clearly labelled and the units of measure displayed.

Example: The graph shows the proportion of Firsts or Upper Seconds as a measure of degree attainment, plotted against standards of A level passes obtained by new students, the data is from the Sunday times survey of HE. It is sighted as evidence that universities with an intake of students with "better" A-levels have an output of students with a higher percentage of "better" classed degrees.

Is this an appropriate way to show this data? Yes, this is paired data, one dot represents the data from one University.

Is the graph labelled adequately? Not bad, but I would have liked an overall title and some indication about how the a-level scores are derived (is big = good on this scale?).

What does it show? It shows that establishments with higher average A-level attainment students at intake tend to award a higher level of degree.



Does this support the above argument? It appears to support the theory that universities with an intake of students with "better" A-levels have an output of students with a higher percentage of "better" classed degrees. However it is only an overall picture, it doesn't preclude the possibility that the worst A-level student could end up with the highest degree classification! It is looking at universities not students.

Task 8

Line graphs.

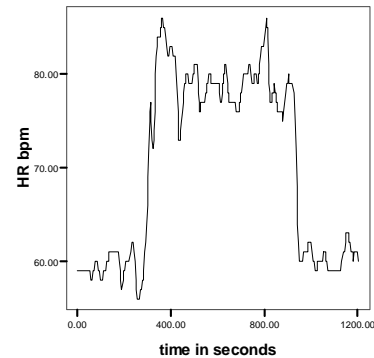
Load the file called “**Oxygen used walking**”

Creating the line graph.


Looking at the graph. It is easy to see when the subject started and stopped walking!

It appears at a glance that the heart rate increases massively during the exercise, by a factor of about six!

Now look carefully at the Y-axis. What is the heart rate when the Y-axis meets the X-axis?

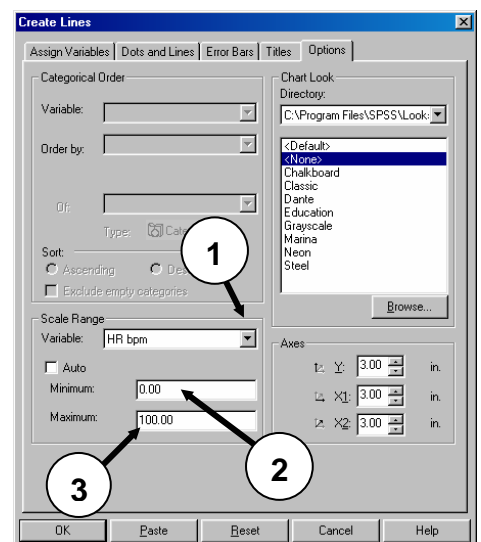


The lowest heart rate is 56 beats per minute, the highest is 86 (you can verify this by calculating the descriptive statistics). SPSS has chosen the range of the Y-axis to show all the data but it not obvious until you inspect the axes whether the scale is starting at zero or not. In this case it is no problem, however a false origin is often used to exaggerate differences, the advertising world is a frequent culprit, but it is not exclusively done in commercial interest! It may sometimes be useful to focus on the differences, but the viewer must be made aware of this. Drawing the graph with no false origin puts the change in heart rate in perspective.

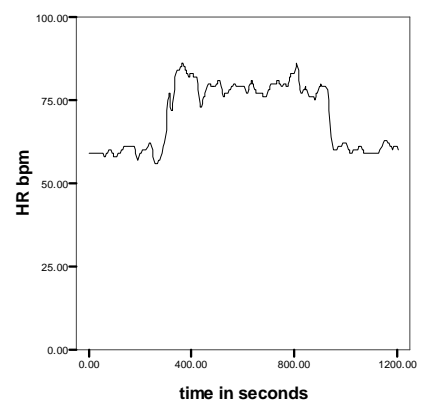
 You can select a previously used dialog box by pressing the Dialog Recall button. When you use the Dialog Recall button, the last used options are shown as a list you can pick from. The names may not be quite the same as you see on the menus – but you’ll get by!

Click on the Options tab at the top of the dialog box, then;

1. Select the Heart Rate variable.
2. Set the minimum to 0.
3. Set the maximum to 100.



This is the same graph having set the minimum to 0 and the maximum to 100 for the HR.



Multiple line graphs.

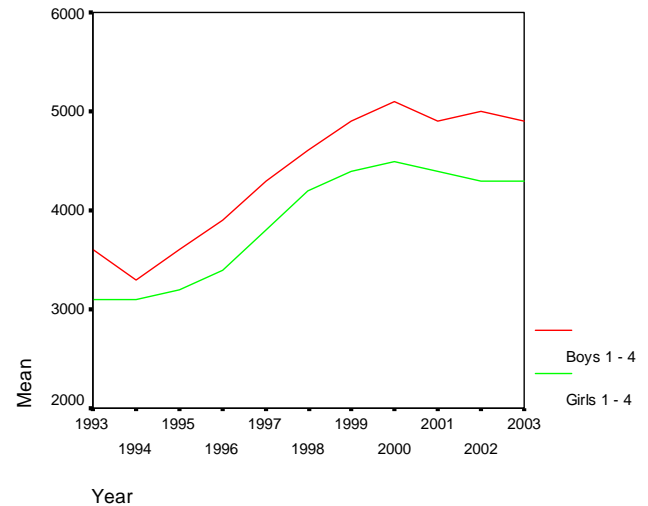
The graph that appears should let you answer the following questions;

1 In the 11 years covered by the data do the numbers of girls and boys looked aged 1 to 4 after by Local Authorities in England appear to increase?

Yes, they've increased from around three to four thousand up to four or five thousand.

2 Are the number of boys and girls in the age group 1 to 4 staying in roughly the same proportion, i.e. do they seem to increase or decrease together?

Yes, the number of girls is constantly a bit less than the number of boys. They have increased together.



Now plot the data for the 16 and over age group, can you see any difference between the girls and boys?

Yes there is a difference.

Try to write down what the difference is (no need to speculate on the cause, just what the data show).

Over the years the difference between the number of boys looked after by the local authority and the number of girls has changed, the division between them has increased, there are now (at the end of this period) considerably more boys in this group than girls, this was not the case in 1993.



Task 9

Reading tabular information.

The percentages should be similar to those in the table we studied before, the slight differences being due to the points discussed.

				1996/97 - 1998/99			
				Percent	Total discharges	Total patients	
England				8.7	2103	27551	
Trust cluster							
Small/medium acute			9.0	577	3463		
Large acute			7.7	3805	9063		
Very large acute			8.1	5733	2643		
Acute teaching			8.3	676	9897		
Multiservice			0.5	5374	0480		

				Page 1			
				disc9699 = Total discharges 1996/99 Sum	pati9699 = Total patients 1996/99 Sum		
typeSize Trust type/size							Average
Acute teaching			9086	18680			49%
Large acute			13805	29063			48%
Multiservice			14530	28437			51%
Small/medium acute			6577	13463			49%
Very large acute			15213	31692			48%
Grand Total			59211	121335			49%

When comparing the tables note that the rows are in different orders.

The Grand Total in the table from SPSS sensibly gives a grand average for the Average column. This Grand average is not simply the average of the averages, this would be meaningless, it is the average worked out from all the data.

Answers to the 9 questions.

1. What percentage of these patients were discharged within 28 days from English trusts in the year 1996/97? *The table says 49.1%*
2. How many patients does this percentage represent? *The table says 20556*
3. The trusts are grouped, Small/medium acute, Large acute, Very large acute etc. of the five groups represented, which, for the three years taken together, achieved the highest rate of discharge within 28 days? *Multiservice.*
4. Which group, over the three years, had the lowest rate of discharge? *Large acute.*
5. In the year 1997/98 which group had the lowest rate of discharge? *Acute teaching.*

6. During the entire three-year period, how many patients in this age group were admitted with a broken neck of femur in England? *127551*
7. Over the three years, for all English trusts, has the number of patients in this age group admitted with a broken neck of femur increased or decreased? *From this data it looks like a slight increase, but this could have been due to specific conditions, e.g. a very icy winter, it may not be a general trend.*
8. Over the three years, for all English trusts, is the percentage of discharges within 28 days for these cases increasing or decreasing? *No – but again it could be due to a specific effect.*

A question the table is not really equipped to answer...

9. Is it good to be discharged earlier or later? *There is an assumption in this set of figures that the early return to home for these patients is generally a good thing. This is partly a politically driven goal, a fence the government have erected to jump in order to indicate good performance to the public. However it must be the case that for some individuals an early trip home is the last thing in their interest, the percentages will never reach 100!*

How well did you do?

- Less than 4 correct – where were you last night! – have another look, try to see if you can see why the answers are what they are.
- Between 5 and 8 – well done, have a look at the ones that got by you, it's easy to look quickly at data like this and pick the wrong number.
- Over 8 – Whitehall beware! Not much gets past you.

Summary: Tables are generally used for structured information, if appropriate the independent variable should be on the left side of the table, dependant variables should be on the right of the table. Tables may be formatted (e.g. the use of indenting and lines etc.) to aid navigation within a table. They should be clearly labelled and the units of measure displayed.

Tables are often used to display aggregated data, quoting totals or means for groups. There are dangers in this, although seeing a mean tells us generally the level of the data, as we have seen earlier this may not tell us how the data are distributed within the population. Similarly "percentages of patients discharged before 28 days" gives us no idea of how long those not discharged within 28 days are spending in hospital and no amount of looking at the aggregated data will tell you!

There are other limits to the amount of re-use aggregated data can be put...

Look at the table below, it gives the mean age of males and females living in a house. If we want to find the average age of people living in the house, regardless of gender can we do it from the information given in this table? It may help to know there are 5 people in the house, so there isn't any chance of the same number of males and females.

<i>Mean Male age</i>	<i>Mean Female age</i>
<i>30.00</i>	<i>20.33</i>

No. You can't find the average age of people living in the house from the data supplied in the table.(This is an issue if wanting to combine results from several studies.)

If you are tempted to just find the mean of the two numbers quoted you would be wrong, this would be $(30.00 + 20.33) / 2 = 25.165$ but assumes equal numbers of males and females!

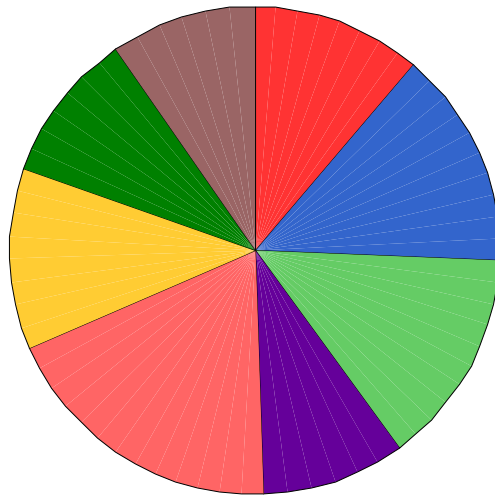
The real ages were two males 45 and 15, three females 40,12 and 9. So the overall mean age is $(45 + 15 + 40 + 12 + 9) / 5 = 24.2$

So in order to find the overall mean the summarised data presented in the table was of little use. Would knowing there are 3 females have helped?

Yes, we could have done a weighted average of the two means - I'll not go into the method, I bet there's a website out there eager to show just how if you need to use it.

Task 10

Pie Charts



Pies show Sums of patien97

SPSS should produce a pie chart that summarises the data in the file called **hip fracture discharges**. The file contains more than one record for each region, it has one record (line) for each hospital. When making a pie chart in this way SPSS can summarise the data.

Of the four types of data; nominal, ordinal, interval and ratio data the Trust Cluster variable is storing nominal data.

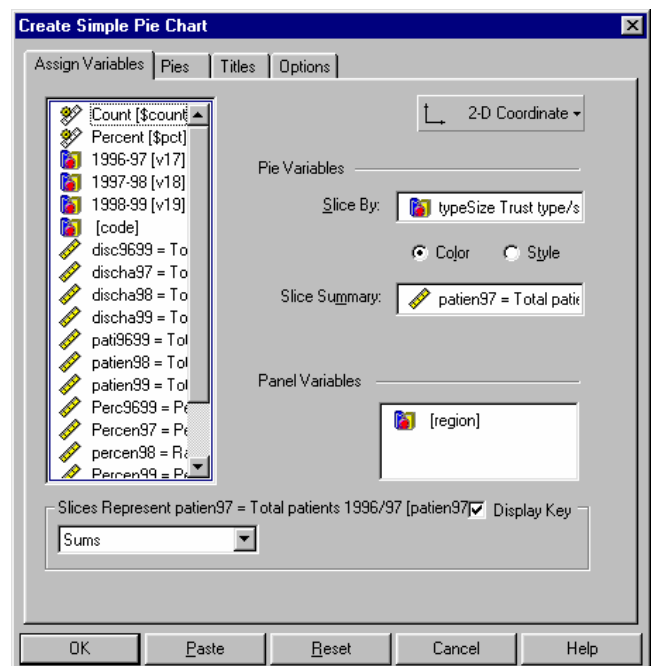
Multiple pie charts.

Creating more than one pie chart at a time can be a useful tool for quickly seeing differences. We can have a go at seeing regional differences in the **hip fracture discharges** data.

Assign the variables as follows;

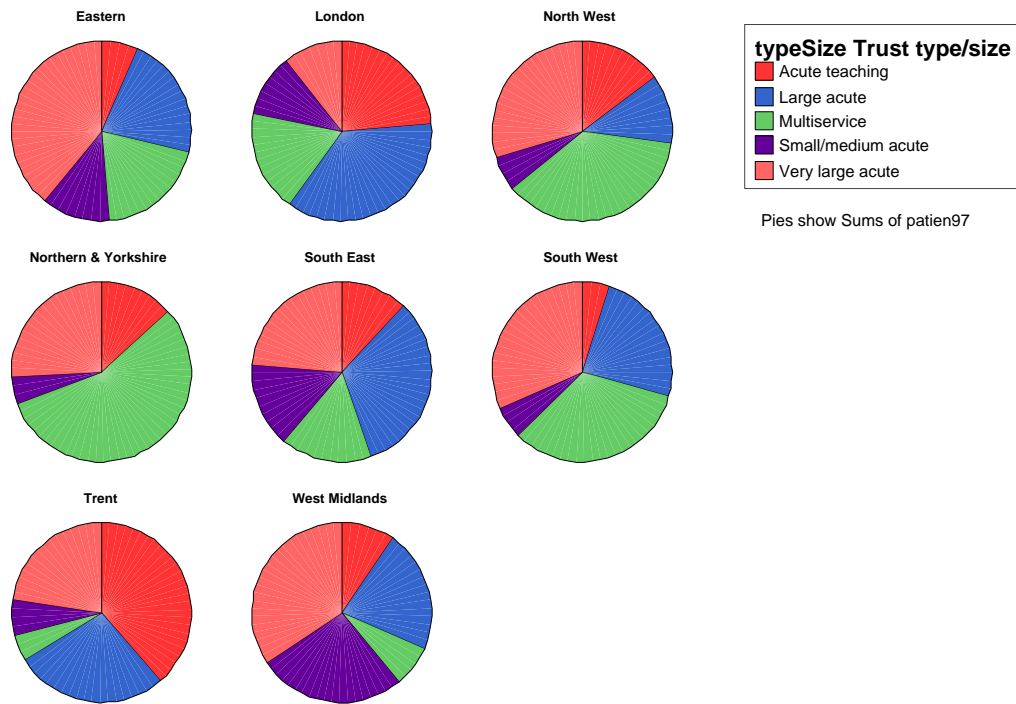
Remember you might need top scroll through the list of variables to find the ones you need.

Slice By:	typesize (typesize = Trust type/size)
Slice summary:	patient97 (patient97 = Total patients 1996/97)
Panel variable:	Region (region = NHS region)



One reason the resulting graphs are different they from the first pie chart you created could be that there are different numbers of each trust in each region. (e.g. There may be less acute teaching hospitals in the South West.)

Would you expect all the regions to be the same? No, the demography and geography of the regions will probably dictate different patterns of provision.



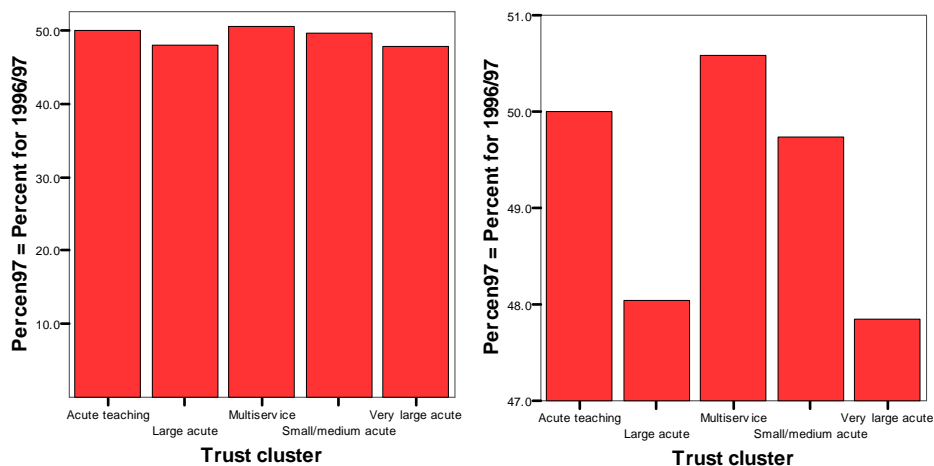
When is a pie chart suitable?

With regard to the data in the file called **hip discharges summary**.

The percentage of patients discharged within 28 days from each type of trust is in the “percent 97” variable.

- 1 Would it be useful to draw a pie chart of this data? I don't think it would be very useful. One useful guide is to see if the percentages would add up to 100. In this case they add up to 246! Pie charts are generally used to show how something is split-up.
- 2 How would you graph the percentage discharge? (Bar chart, line graph etc.)

A bar chart is the best option. The two below are drawn from the data in question. Notice the visual effect of the false origin in the second graph – it certainly exaggerates the differences. (Notice I chose Bar Chart not Histogram – why? Consult the glossary.)



Summary: Pie charts, are used to show proportion, e.g. the number of votes cast for each party in an election. The pie should add up to 100% of the observed data. Pie charts should be clearly labelled and the units of measure displayed.

Which of the following do you think are suited to presenting on a single pie chart?

- *The amount of funding for each department in a hospital. **This is typical pie chart material!***
- *The change in funding over a three year period. **Histogram or line graph would be most appropriate.***
- *The percentage of patients being readmitted within one week of discharge. **This could be a useful pie chart, it could be done using multiple graphs to compare different hospitals etc.***

COMMENTS ON THE TASKS WITH ANSWERS.	2
TASK 1	2
<i>Entering and saving Data.</i>	2
TASK 2	2
<i>Looking at Data:</i>	2
<i>Drawing Boxplots.....</i>	3
<i>Using Descriptive Statistics.....</i>	3
<i>Different types of data</i>	5
<i>The difference between Mean and Median.....</i>	5
TASK 3	6
<i>Standard Deviation (S.D.) using SPSS.....</i>	7
TASK 4	9
<i>Histograms and the Normal Distribution.....</i>	9
TASK 5	13
<i>Totals, averages and Percentages</i>	13
TASK 6	16
<i>Using Scattergrams to look for Changes</i>	16
TASK 7	17
<i>Correlation.....</i>	17
TASK 8	19
<i>Line graphs.....</i>	19
TASK 9	21
<i>Reading tabular information.....</i>	21
TASK 10.....	24
<i>Pie Charts.....</i>	24