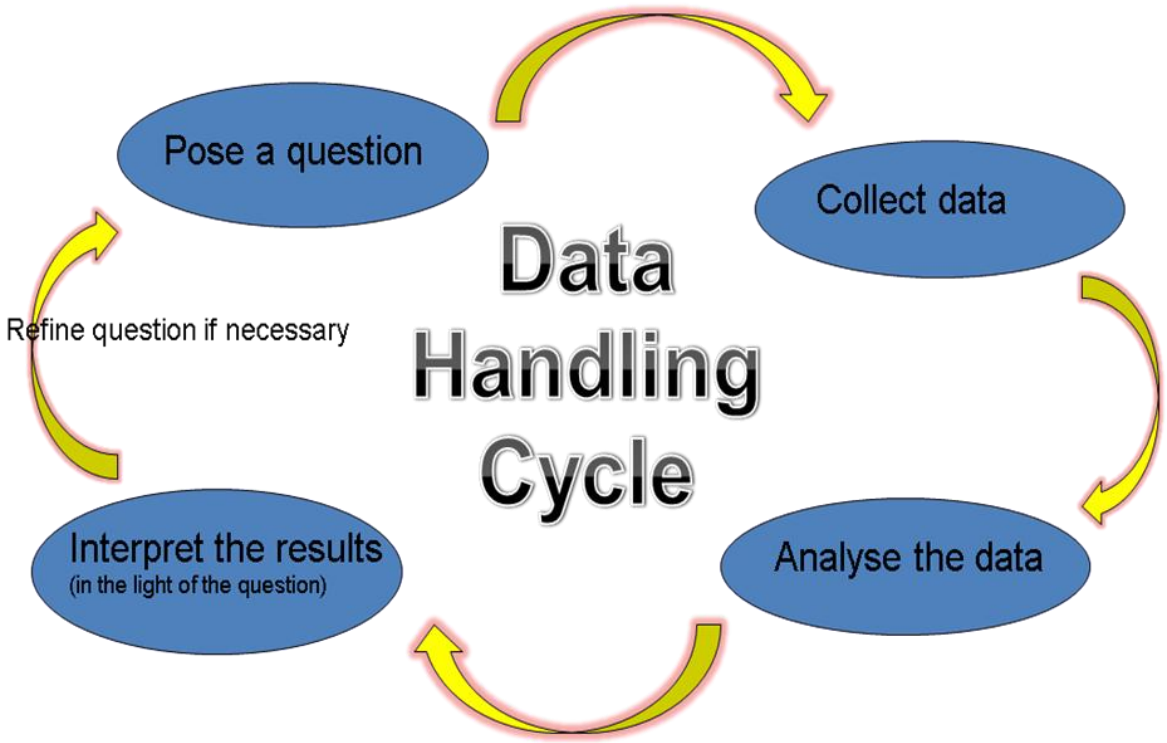
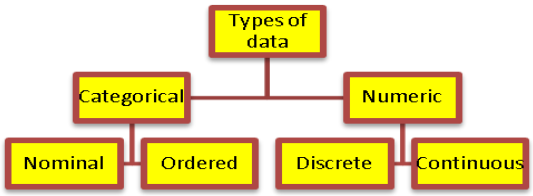


© Project Maths Development Team 2009

Types of Data



For more detail on all aspects of Statistics see http://www.nce-mstl.ie/index.php?option=com_remository&Itemid=79&func=select&id=39
Summer Course on Probability and Statistics NCE –MSTL

Categorical Data: The answer to "what colour is your hair?" produces categorical data, which fits into the categories "black", "brown", "red", "blonde", "other".

- **Nominal** e.g. naming or classifying e.g. blue eyes, brown eyes, blood group types, makes of car Examples from Census at School**:** gender, favourite subject/sport, pets.
- **Ordered** – involves some order e.g. first, second, third, Jan., Feb., March, schoolwork pressure – a lot, some, very little, none.

Categorical	**Nominal** Can be identified by particular names or categories, and cannot be organized according to any natural order.	**Examples**	**Suitable graphical representation**
		Gender : female or male **Hair colour:** black, blonde etc **Favourite sport:** soccer, rugby etc	Bar Chart, line plots, pie chart
	Ordered Identified by categories which can be ordered in some way	**Watching TV:** never, rarely, sometimes , a lot	Bar Chart, line plots, pie chart

Numeric Data: Data represented by real numbers

- **Discrete** – distinct values, e.g. how many people live in each household i.e. cannot have 2.75 people in a household
- **Continuous** – infinite number of values between any 2 given values e.g. heights, weights, lengths in the long jump

Numeric	**Discrete** Data can only have a finite number of values	**Examples**	**Suitable graphical representation**
		Number of peas in a pod, Age in years (as opposed to age)	Bar Chart, pie chart, stem plot
	Continuous Data can assume an infinite number of values between any 2 given values. Students height may be 1.4325m	Height, arm span, foot length. *(Note: In census at school students are required to record their answers to nearest cm or mm so their responses actually give discrete data*)	Histogram, stem plot

In practice no scale is truly continuous because measurement is restricted by some level of accuracy.

Look at all the data collected through Census at School and categorise as above.

Note: Continuous data are often **grouped into class intervals** to make them easier to handle. Instead of displaying every height measured in a class of 30 students, it is more effective to display group categories such as 130 to 139cm, 140 to 149 cm etc.

Discrete data may also be grouped or ungrouped but if there are a small number of responses it is often better to leave them ungrouped.

A grouped frequency table

(Note: a frequency table can be used for discrete variable or categorical data)

The grouped frequency table records the number of occurrences of values within specified intervals. The range of each interval is called the *class interval.*

Notes on class intervals:

Height(m)	Tally	Frequency
1.35-1.40$^-$		
1.40-1.45$^-$		

1.40-1.45$^-$ contains any value from 1.40 up to and not including 1.45.

Single variable vs. Paired Data

Single variable data: Only one item of one of the above types of data is collected for each student e.g. height

Paired Data: Data (of any of the above types) collected in pairs from each student **to see if there is a relationship** between the variables e.g. height and hand span, distance from school and number of days students have been late, mobile phone bill and age, index finger length and height.

Data: Category, discrete variable or continuous variable + single variable (if collected singly) or paired (if collected in pairs).

Category paired data: colour of eyes and gender.

Discrete paired: Number of bars eaten per week and number of tooth fillings

Continuous paired: Height and weight

Category and discrete paired: Type of dwelling and number of occupants etc.

© Project Maths Development Team 2009

Correlation, Paired Data and Scatter Graphs

Correlation is about assessing the **strength of the relationship** between two sets of data.

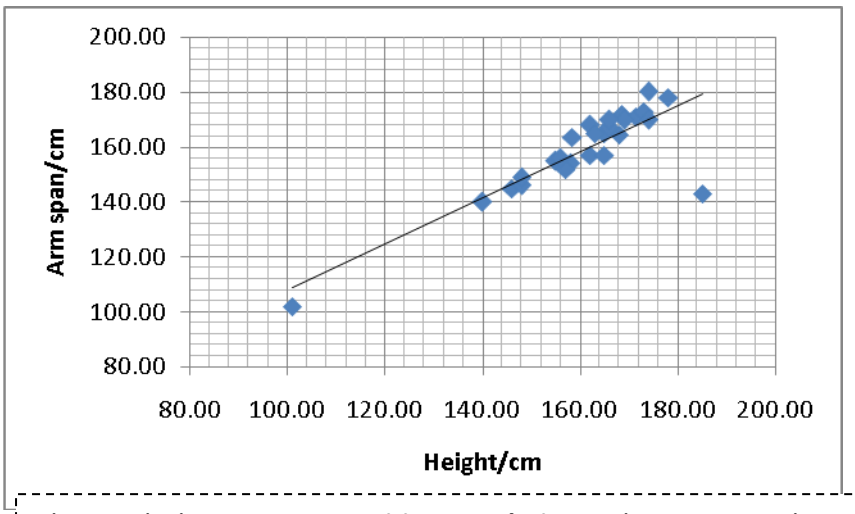
Scatter graphs can show the relationship between 2 variables using ordered pairs plotted on a coordinate grid. The data points are **not joined.** The resulting pattern shows the **type and strength** of the relationship between the two variables. Where a relationship exists, a line of best fit can be drawn between the points.

The line of best fit goes roughly through the middle of the scatter of the points. It should have as many points on one side of the line as the other, and it doesn't have to go through any of the points. It can go through some, all or none of the points. Strong correlation is when the scatter points lie very close to the line. It also depends on the size of the sample from which the data was chosen.

Scatter graphs can show positive or negative correlation, weak or strong correlation, outliers and spread of data.

Example 1

- Pose the question: Is there a correlation between height and arm span?
- Predict the answer.
- Collect the data. We can use data from census at school.
- Analyse using a scatter graph.
- Interpret the results.



This graph show a **strong positive correlation** with most points lying very close to the line of best fit, and one outlier, which may be a valid rare occurrence or due to an error in measurement.

As height increases, arm span tends to increase

(As x increases, y increases)

(e.g. The more I exercise the fitter I become – positive correlation)

© Project Maths Development Team 2009

The main use of the line of best fit is for **predictions.** Providing there is a strong correlation i.e. most points lying close to the line of best fit, then we could use the above graph to predict the arm span of a person of height of 130 cm which as we can see from the above is close to 130cm(Vitruvian man).

Interpolation is when predictions are made within the range of known values, which is reasonable, given a strong correlation. However, **extrapolation** which involves predictions outside the known range of values should be treated with caution as we cannot be certain the trend will continue beyond the known range.

When plotting a scatter graph, which variable should go on which axis?

We usually follow the convention that the independent variable goes on the x-axis and the dependant variable on the y-axis. In the above case it really doesn't matter although it is usually more reasonable to explain a person's arm span by their height rather than the other way around.

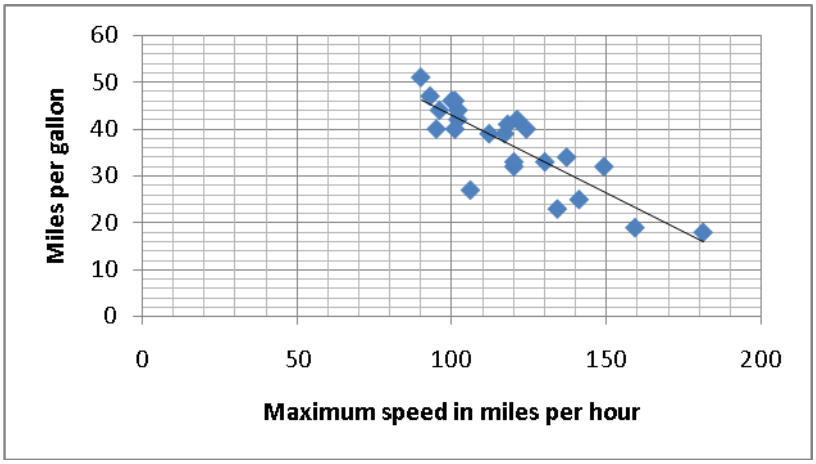
Example 2

Question: Is there a relation between maximum speed achieved by different makes of cars and fuel economy (mpg)?

Predict the answer

Collect the data – car magazine

Analyse the data using a scatter graph. Interpret the results.



There is a reasonably strong negative correlation between the max speed and mpg.

(Negative correlation: The longer I spend playing play station the less time I spend studying. The *more* of one the *less* of the other)

Example 3

Question: Is there a correlation between reaction time and height?

Predict the answer.

Collect the data – census at school.

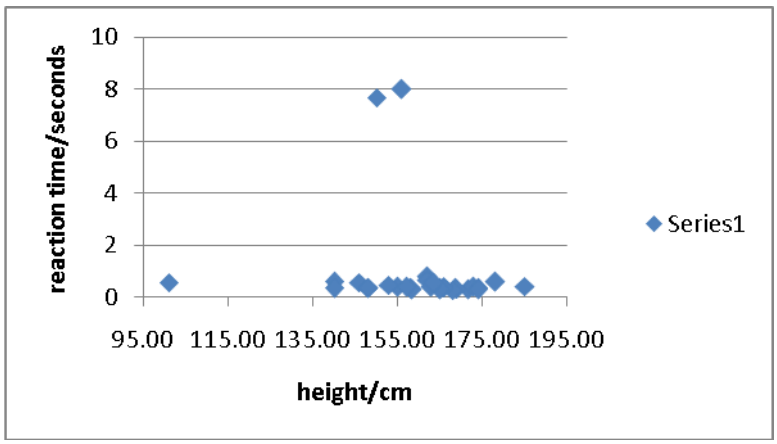
Analyse the data using a scatter graph.

The graph shows **no correlation** between reaction time and height. As students height increases their reaction time stays roughly the same. There are a couple of outliers but there are not enough data points to suggest that height is causing such slow reaction times.

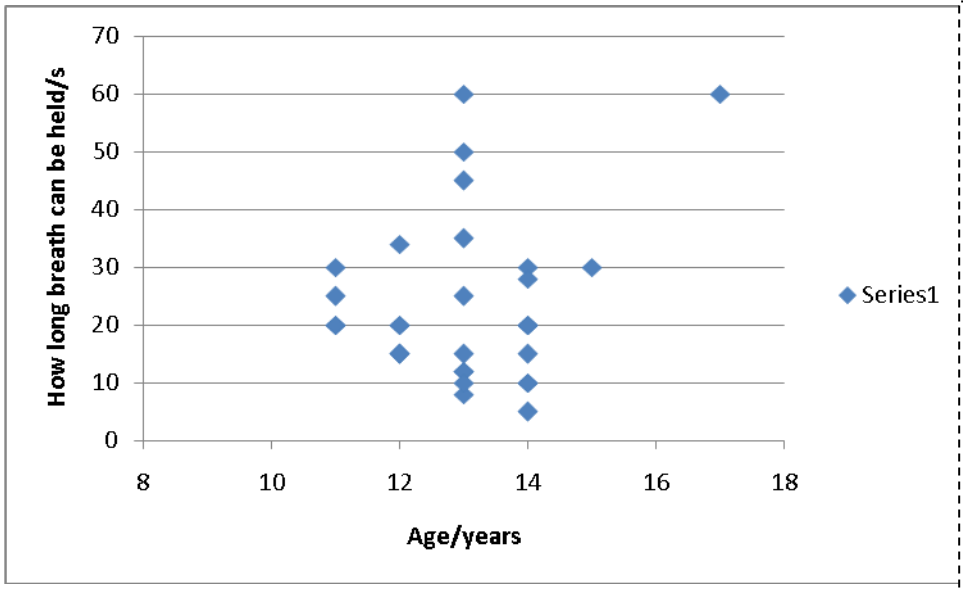
No point in trying to draw a line of best fit.

© Project Maths Development Team 2009

Interpret the results.



Q: Is there a correlation between age and how long you can hold your breath for?



The graph shows **no correlation** between age and how long breath can be held for. We need to be aware that this is a relatively small spread of ages from 11 to 18.

Examples and questions on correlation

http://www.cambridge.org/uk/education/secondary/mathematics/ks4/smp_interact/downloads/22%20Paired%20data.pdf ------Look at this!

Correlation vs Causation

Even strong correlation may be no more than a statistical association and does not always imply causality.

Correlation is a mathematical relationship between two variables which are measured.

A correlation of 1 means two things are completely related and that one can predict the other.

A correlation of 0 means two things are unrelated and knowing one does not allow prediction of the other.

© Project Maths Development Team 2009

However, just because there is a strong correlation between two things does not mean that one causes the other. A consistently strong correlation may **suggest** causation but does not prove it. Look at the examples below which show a strong correlation but do not prove causality.

- With the decrease in the number of pirates we have seen an increase in global warming over the same time period. Does this mean global warming is caused by the decrease in pirates?
- When sales of ice cream increase, rates of deaths by drowning increase. Ice cream causes drowning?

© Project Maths Development Team 2009

A statistical Investigation

1. Pose the question –

- What *exactly* is your question?
- What are the factors?
- What sort of results are you expecting?
- How are you going to find out?

Posing the question – students should have a clearly articulated problem in mind. They may be attracted to an area like sport or movies without a clear question in mind. **Teacher's role** should be to encourage pupils to move away from just thinking in terms of an area of interest to **forming a question which requires an answer.** The actual wording of the question is very important, and it may need to be refined more than once. Look ahead to the data collecting stage and anticipate problems or sources of error. Is it possible to answer the question?

e.g. investigate use of vowels in language – too open ended. Refine – which is the most or least commonly used vowel? Do vowels mostly occur in the middle or at the end or beginning of words? Collecting data – choose a page from the newspaper or your English book and count how often or where vowels occurred. Analyse – tabulate the data and graph it. Interpret the result.

2. Collect the data

Primary sources: You collect the data yourself. Primary data can be got from an experiment, study or survey. What will you need to measure? How? In what units?

Sampling How big a sample will you choose? How will you select it?

Samples must be *randomly chosen* and *reasonably large.*

A sample is considered to be random if every member of the population from which it is chosen has an equal chance of being selected and the selections are made independently.

Asking 10 boys who are all friends of the same age, what is their favourite sport and football team , is not an indication of peoples' sporting preferences in general.

Asking people do they believe in God outside of church on Sunday is not likely to be a representative sample.

There is little point in conducting a survey on attitudes to gambling in a casino.

(For more detail on the different types of sampling (simple random sampling, stratified, cluster, quota etc.) see "Summer course in Statistics and Probability" Ailish Hannigan http://www.nce-mstl.ie/index.php?option=com_remository&Itemid=79&func=select&id=39

© Project Maths Development Team 2009

Designing a questionnaire

The questions need to be clear and logically ordered and the survey should provide interesting and usable information. It should be possible to record the responses quickly and accurately. It should first be tested out on a small sample.

Recording the data: It may be a good idea to have a tally chart or stem chart ready beforehand to record data. Hence students will need to have some idea of the categories of data or the range of data so doing a quick pilot survey first can help here. Accuracy required in measurements needs to be thought out first also.

Secondary sources of Data: Census at School, the internet, newspapers and magazines

3. Analyse the data

What sort of graph is most suitable to display your results?

Can you state one interesting or unexpected result from your data?

What trends do you see?

Matching data to graph type

How data should be represented depends on the type of data, and the purpose of the investigation.

Scatter plot – ideal for paired data

Stem plot – can show discrete and continuous data but not categorical

		Bar chart	Frequency Table	Histogram	Line of fit	Pie chart	Scatter plot	Stem plot
Single variable	Categories	✓	✓			✓		
	Discrete Variable	✓	✓			✓		✓
	Continuous Variable		✓	✓				✓
Paired Variable	Discrete		✓		✓		✓	
	Continuous		✓		✓		✓	

Students could try out different representations, using bar charts, stem plots, histograms, scatter plots etc. and see which is best and why. Using software here could help, before they actually plot the data themselves.

© Project Maths Development Team 2009

Support in asking the right questions about the representation used.

Representation	Possible questions for analysis
Table	How do the rows compare? How do the columns compare? What are the row and column totals? Investigate writing row and column totals as percentages.
Stem plot	How spread out are the values? Are there any outliers? Are they clustered around the middle? What is the median? If plotting back to back stem plots, are there differences in spread or magnitude?
Scatterplot	Can you see a trend in the points? Is it a straight line? Do the points lie close to the trend line? Does the trend line have a negative or positive slope?(sloping down/up to the right)

4. Interpret the data-

Do your results answer your original question?

Could there be another possible interpretation?

Look back at the appropriateness of each step-

Was the question refined enough to answer the problem?

Was the data collected appropriate to the problem?

Was the graphical method used for analysis suitable – did it reveal trends?

Were there other factors influencing the data? Show scepticism

Pupils should **write up** their conclusions clearly and logically and present them to the class.

When doing so they should list the 4 stages, making it easier to evaluate the investigation.

Spending time and being careful in the write up stage, reflecting on and discussing what they have done will be a rich source of learning.

If they find out that there has been a weakness at any stage of the investigation this will have been a major source of learning as opposed to failure because they will have understood what is involved in a statistical investigation.

© Project Maths Development Team 2009

Other statistical investigations

What is the relationship between the diameter and circumference of circular objects? What data will be collected? What type of data is this and how can it best be represented? How will the relationship be calculated?

The county council needs to monitor the traffic at a particular junction to see if an alternative route is needed. What data might they need to collect? Will they need single variable or paired data? How will they use the data to come to a conclusion? They will need a standard to compare the results to e.g. a certain number of cars in a particular time interval is unacceptable.

Teacher brings in two cuttings from 2 different newspapers- tells students that one is from the Times and one from the Sun. What could you measure about the two cuttings that would distinguish them without knowing which was which? (e.g. average sentence length i.e. number of words in a sentence, average word length)

How many sentences to sample, one hundred, fifty? Which section of the paper to choose – business, sport, foreign affairs, and fashion?

Would your skill at estimating length of lines and size of angle improve with practice?
Given 5 lines of different lengths, A, B, C, D, E, F – all students estimate length of line A. Students draw a stem plot of results. Give the exact measure of the line. Look at outliers, spread of estimates. Now ask them to estimate length of line B and repeat above. Do the same for lines C, D, E, F. If their hypothesis is true estimates for Line F should be more clustered around the true value and outliers probably will have disappeared.
A similar investigation could be done for angles.

(When measuring lengths such as index finger length issues will arise as to the **level of accuracy** – to the nearest cm / mm? Is it possible to measure to the nearest mm?)

© Project Maths Development Team 2009

Issues Arising – a quantitative approach to decision making

Pupils will realise that it can be difficult to test a hypothesis and it may be difficult after an investigation to decide whether to reject or accept a particular hypothesis.

In the previous investigations we did not make quantitative measures of spread, and correlation, so the next step would be measuring **standard deviation** and **correlation coefficient**. Testing whether or not a result is significant will involve probabilities.

If you toss a fair die you would expect to get a 6 one sixth of the time. If you tossed the die 30 times you would expect to get on average 5 sixes, but you might for any set of 30 tosses get 10 sixes or 1 six only. So the question is, outside of what *range of values* would your result have to be before you conclude that the die is probably unfair/loaded. This depends on the range of values that would be acceptable for a fair die.

If for a fair die a 6 occurs mostly between 2 and 8 times then the criterion for loading for the number of sixes in 30 tosses might be outside the range 2 to 8 or if there was a wider variation in the number of 6's got in several sets of 30 tosses for a fair die then the criterion for loading would have to be wider. Usually a statistically significant result at a 5% level of significance is one which would have a probability of less than .05 of occurring i.e. it could only occur *by chance alone* in 1 of every 20 trials or less.

© Project Maths Development Team 2009

Representing Data

What is data? A **datum is a fact** usually expressed as a number e.g. the cost a particular pair of jeans is 90 euro, someone's height is 160 cm.

Table of data: Presenting data in an organised way with rows and columns, a clear title at the top and the source of the data written at the bottom of the table.

e.g Take a quick survey of favourite singer/group from the class.

Put the data on the board and get students to organise it into a table.

A **bar chart** is appropriate to represent this data as the data fall into discrete **categories.**

In the probability section students will have collected data by tossing a die 30 times. Again they will have this organised into a table. This time the values are **discrete variables** so a bar chart is again suitable.

Characteristics of a bar chart

1. The **height** of each bar is a measure of its frequency (the number of observations in that category). Bars must be of equal width so that this is the case.
2. The bars should be separated to show the discrete nature of the data.

Avoid 3 dimensional bar charts because the added depth can make it more difficult to read the data accurately.

Pie Chart

The above data on the favourite groups could be represented in the pie chart.

Characteristics of a pie chart

1. It is circular as opposed to linear
2. The angle at the centre and the area of each sector is proportional to the frequency.
3. The chart cannot be drawn until **all** the information has been collected.

Pie charts are useful when they represent a unified whole – in this case adding up the numbers gives the total numbers of students surveyed in a class.

Using a pie chart for the costs of different brands of cars is not as useful since adding up the prices does not represent an interesting representation.

All the different sources of background radiation could be represented in a pie chart because it represents **all** the different sources.

Bar charts and pie charts are for categorical or discrete variable type data.

© Project Maths Development Team 2009

Histogram (only histograms with equal class intervals to be considered)

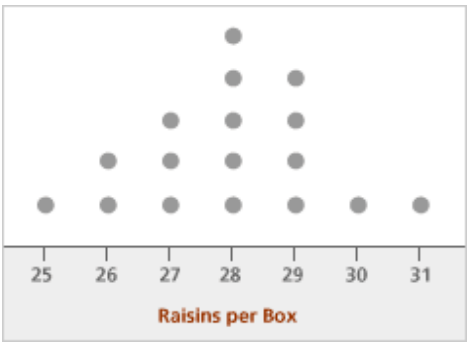
Represents continuous data (e.g. height, weight, time) which has been organised as a frequency table.

- The columns should touch as the data is continuous

Line Plot A line plot shows data on a number line with an **x** or a dot to show frequency.

The following example shows a line plot. X's could be used instead of ○

Example (NCCA materials for Strand 1)



Students were investigating the number of raisins contained in individual mini-boxes of Sun-Maid raisins. They recorded their results in the diagram shown.

(a) (i) How many boxes of raisins did they survey?

(ii) What was the modal number of raisins per box?

(iii) What is the median number of raisins per box? Explain how you found this answer.

(b) If the students chose a box at random from all the boxes they surveyed what is the probability that the box contained 29 raisins?

© Project Maths Development Team 2009

Stem plot or Stem and leaf diagram.

(A type of histogram that retains the raw data.)

It can be used for discrete or continuous data and is a graphical representation of a frequency table.

It is used where we want to understand how data looks **without losing the individual data points.**

Given a set of 20 test marks ranging from 0 to 100:

98,92,92,87,84,82,80,80,80,70,66,65,62,60,50,45,42,42,41,40

Mark	frequency
40	1
41	1
42	2
45	1
50	1
60	1
62	1
65	1
66	1
70	1
80	3
82	1
84	1
87	1
92	2
98	1

We could get a feeling for the distribution of scores by presenting them as above but this would be awkward if the number of scores was large.

Constructing a stem and leaf plot

The stem is the vertical column of numbers on the left hand side. The stem contains the first digit (or more) and that digit's corresponding list(leaves) are on the right. They are called the leaves as they are attached to a particular stem.

The leaves may not initially be sorted into numerical order and a subsequent ordered version may be drawn.

Write in the sample size at the top of the table (e.g N=20) and a title (what the data represents).

© Project Maths Development Team 2009

N=20(sample size)

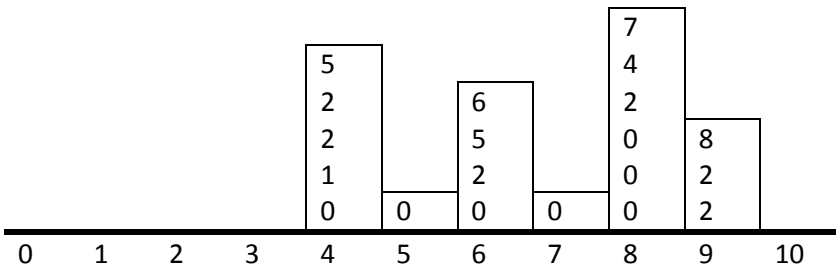
Test scores

0	
1	
2	
3	
4	0 1 2 2 5
5	0
6	0 2 5 6
7	0
8	0 0 0 2 4 7
9	2 2 8
10	

9	2	Represents 92

We can easily see that the mode is 80 and the median is 68 (mean of the 10th and 11th scores)

If we rotate this through 90 degrees we produce a **histogram** of the data.



If we were representing heights of pupils in a class which ranged say from 1.34m to 1.58m, we would use the first 2 digits as the stem and the third digit as the leaf.

Write in a key at the end to explain what the stem and leaf represent.

If the heights had a large number of values between 1.30 m and 1.40 m it would be a good idea to split the values between 1.30 to 1.34 inclusive and another from 1.35 to 1.39 inclusive.

13	3
13	5 5 6
14	0 0 1 1 1 2
14	6 6 7 7 7
15	0 1 2 2 2 3
15	5 5 6 7 7
15	5 is 1.55m

© Project Maths Development Team 2009

Calculate quartiles and interquartile range from an ordered set of data

The median divides the data into two halves. To divide the data into quarters, you then find the medians of these two halves. If you have an even number of values, where the first median was the average of the two middle values, then you include the middle values in your quartile computations. If you have an odd number of values, where the first median was an actual data point, then you do not include that value in your quartile computations. That is, to find the quartiles, you're only looking at the values that haven't yet been used.

e.g. 4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4

The set is ordered

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The first number I need is the median of the entire set. Since there are seventeen values in this list, I need the ninth value:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The median is 4.4.

The next two numbers I need are the medians of the two halves. Since I used the "4.4" in the middle of the list, I can't re-use it, so my two remaining data sets are:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 and 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The first half has eight values, so the first/lower quartile is the average of the middle two:

$(4.3 + 4.3)/2 = 4.3$

The upper/third quartile is:

$(4.7 + 4.8)/2 = 4.75$

The interquartile range is $4.75 - 4.3 = 0.45$

© Project Maths Development Team 2009

Calculate quartiles and interquartile range from a stem plot (using the stem plot from page 16)

13	3					
13	5	5	6			
14	0	0	1	1	1	2
14	6	6	7	7	7	
15	0	1	2	2	2	3
15	5	5	6	7	7	
15	5 is 1.55m					

There are 26 data items which are ordered.

1.33,1.35, 1.35, 1.36, 1.40, 1.40, 1.41, 1.41, 1.41, 1.42, 1.46, 1.46, 1.47, 1.47, 1.47, 1.50, 1.51, 1.52, 1.52, 1.52, 1.53, 1.55, 1.55, 1.56, 1.57, 1.57

Because of the even number of data items the **median** is the average of the 13^{th} and 14^{th} data items i.e. (1.47+1.47)/2 = 1.47 m

The next two numbers I need are the medians of the two halves.
Lower half: 1.33,1.35, 1.35, 1.36, 1.40, 1.40, 1.41, 1.41, 1.41, 1.42, 1.46, 1.46, 1.47

Because this half has an odd number of data items the median of the lower half is the middle item i.e. 7^{th} data item . This is the lower quartile = 1.41 m

Upper half: 1.47, 1.47, 1.50, 1.51, 1.52, 1.52, 1.52, 1.53, 1.55, 1.55, 1.56, 1.57, 1.57

Because this half also has an odd number of data items the median of the upper half is the middle item i.e. 20^{th} data item . This is the upper quartile = 1.52 m

The interquartile range = 1.52 – 1.41 =0.11 m. This is the range of the middle 50% of the data.

© Project Maths Development Team 2009

Back to back stem plots

Comparing 2 sets of data.

Heights for girls						Heights for boys					
				0	13	3					
				5	13	5					
			0	0	14	0	0	1			
7	6	6	5	5	14	6	6	7			
	4	3	3	3	15	0	1	2	2	2	3
				5	15	5	5	6	7	7	
					15	5 is 1.55m					

Percentiles

A percentile is a comparison score. It represents how well or how poorly a student performs compared to other students. It does not represent the percentage of questions the student answered correctly in a test. A percentage gives a number related to your own performance in a test whereas a percentile relates your score to the performance of 100 similar students.

If you do a test with 100 problems, scoring 1% each, and if you get 85 correct, then your score is 85%. This does not relate your score to anyone else who took the test. If your score was given in percentiles it would be based on the number of students who scored below you in the test. If your score was the highest score in the class, and there were 100 students doing the test, then you would score in the 99^{th} percentile i.e. 99% of students scored below you or you had a score better than 99% of the class. Percentiles show how most people are performing as well as how individuals are performing. Percentiles allow us to evaluate a test – if everyone is scoring badly on a particular question then perhaps we have not prepared them well enough for the test or the question may be phrased badly.

Percentiles are very useful for checking testing in a population e.g. if one school scores badly, percentile wise compared to the neighbourhood, then perhaps deeper issues need to be addressed e.g language difficulties etc.

If you score in the 80^{th} percentile then that means that you scored better than 80 out of 100 people who took the test.

© Project Maths Development Team 2009