Statistics

Statistics

Topics:

- Sources and collection of data
- Sampling data
- Describing data
  - Graphs, diagrams & charts
- Computing
  - mean, standard deviation, median, mode & inter-quartile ranges

Learning outcome:

- The calculation of statistical measures
- The interpretation and presentation of data using statistical graphs and charts

Statistics

- What is statistics?

  ‘Collection of methods for planning experiments, obtaining data and then organising data, summarising, presenting, analysing, interpreting, and drawing conclusions’

Sourcing & Collecting Data

- Collecting the statistical data may seem to be the easiest task of the statistician.
  - At it simplest, it may require a trip to the local library, while at its worst, it may involve a ten year expedition to Antarctica !
- Statistics are rarely produced for one's own sake – they are generally needed by someone, somewhere, for a particular reason.

- Suppose we wanted to make a statement about the ages of students at IT Blanchardstown.
  - One possibility would be, "The average is 20.35 years". To obtain that figure, or to estimate it reasonably accurately, we need to know the ages of all of the students.
  - If we have access to a computerised database with all the students' dates of birth, we can probably extract the figure we want in minutes.
  - If we have to send a questionnaire to the students, the job could take weeks – and almost certainly all of them would not reply.
  - We may be satisfied by an alternative form of statement such as, "Most of the students are between 18 and 22 years of age". We can obtain this information by a show of hands in four or five lecture theatres.

Collecting Data

- Collecting data can be time consuming and often expensive

- **Primary data:** Data collected from our original sources for the purpose of a study/experiment
  - Advantage: able to control better what is collected and can monitor, influence or control extraneous conditions which may affect the data

Collecting Data

- **Secondary Data**: Data compiled for another purpose, but which we can utilise for our own study
  - May not know the precise conditions under which the data was collected – treat with caution
  - Advantage: less time consuming to collect
  - In some cases more data than you would hope to collect yourself (e.g., census data)

- Traditional sources of data
  - Papers, books, thesis...
- Computerised data
  - database, CD_ROM

Collecting Data

- **Observational Data**: Data in which subjects are observed and studied, but no attempt is made to manipulate or modify the subjects.
  - e.g. In a study to determine how fit they are, various people are asked to run around a field and then their heart rate, blood pressure etc. are recorded. Other variables such as age, height, weight are also recorded. Suppose we discover that overweight people are less fit than thin people. Does this mean that being overweight makes you less fit? If they lost weight would they be fitter? Or is the problem that they do not do enough exercise, and as a result a) they become overweight and b) they are not fit. The difficulty we face in this study is that we cannot control who is overweight and who is thin. We can only <u>observe</u> the association between overweight and fitness.

Collecting Data

- **Experimental Data**: Data in which a treatment is applied (have control), and then its effects on the subjects are studied.
  - e.g. If the same group of people volunteers to be allocated at random (or maybe by some design), to two groups – one of which would be starved and the other overfed. This would produce two distinct groups, one very thin group, and one overweight group. We would then be able to investigate if being overweight because of overeating caused unfitness

- we can discover *association* between variables from observational data
- and determine *cause and effect* from experimental data

Sampling

- **Population**: All subjects possessing a common characteristic that is being studied.
  - Depending on the data – can be quite large

- **Sample**: A subgroup or subset of the population.
  - Have different methods of selecting a sample from a population

Sampling

- **Random Sampling**: Sampling in which the data is collected using chance methods or random numbers.
  - Every member of the population has an equal chance of being in the sample
  - e.g. To estimate the average total length of 18-hole golf courses in Ireland using a random sample of size 20, you would have to list all of the 18-hole courses (the population) select a random sample of 20 of them (using random numbers for example) and find the information on those selected. The average of the sample of 20 would then be an *estimate* of the population average.

Sampling

- **Stratified Sampling**: Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.
  - More sophisticated form of random sampling
  - e.g. Suppose the committee of the local tennis club wants to find out the opinion of its members on plans for improvements to the club. Its membership is made up of 100 men and 50 women, and they plan to talk to a sample of 30 people. A random sample may produce a majority of women, or no women at all. To avoid the possibility of such bias, the committee takes a stratified random sample by selecting 10 women at random from the 50 female members and 20 men from the 100 male members. In other words, the *proportions* of men and women are the same in the sample as in the population. This sort of stratification may be performed on several variables simultaneously.

- If used correctly, statistics can result in changes which are of great benefit, for example in
  - Medicine: testing of new drugs, casual links such as smoking & cancer
  - Science: forensic evidence
  - Legislation: data on seat-belts, drinking habits
  - Industry: market research, quality control.

- However, statistics may also be used in a misleading way (can be manipulated!!).
  - Suppose a company prints the statement, “Of the washing machines sold by our company in the last 20 years, 90% are still in daily use”. Reading this you might think that there is a good chance your machine will last for 20 years. What the firm ‘forgot’ to mention is that the sales have increased considerably over the years and only 5% of the machines were sold over10 years ago, and none of them are still in use. In fact the average life of a machine is about 5 years !!

Calculating Statistical Measures

Statistical measures - numerical

- Describing data using numerical computation
  - We have already seen how to represent these data graphically – but how would you attempt to describe them without the aid of pictures – perhaps using only a few summery number ?

- calculate some quantities: mean, median, standard deviation which describe the data

Indexed Lists and the Summation Notation

- As we study Statistics, we will often be dealing with additions of large sets of numbers. It is therefore necessary to introduce some notation which will make it easier for us to denote such lists, and to indicate that the numbers in the list are to be added.
- Consider a list of 40 figures, for example, the ages of a group of students. We can’t use a different symbol to denote each one, as we would soon run out of letters. Even if we brought in other alphabets, it would still not be a very convenient way of dealing with a list. Instead, let us use one symbol to denote the age of a student, say x.
- The ages of each student can then be denoted by the letter x, with a subscript assigned to set them apart. Thus the 40 ages are referred to as

$$x1,\ x2,\ x3,\ldots,\ x40.$$

Each of the symbols $x1$, $x2$, $x3$, up to $x40$, is a separate variable representing a particular students age.

The fact that all the numbers in this list are ages of students is reflected in the fact that the same symbol represents them all. Since they are all to be dealt with together, and are only differentiated by their place in the list, this notation is suitable.

We can now talk about the list of 40 figures as

$$x_1, x_2, \ldots, x_{40},$$

or they can be written as the figures

$$x_i, \text{ where i = 1 to 40.}$$

In the second case, the variable *i* is called the index, or an index variable.

Summation

- The first extension of the previous notation is to allow a way of indicating that the numbers in a list are being added. Such a notation will be vital for statistics and topics such as series approximations.
- For our list

$$x1,\ x2, \ldots,\ x40,$$

we will indicate that these 40 numbers are to be added by writing the following expression:

$$\sum_{i=1}^{40} x_i$$

- The symbol S, called sigma, is the Greek equivalent of capital S.
- The fact that '*i* = 1' is written below the S and '40' is written above it means the summation goes from *x*1 to *x*40. Thus the result of the calculation is the total of the values in the list.

Summation

Note that this notation is not saying to multiply a value *x* or a series of values by some number ∑. The symbol ∑ is an instruction to add these numbers, somewhat like an integration sign instructs us to integrate an algebraic function.

$$\sum_i x_i$$

If it is understood that the entire list is to be added, the sum can be written as:

$$\sum x_i$$

The general variable *x* is included, and the index variable is shown below. The index variable could even be dropped, to leave

Since there is only one index variable, it is the one listing the numbers to be added.

Example Let x_i, for i = 1 to 20, be the percentage marks of 20 students in an assessment:

85, 65, 40, 55, 64, 75, 80, 66, 57, 86,
47, 94, 81, 72, 83, 51, 63, 77, 36, 68.

Find the value of

$$\sum_i x_i$$

Solution:

Finding the value of $\sum_i x_i$ simply means adding up the list of 20 numbers. This gives 1345, so

$$\sum_i x_i = 1345.$$

Now consider the quantity

$$\sum_i x_i^{\ 2}$$

This means the sum of the squares of the numbers in the list.

Note that the squared sign is just over the x_i, so the values of x are being squared before they are added. Thus each number is squared, and the results added up.

In the case of the numbers listed in the previous example for the marks in the assessment, the result of this calculation is 95,295. Thus

$$\sum_i x_i^{\ 2} = 95{,}295$$

How does it compare with the number

$$\left(\sum_i x_i\right)^2 ?$$

Here the numbers have been added, and then the result squared. For example, in the case of the assessment marks,

$$\left(\sum_i x_i\right)^2 = 1{,}809{,}025$$

Generally the first number, the sum of squares, is smaller than this number, the square of the sum.

For a list of *n* numbers x_i, the difference between the two quantities

$$\sum_i {x_i}^2 \quad \text{and} \quad \left(\sum_i x_i\right)^2$$

is a vitally important quantity. As we will see it is a measure of how widely distributed the numbers are.

Mean and Standard Deviation

Given a large set of figures, for example a list of sales figures or ages, simply looking at the numbers themselves can tell very little about any trends or patterns in the data.

How can such a collection of figures be described, or summarised, to give some idea of what they mean?

Consider the list of n figures $x_1, x_2, \ldots, x_n$.

The first number we might quote to describe this data list is the **average**. It is given by:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

This is the total of the values, divided by the number of values in the list.

But this number does not say anything about how widely spread the numbers are – a second number is needed to quantify this.

The best such measure is the **standard deviation**, usually denoted *s*, which is given by:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}$$

This number takes the difference between each value of *x* and the average, squares this difference, adds them, and then divides by *n*-1. Squaring the differences ensures they build up rather than cancel.

In fact, the standard deviation can be more quickly calculated as

$$s^2 = \frac{\sum_{i=1}^{n} x_i^{\ 2} - n\left(\overline{x}\right)^2}{n-1}$$

This equation is the same as the previous one for *s*, except for some algebra.

Recall the list of the percentage marks of students in an assessment:

85, 65, 40, 55, 64, 75, 80, 66, 57, 86, 47, 94, 81, 72, 83, 51, 63, 77, 36, 68

We can calculate the average and standard deviation for these figures, and do so quickly using the second, simpler form of the equation for *s*. With this in mind, the terms to be calculated are:

$\sum_i x_i$ to give the mean, and then $\sum_i {x_i}^2$

The value of the standard deviation *s* then comes from the equation

$$s^2 = \frac{\sum_i {x_i}^2 - n\left(\overline{x}\right)^2}{n-1}$$

The values found are: $\sum_i x_i = 1{,}345$ and $\sum_i x_i^{\ 2} = 95{,}295$

The calculation goes as follows:

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{1{,}345}{20} = 67.25$$ **Mean**

For s: $$s^2 = \frac{\sum_i x_i^{\ 2} - n\left(\bar{x}\right)^2}{n-1} = \frac{95{,}295 - 20 \times 67.25^2}{19}$$

The details of the top-line calculation are: $67.25^2 = 4522.5625$ and so then multiplying by 20 gives 90,451.25.

Thus $$s^2 = \frac{95{,}295 - 90{,}451.25}{19} = \frac{4843.75}{19} = 254.934$$

Taking the square root to get s gives $s = 15.97$. **Standard Deviation**

This number is quite high, indicating that the percentage marks were quite widely distributed.

Looking through the list, there were several marks in each ‘decade’ from the 30’s, 40’s and on up to the 80’s. The standard deviation reflects this variety in the numbers.

Now for the opposite case, consider the following list of ages in a first year college class. There are:

- 7 students aged 17,
- 20 aged 18,
- 11 aged 19, and
- 2 are aged 20.

Adding the numbers gives:

7x17 + 20x18 + 11x19 + 2x20
= 139 + 360 + 189 + 40 = 728.

Thus the mean age is: 728/40 = 18.2.

To get the standard deviation, the sum of the squares is needed. It is:

7x172 + 20x182 + 11x192 + 2x202
=7x289 + 20x324 + … + 2x400
= 13,274.

The standard deviation is then given by putting this value into the equation:

$$\frac{13274-(40\times 18.2^2)}{39}=\frac{13274-13249.6}{39}$$

so $s^2 = 0.63$, and the standard deviation $s = 0.79$.

This value tells us that the students ages are quite close together.

In summary,

To get the mean:

- Add up all the data
- Divide it by the number of terms.

To get the standard deviation:

- Square all the items of data
- Add all these values together   (value 1)
- Find the value of the the mean squared
- Multiply this value by the number of items in the data   (value 2)
- Subtract value 2 from value 1
- Divide this answer by the number of items -1   (value 3)
- Get the square root of this value

Grouped Data

Consider the following case of 449 people attending a film. Their ages are known, but grouped by how many are in the teens, twenties, etc. for each decade:

10 – 20	51
20 – 30	120
30 – 40	150
40 – 50	75
50 – 60	43
60 – 70	10

The numbers in each group are called frequencies as before, and this kind of grouped data is again known as a frequency distribution.

The question now arises as to how we can calculate a value for the mean or the standard deviation for a frequency distribution.

Clearly, since we do not have the original values, the average and standard deviation can only be estimated. The way to proceed is as follows.

To estimate the sum of all the ages, work as if each person in the age group 10 to 20 is age 15, each person in 20 to 30 is 25, etc.

So we will multiply each midpoint, by the number of people in that group, the frequencies. These numbers are totalled to give an estimate of the total of all the ages.

We are now working from this table:

Age Classes	**Frequencies**	**Midpoints**
10 – 20	51	15
20 – 30	120	25
30 – 40	150	35
40 – 50	75	45
50 – 60	43	55
60 – 70	10	65

The first, trivial calculation is to find the total *n*. This is simply the sum of the frequencies:

$$n = \sum_i f_i = 449$$

The sum of the midpoints times the frequencies is:

$$51\times15 + 120 \times 25 + 150 \times 35 + 75 \times 45 + 43 \times 55 + 10 \times 65 = 15{,}405.$$

This is an estimate of the sum of the ages.

This in turn will give us an estimate for the mean, by dividing it by the total *n*:

$$15,405 / 449 = 34.31$$

An approximation can be found for the sum of the squares in a similar way:

$$51\times225 + 120\times652 + 150\times1225 + 75\times2025 + 43\times3025 + 10\times4225$$
$$= 597,665$$

An estimate for the standard deviation can now be calculated. Substitution in the equation for the standard deviation gives

$$s^2 = \frac{597665 - (449\times34.31^2)}{448} = \frac{69113}{448} = 154.27$$

$$\Rightarrow \quad s = \sqrt{154.27} = 12.42$$

A Summary of Grouped Data Calculations

Let us rewrite the equations we have been using for the mean and standard deviation for the case of Grouped data.

The mean and standard deviation may be calculated with some error but in less time as follows.

Let there be M classes in the frequency distribution.

Let the numbers $f_1, f_2, f_3,, \ldots f_M$, be the frequencies, and let the numbers $m_1, m_2, m_3,, \ldots m_M$, be the midpoints of the groups.

The *frequency average* is

$$\bar{x} = \frac{\sum_{i=1}^{M} f_i m_i}{M}$$

In a similar way, the *frequency standard deviation* can be calculated as

$$s^2 = \frac{\sum_{i=1}^{M} f_i m_i^{\,2} - M\left(\overline{x}\right)^2}{M - 1}$$

This equation is the same as the previous one for *s*, except that the approximation for the sum of the squares found from the midpoints and frequencies is used instead of the actual values.

Example The following are the lifetimes of components produced by two companies. Find the average and standard deviation for each data set.

Lifetimes (hrs)	**A**	**B**
0 to 5	2	58
5 to 10	8	25
10 to 15	19	20
15 to 20	59	7
20 to 25	26	6
25 to 30	7	4
30 to 35	3	2

For the first data set, the sum of the midpoints times the frequencies is:

$$2.5 \times 2 + 7.5 \times 8 + \ldots + 32.5 \times 3 = 2210$$

Dividing this by the total, 124, gives the estimate for the mean of 17.82.

The estimate of the standard deviation goes as follows. The ‘sum of the squares’ is:

$$\sum_i f_i m_i^{\,2} = 2\times 2.5^2 + 8\times 7.5^2 + \ldots + 3\times 32.5^2 = 43125$$

Putting this in the equation for *s*:

$$s^2 = \frac{\sum_i f_i m_i^{\,2} - n\left(\overline{x}\right)^2}{n-1} = \frac{43125 - 124\times 17.82^2}{123}$$

This works out to be $s^2 = 3748.5 / 123 = 30.476$, so $s = 5.52$.

Do the same calculations for the second data set now!!

For the second data set, the first sum, of the midpoints times the frequencies, is 1,015, giving an estimate for the mean of 8.32.

The estimate of the standard deviation goes as follows. The ‘sum of the squares’ is:

$$\sum_i f_i m_i^{\,2} = 58 \times 2.5^2 + 25 \times 7.5^2 + \ldots + 2 \times 32.5^2 = 15212.5$$

Putting this in the equation for s:

$$s^2 = \frac{\sum_i f_i m_i^{\,2} - n\left(\bar{x}\right)^2}{n-1} = \frac{15212.5 - 122 \times 8.32^2}{121}$$

This works out to be $s^2 = 55.93$, so $s = 7.48$.

Medians and Quartiles

Aside from the mean and standard deviation, there are other numbers that describe the distribution of a large dataset.

These are the **median** and the **quartiles**. They will be useful in their own right, and when they are compared to the mean.

The essential difference between these numbers and the mean and standard deviation is that the median and quartiles are concerned with the order of a dataset.

The Median

The median is that value which is mid-point in the data. Half of the numbers are greater than median, and half are less.

This is straightforward in the case of an odd number of values, as in the following case:

1, 4, 5, 6, 10, 11, 12

These numbers are arranged in order, so it can be seen that 3 numbers are above 6, and 3 below. Thus 6 is the median value.

The situation is slightly different if there are an even number of values, for example

4, 5, 6, 9, 12, 23, 25, 30

Here the 9 is the 4th number and the 12 the 5th, out of 8 numbers.

Therefore the midway point of the data lies between these two values.

For the median, the value actually quoted is the midpoint of 9 and 12:

$$(9+12)/2 = 10.5$$

The Mode

The mode is the value that occurs the most frequently in a data set

The mode is in general different from the mean and median, and may be very different for strongly skewed distributions.

The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most ambiguous case occurs in uniform distributions, wherein all values are equally likely

The Quartiles

Once a list of numbers has been divided into two groups by the median, the quartiles take this a step further.

The first quartile divides the lower half, the third quartile divides the upper group, in exactly the same way as the median did for the original list.

Example Find the median and quartiles for the following list of numbers: 7, 8, 12, 6, 5, 3, 9

Step 1: The first step is to arrange them in order: 3, 5, 6, 7, 8, 9, 12

Step 2: Find the median: There are 7 numbers so the 4th is the median, 7

Step 3: Get the 1st quartile from the first half: 3, 5, 6, 7

The question arises here as to whether the 7 should be included in the lower list or the higher list.

Logically it should be in both or neither – the convention adopted is that it will be included in both.

The quartile is then between 5 and 6, so it is (5+6)/2 = 5.5

Step 4: Get the 3rd quartile from the second half: 7, 8, 9, 12

The quartile is between 8 and 9, so it is (8+9)/2 = 8.5

Example What are the median and quartiles of the following set of heights: 1.7, 1.8, 2.0, 1.6, 1.5, 1.8, 1.9, 1.7

Do this question now yourselves!!

Sorting these in order gives: 1.5, 1.6, 1.7, 1.7, 1.8, 1.8, 1.9, 2.0.

The median will be between 1.7 and 1.8, i.e. 1.75.

The first quartile will be between 1.6 and 1.7, and so is 1.65.

Similarly the third is between 1.8 and 1.9, and so is 1.85.

Comparing the Mean with the Median and Quartiles

The value of the mean compared to the median and the two quartiles tells us something about the distribution of the numbers in the data.

We are comparing the values of numbers, with their order.

The conclusions are:

- If the average is close to the median, then there are as many figures above average as below.

- If the average is closer to the 1st Quartile, then a few of the lower figures are ‘dragging’ the average down.

- If it is closer to the 3rd Quartile, then a few of the higher numbers are pulling the average up.

The Median for Grouped Data

Recall the frequency distribution of ages of a group attending a film:

Age classes	Frequencies
10 – 20	51
20 – 30	120
30 – 40	150
40 – 50	75
50 – 60	43
60 – 70	10

In a similar way to the mean and standard deviation, the median and quartiles for grouped data must be estimated, since the numbers in the list above are unknown.

If the numbers in the list, going from 10yrs to 70yrs, were equally spaced out, a good estimate of the median would be

$$60\text{yrs}/2 + 10\text{yrs} = 40.0\text{yrs}$$

But a better estimate could be found, if we repeat this procedure for the group the median is in. In other words, see what its place is within one of the groups, and from this, estimate how far it is above that groups’ lower bound.

To see which group the median is in, work out the *cumulative frequencies*, that is, the number in the group plus those in previous groups.

Here are the ages of the cinema group, with the cumulative frequencies:

Age Classes	Frequencies	Cumulative Frequencies
•10 – 20	51	51
•20 – 30	120	171
•30 – 40	150	321
•40 – 50	75	396
•50 – 60	43	439
•60 – 70	10	449

In this case, there are 449 numbers, so the median would normally be the 225th number. Age group 30 – 40 has the 172nd to the 321st numbers, so the median is in this group.

What is the median's place within the group? It is the 225th number, there are 171 in the previous 2 groups, so the median is the 225 – 171 = 54th number in group 30 – 40.

If the numbers *within* this group were evenly spread out, the average separation between them would be 10yrs/150.

Thus the median is the 54th number in the 30 – 40 age group, and the numbers, on average, are separated by 10yrs/150, so the estimate is 54×10yrs/150 =3.6yrs above the lower bound.Thus is the required estimate for the median is:

$$30\text{yrs} + 3.6\text{yrs} = 33.6\text{yrs}.$$

Repeating this procedure more generally will give us an equation for this estimate.

The steps are:

From the frequency distribution, work out the cumulative frequencies.

Decide which class the *n*/2 data point occurs in. The Median is given by the formula

$$\text{Median} = \left(\frac{n}{2} - cf\right)\frac{w}{f} + L$$

where the symbols are:

- *cf* is the cumulative frequency of the class below that of the median value,
- *f* is the frequency of the class containing the median value,
- *L* is the lower bound of this class,
- *w* is the width of each class.

Why does this work?

Consider a data set, a list of n figures $x1$, $x2$,…, xn, for which a frequency distribution is available, and the raw data is unknown.

The first step in the procedure to estimate where the median might be is to decide, using the cumulative frequencies, which class it falls in.

Let L be the lower boundary of this class, and let w be the width of the classes. The values of the data in this class are then from L to $L + w$.

If cf is the cumulative frequency of the class before, and there are f data values in this class, then the numbers in this class are the cf + 1st to the cf+fth places.

The median is the $n/2 – cf$ number in this class.

Now assume that the data points are evenly distributed within this class. They are each separated by a step of,

$$\frac{w}{f}$$

where *w* is the width of the class.

To work out how much above the lower bound the median value is, its place in the class is multiplied by this increment.

The estimate is then the place number, times the increment, added to the lower bound. The resulting equation for the estimate *MD* is then,

$$MD = L + \left(\frac{n}{2} - cf\right)\frac{w}{f}$$

The Quartiles for Grouped Data

To calculate the quartiles, as with the median, work out the cumulative frequencies. Decide which class the $n/4$ and $3n/4$ data points occur in. The first and the third quartile are given by the formulae:

$$\mathrm{Q1} = \left(\frac{n}{4} - cf\right)\frac{w}{f} + L \quad \text{and} \quad \mathrm{Q3} = \left(\frac{3n}{4} - cf\right)\frac{w}{f} + L$$

where

- cf is the cumulative frequency of the class below that of the quartile,
- f is the frequency of the class containing the quartile,
- L is the lower bound of this class,
- w is the width of each class.

Example Recall the lifetimes of components produced by a company, sorted into groups. The frequency distribution was:

Classes	**Frequency**	**Cumulative Frequency**
•0 to 5	9	9
•5 to 10	15	24
•10 to 15	20	44
•15 to 20	30	74
•20 to 25	20	94
•25 to 30	6	100

Find the median and quartiles for each data set.

The cumulative frequencies are shown above, calculated from the frequency distribution. In this example, there are 100 numbers, so the median would normally be the 50th number.

The first three groups contain 44 numbers, so the 50th is in the 4th group, 15 to 20. The numbers *f*, *cf*, *w*, and *L* must now be found.

The frequency of the class containing the median is 30, so:
$f = 30$.

The cumulative frequency of the previous class is 44, so:
$cf = 44$.

The lower bound of the class containing the median is 15, so:
$L = 15$.

The width of the classes is 5:
$w = 5$.

The Median is given by the formula: $\text{Median} = \left(\frac{n}{2} - cf\right)\frac{w}{f} + L$

Putting in the values gives: $\text{Median} = \left(\frac{100}{2} - 44\right)\frac{5}{30} + 15$

$$= (50 - 44)\frac{5}{30} + 15 \qquad = \frac{6 \times 5}{30} + 15 = 1 + 15 = 16$$

The median is then 16.

For the first quartile, $n/4$ is 25, so Q1 would normally be the 25th number. This means it is in the group 10 to 15.

Again, the value of w is 5.

The frequency is $f = 20$.

The value of the lower bound is $L = 10$.

The previous cumulative frequency is $cf = 24$.

The first quartile is given by the formula: $Q1 = \left(\frac{n}{4} - cf\right)\frac{w}{f} + L$

This becomes: $Q1 = (25 - 24)\frac{5}{20} + 10$

$$= 1 \times \frac{5}{20} + 10 = {}^{1}\!/_{4} + 10 = 10.25$$

The first quartile is 10.25.

For the third quartile, $3n/4$ is 75, so Q3 would be the 75th number. This means it is in the group 20 to 25.

As before, $w = 5$.

The frequency is $f = 20$.

The value of the lower bound is $L = 20$.

The previous cumulative frequency is $cf = 74$.

The third quartile is given by the formula: $\mathrm{Q3} = \left(\frac{3n}{4} - cf\right)\frac{w}{f} + L$

This becomes: $\mathrm{Q3} = (75 - 74)\frac{5}{20} + 20$

$$= 1 \times \frac{5}{20} + 20 = {}^{1}\!/_{4} + 20 = 20.25$$

The third quartile is 20.25.

Statistical Graphs and Charts

Describing Data

- Can describe (e.g. the distribution of a set of data) data in two general ways
  - Graphical: pictures of the data
  - Numerical: calculate quantities which summarise data

The following numbers show the percentage return on an ordinary share for 23 consecutive months:

0.2 -2.1 1.0 0.1 -0.5 2.4 -2.3 1.5 1.2 -0.6 2.4 -1.2
1.7 -1.3 -1.2 0.9 0.5 0.1 -0.1 0.3 -0.4 0.5 0.9

If you were an investor or financial journalist, it would be useful to be able to make some general statements about the returns on this share – in other words, to describe the returns in some way.

How would you describe these figures? Since there are only two lines of numbers, it is easy to see that the largest is 2.4 and the smallest is -2.3, but this doesn’t give any information about how the numbers are distributed between this maximum and minimum.

Describing Data

- **Frequency Distribution**
  - To ‘draw a picture’ of a set of data we must first split an interval enclosing the smallest and largest values into several non-overlapping classes of equal width. For the ordinary share return data the classes could be:

-3 to less than -2

-2 to less than -1

-1 to less than 0

0 to less than 1

1 to less than 2

2 to less than 3

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins

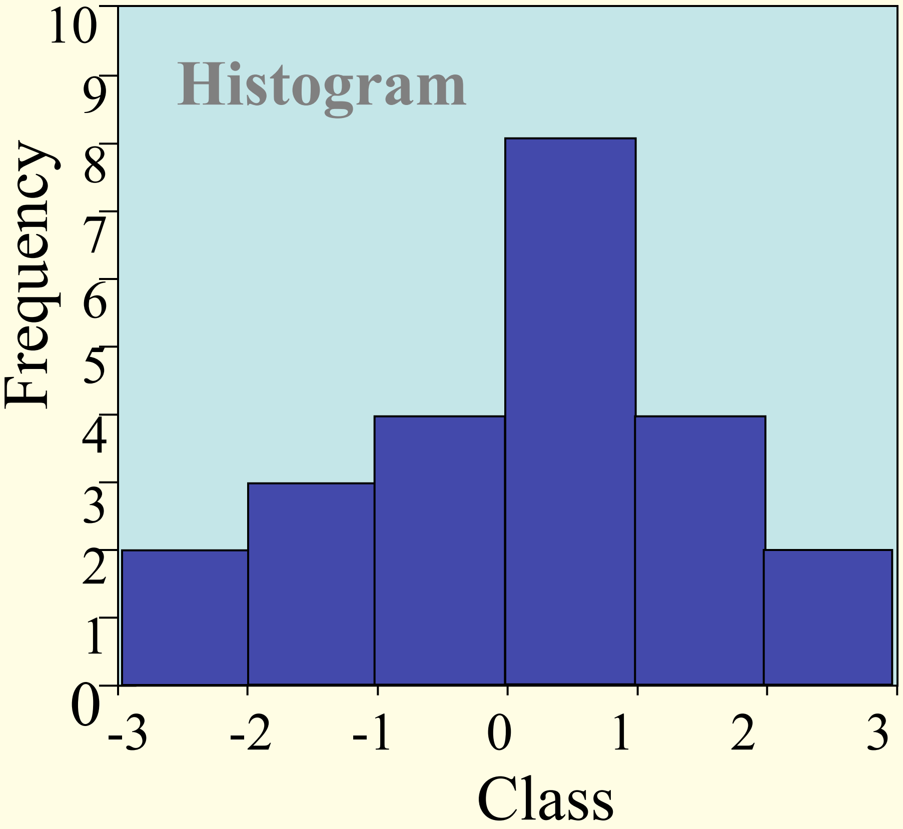
- **Relative Frequency**: Is the proportion of the data which falls in each class
  - We calculate the relative frequencies by dividing by the total number of items of data.
  - The relative frequencies add up to 1.

- For the returns data we obtain the following

Class	Frequency	Relative Frequency
-3.0 to -2.1	2	2/23
-2.0 to -1.1	3	3/23
-1.0 to -0.1	4	4/23
0.0 to 0.9	8	8/23
1.0 to 1.9	4	4/23
2.0 to 2.9	2	2/23

Histograms

Either the frequencies or the relative frequencies can be drawn pictorially in a **histogram**. Below is a histogram of the frequencies of the returns data.



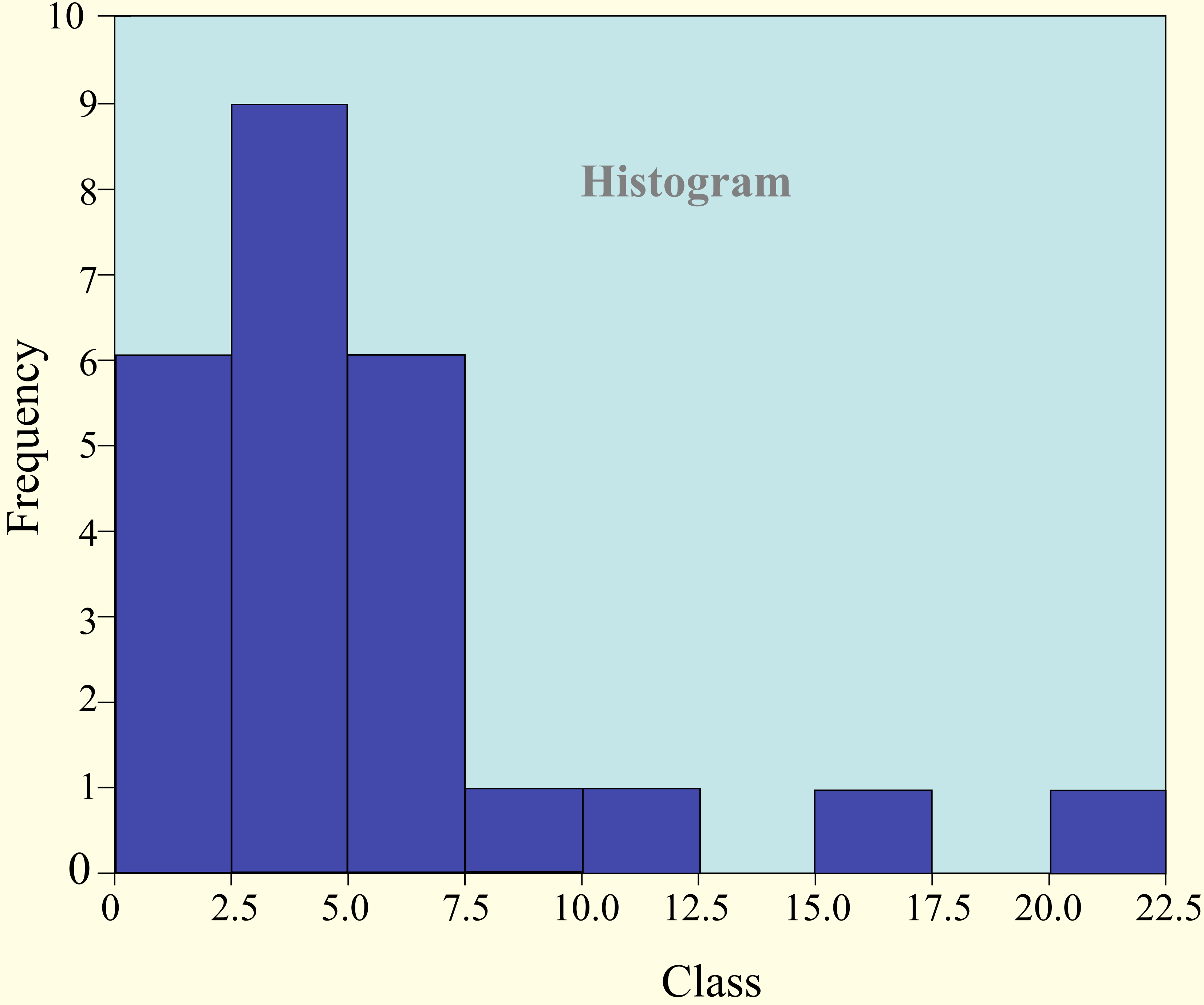
It is now immediately apparent that the distribution of the returns lies between -3 and 2.9 and the histogram has an inverted U-shape.

A histogram of the relative frequencies has exactly the same shape with different labelling of the *y*-axis.

Example The following data gives the time in days it takes a manufacturing firm to supply price quotes to customers. Work out the frequencies and relative frequencies and draw a histogram.

2.36 5.73 6.60 10.05 5.13 1.88 2.52 2.00 4.69 1.91 6.75
3.92 3.46 2.64 3.63 3.44 9.49 4.90 7.45 20.23 3.91 1.70
16.29 5.52 1.44

Class	Frequency	Relative Frequency
0.00 to 2.49	6	0.24 (6/25)
2.50 to 4.99	9	0.36
5.00 to 7.49	6	0.24
7.50 to 9.99	1	0.04
10.00 to 12.49	1	0.04
12.50 to 14.90	0	0.00
15.00 to 17.49	1	0.04
17.50 to 19.99	0	0.00
20.00 to 22.49	1	0.04



Histogram

Frequency
0
1
2
3
4
5
6
7
8
9
10
Class
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5

Histograms

- From the histogram we can see that it most common for price quotes to take only a few days but occasionally some may take much longer.

- **Skewed Distribution**: A histogram does not have a peak in the centre but is peaked on one side.
  - ***skewed to the left*** - when the peak is on the right
  - ***skewed to the right*** - **w**hen the peak is on the left

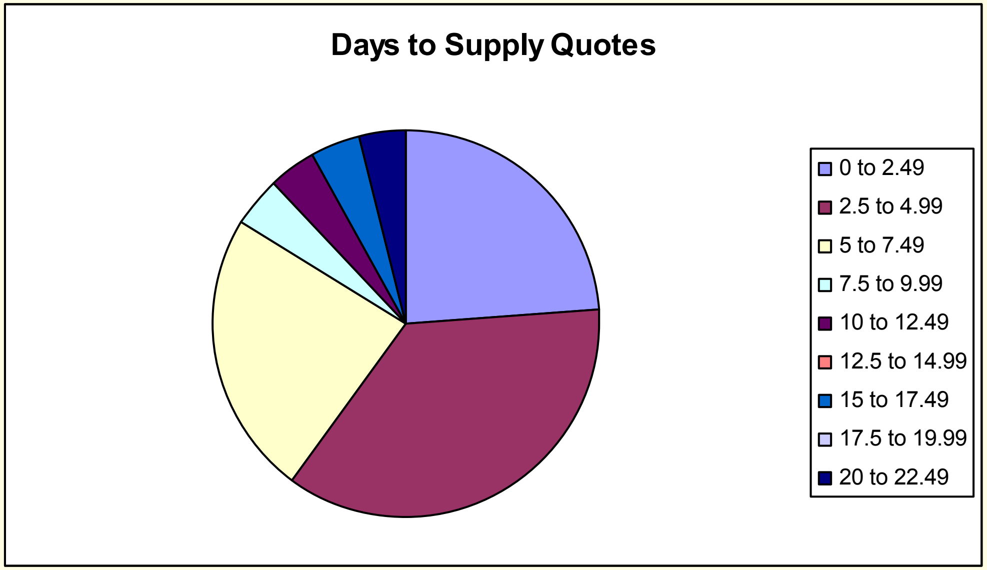
- **Symmetric**: When there is no obvious skew, we say that the distribution is roughly symmetric

Choosing the classes

- What factors should we consider when choosing the class widths?
- The first thing to realise is that there are no hard and fast rules !
- It is usual, where possible, to choose equal class widths, and the number of classes should be large enough to reveal details in the structure of the data, but small enough to ensure that some classes contain a reasonable percentage (perhaps 10%) of the data points.
- Typically there are between five and fifteen classes.
- Finally, the class widths should be based around whole numbers or ‘sensible fractions’ such as halves or quarters.

Pie Chart

- Another pictorial representation
- A **pie chart** is a circular depiction of data where the whole 'pie' representing 100% of the data, is divided into 'slices' whose sizes are proportional to the size of the categories they represent



This pie chart represents the price quote data. It is immediately apparent that the categories 0 to 2.49 and 5 to 7.49 each contain just under a quarter of the data, and that the category 2.5 to 4.99 contains just over a third of the data.

Pie Chart

- To draw the pie chart we simply divide the total angle of the circle, that is 360°, into the appropriate percentages.
- For example, the relative frequency of the first category of the price quote data is 0.24.
- We therefore multiply 360 by 0.24, which gives us an angle of 86.4°.
- This procedure is then carried out for each of the classes and the pie chart can then be drawn.
- Note that if there are no data in a particular category, then since the relative frequency will be 0, so to will the angle. Therefore this category does not appear on the pie chart as expected.

Bar Charts, Time Series Plots and Scatter Plots

- Here we will introduce some other ways of displaying data which are only suitable when the data has one of the following characteristics:

(i) The data is not numerical, but comes in categories,

(ii) The order of the data is important – usually because the data set is a series of values occurring through time, such as monthly inflation rates,

(iii) The data occur in pairs and the relationship between the values in each pair is of interest.

Bar Charts

- The data we have used so far has been numerical or quantitative, so to draw a histogram of stem and leaf diagram we had to split it into classes.

- If the data is not numerical, but records an attribute or quality, then it is in classes already. We call such data qualitative or categorical data.

- Some examples of qualitatice data are:
  - The counties from which students in your college are from
  - The types of cars driven by staff of a company
  - The preferred drink of students at ITB

- As with quantitative data, the number of data items in each class are called frequencies and the proportion of data items in each class is called the relative frequency.

Example

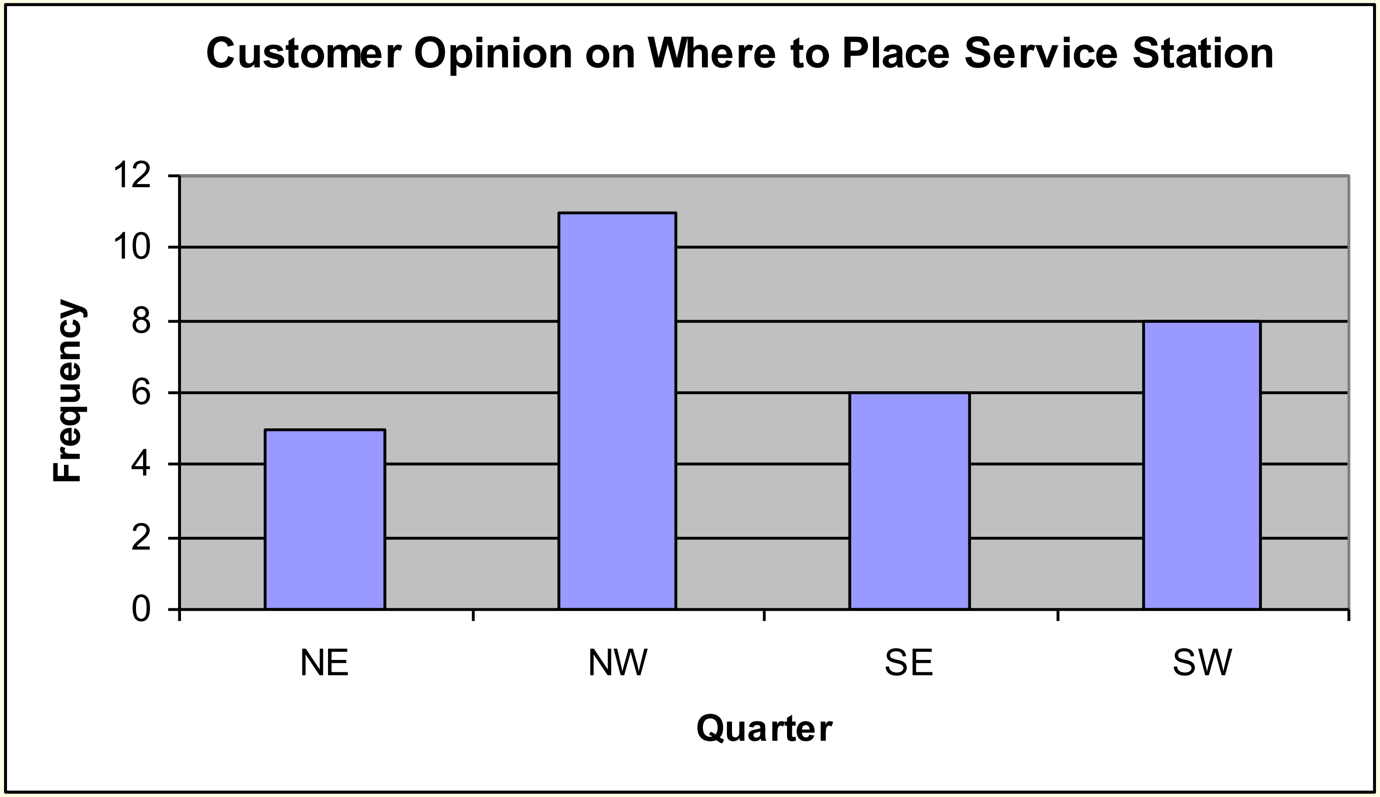
An oil company wants to open a new service station to serve the resident population of a city. There are four possible sites which lie in the NW, NE, SW and SE quarters of the city respectively. In an initial survey the company stop 30 motorists in the city centre and ask them which site they Would be most likely to use. The results were as follows:

Customer	Quarter	Customer	Quarter	Customer	Quarter
1	NW	11	SW	21	NW
2	NE	12	NW	22	SW
3	SE	13	SE	23	SE
4	NW	14	SW	24	SW
5	NW	15	NW	25	NW
6	SE	16	NW	26	SW
7	NE	17	NE	27	SE
8	NE	18	NW	28	SE
9	NW	19	SE	29	NW
10	SW	20	SW	30	NE

The frequencies and relative frequencies are then:

Quarter	Frequency	Relative Frequency
NE	5	0.167
NW	11	0.367
SE	6	0.200
SW	8	0.267

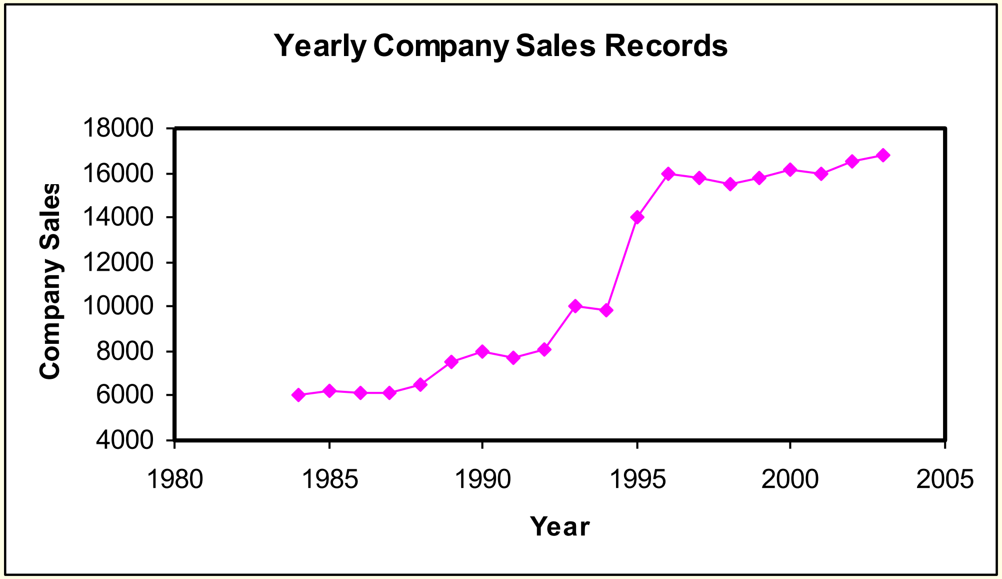
And the **bar chart** looks like this:



A bar chart merely displays these frequencies as shown. Although it looks very similar to a histogram, it is important to realise that a bar chart has a finite number of categories along the axis, whereas a histogram has a continuous numerical scale.

Time Series Data

- Sequential Data: **Time Series** Data - graph usually has time (months or years) on the horizontal axis and the values of interest on the vertical axis.
- e.g. The graph below shows company sales data for the last 20 years.



Such data is usually recorded at equal intervals of time, so it is called a time series.

The crucial element is that the sequence of the data must be preserved.

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide

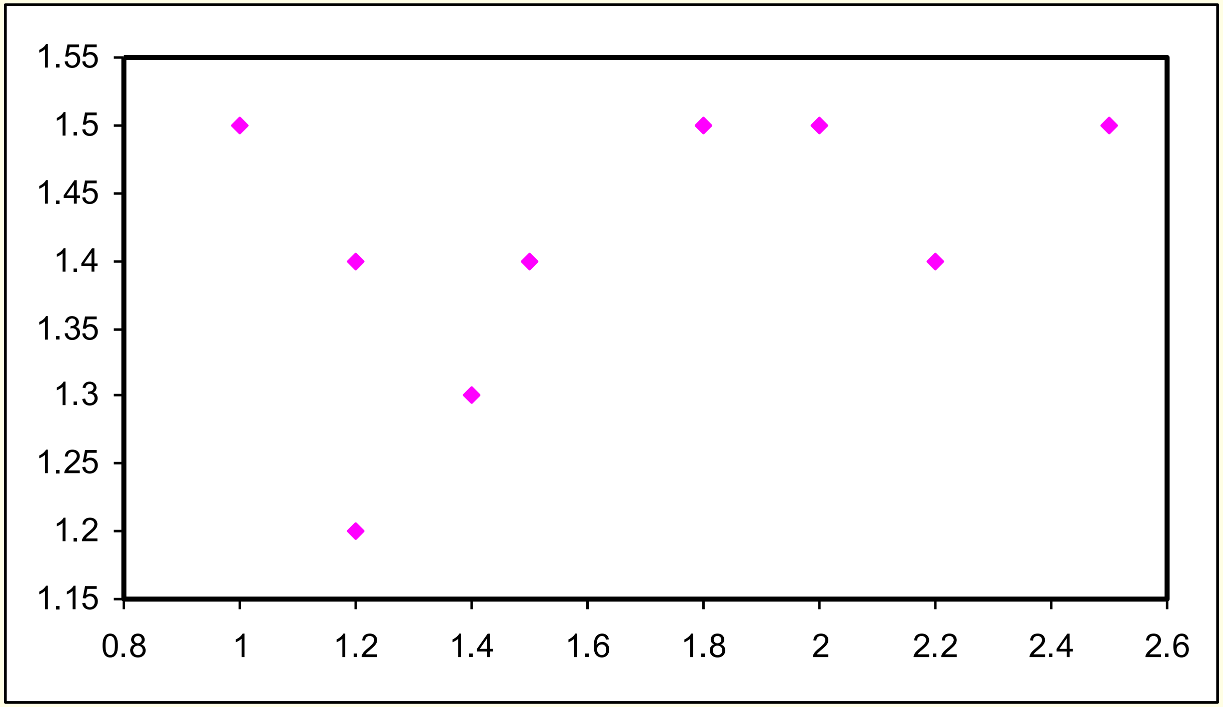
Scatter Plots

- Data which occurs in pairs: **Scatter plots** - The following data gives the percentage returns on 2 ordinary shares for 9 consecutive months:

- Notice that the data comes in pairs, a pair for each month (1.4, 1.3), (1.2, 1.4), and so on.

Month	Share 1	Share 2
1	1.4	1.3
2	1.2	1.4
3	2.2	1.4
4	1.5	1.4
5	1.0	1.5
6	1.2	1.2
7	1.8	1.5
8	2.5	1.5
9	2.0	1.5

The most effective way of presenting paired data like this is to plot the pairs as coordinates on a graph. This is called a **scatter plot**. A scatter plot of the share return data is shown below:



This shows us that when the first share's return is large, the second share's return tends to be large also. Therefore, we can say that the returns on the two shares appear to be correlated.