

# SECTION 1

## DESCRIPTIVE STATISTICS

### 1.1 INTRODUCTION

#### 1.1.1. The Role of Statistics

Much of the information that we have to deal with in life is **VARIABLE**.

e.g. Profits/Sales/Costs vary with time

Heights/Weights/Opinions vary between individuals

Mortality rates/social structure/income vary between regions

Performance/Quality/Durability vary between manufactured items.

In some situations, (e.g. physical appearance, opinions and beliefs) variation is **desirable**. However in other situations (e.g. in a manufacturing environment, in medicine) it is not. In all situations it is useful to be able to *measure variation, know what is causing it and make allowances for it*.

Consider, for example, a radio battery. Performance (as measured say by lifetime) will vary from battery to battery. Reasons for this include:

- *raw material from which it is made*
- *processes involved in manufacture*
- *storage conditions*
- *transportation*
- *usage*

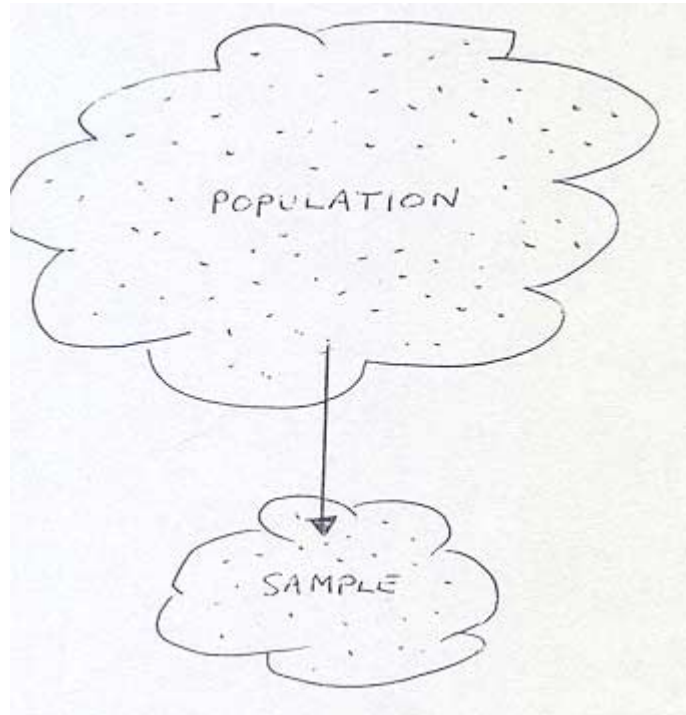
**Statistics is about understanding and communicating about VARIABILITY.**

To put it another way, statistical methods are needed to deal with situations where there is an element of **UNCERTAINTY**.

In this course, we examine some statistical techniques for dealing with variability and introduce the ideas of probability for modelling uncertainty.

### 1.1.2. Populations and Samples

A fundamental distinction must be made between a **POPULATION** and a **SAMPLE**.



A population is the entire collection of units/individuals/outcomes in which we are interested. It is usually very large (and sometimes infinite) so to find out what's going on in the population we observe a sample - a representative subset.

The key word here is **representative**. A sample should be 'the population in miniature'. Then by examining a sample we can draw conclusions about the population. Such conclusions, however, cannot be made with 100% certainty and are stated in terms of probabilities.

To summarise:

*We really want to know about populations. To find out about them, we look at samples.*

**STATISTICS** takes information obtained from a sample of the data, and tries to make inferences about the corresponding population from which the sample was drawn (sample → population).

**PROBABILITY** takes population information and tries to make predictions about what will occur within any particular sample (population → sample.)

### 1.1.3. **Structure of Module**

This module consists of:

Section 1	Describing samples
Section2	Modelling populations
Section 3	Applications to Reliability
Section 4	Drawing conclusions about populations from samples
Section 5	Looking at relationships

More details can be found on the Module Plan.

## 1.2. TYPES OF DATA

### 1.2.1. Some Definitions

A **VARIABLE** is defined as any characteristic which varies from one member of the population or sample to another.

eg. weight, length, lifetime of a particular manufactured item.

**DATA** are then a set of observations taken on a variable.

A **RANDOM VARIABLE** is something which varies from unit to unit with an element of 'randomness' or unpredictability. (For example, the number of days in a year is variable, either 365 or 366, but it is not a random variable). We will be concerned with analysing the behaviour of random variables.

As far as a set of data is concerned, Statistics is concerned with:

- *collection*
- *display*
- *analysis*
- *interpretation.*

Data can be divided into two basic types:

#### QUANTITATIVE

Essentially numerical data which can, for example, be arranged in order and averaged.

(e.g. weight of item, age of person)

and

#### QUALITATIVE

This is non-numerical data. Such data may be coded to numbers but they only serve as labels and cannot, for example, be arranged in order.

(e.g. Is item of adequate quality? What is persons hair colour?)

Within each of these two categories, data can be further subdivided, as follows:

### 1.2.2. Quantitative Data

(ie. Numerical data) can be subdivided into

(i) Discrete

- Data which can only take specific numerical values.  
(e.g.number of machine breakdowns in a week: 0,1,2,...)
- Usually arises from counting something.

and

(ii) Continuous

- Data can take any value within a range.  
(e.g. Height of person: 4' - 7'; Lifetime of component: 0→? hrs).
- Usually arises from measuring something.
- Data has been rounded.

### 1.2.3. Qualitative Data

(ie.Non-numerical data) can be subdivided into:

(i) Categorical data

e.g. Item is manufactured in 4 different colours; Red, Blue, Green and Yellow. Let X, the variable we are 'measuring' (or our *random variable*, or 'variable of interest') be the colour of any item selected for inspection. Then X can take the 4 values R, B, G, or Y.

and

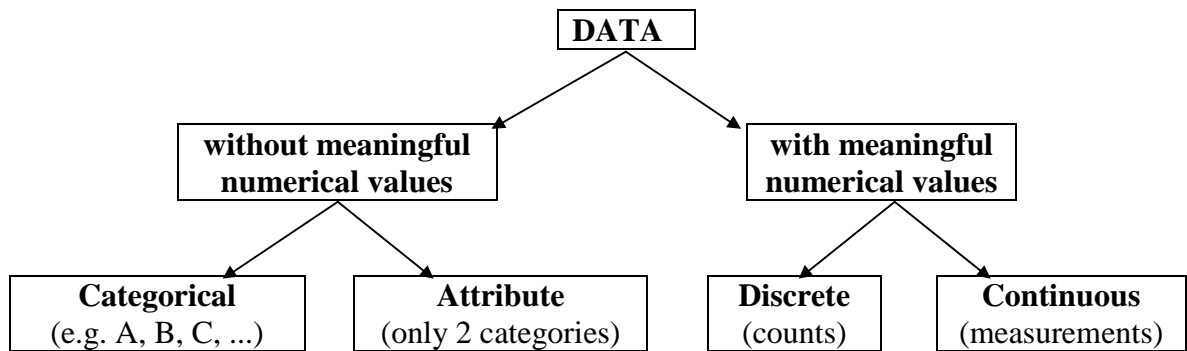
(ii) Attribute Data

e.g. Let X = whether item is accepted as OK by Quality Control.

Then X can only take 2 values; Yes or No.

Attribute data is really an important special case of categorical data.

This information can be summarised as follows:





In practice we often collect data in the form of observations or measurements on some variable of interest. The entire collection or set of measurements is known as the **raw data**. Very often, the raw data set is extremely large and unwieldy to deal with and is difficult to interpret. Therefore, we often try to reduce the amount of information by summarising it in some way. There are basically three ways of doing this:

- by tables (section 1.3)
- by graphs (section 1.4)
- by numbers (section 1.5)

## 1.3 TABULAR REPRESENTATION OF DATA

A first step is often to construct a **frequency distribution**. This is a table that gives the number of times each value occurs in the data (that is, the frequency of occurrence) or, if there are many different values, the numbers of occurrences within certain ranges of values.

Thus it shows how the total frequency (ie. the total number of observations we have) is distributed across the range of values.

- The purpose of a frequency distribution is to show the pattern or 'shape' of the data.
- The construction of a frequency distribution from raw data is best achieved by using a **tally chart**.

### 1.3.1. Ungrouped Distributions

When we are dealing with qualitative data, or discrete data with a fairly small number of distinct values, we can count the number of observations (frequency) for each value.

As well as calculating the frequency, or number of occurrences of a particular value of the random variable, we may also be interested in the **relative frequency** of that value of the random variable. This is calculated as

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Total frequency}}$$

and is often expressed as a percentage.

Relative frequencies are particularly useful when comparing sets of data.

In addition, with quantitative data, the **cumulative frequency**, (defined as the frequency up to and including a particular value) is a useful measure of a data set, and is commonly calculated instead of the frequency.

#### Example

Referring to the data in the previous example, we can calculate for example the *frequency distribution of re-work possibility*

Random Variable, X	Tally	Frequency	Relative Frequency (%)
Rework possible		12	40%
Rework not possible		18	60%

For example, in the present case there are a total of 30 observations, of which 12 have the possibility of a rework.

i.e.  $\frac{12}{30} = 0.4$  of the total observations

(or 40%) have the possibility of a rework

### 1.3.2. **Grouped Distributions**

If the data is continuous, (or discrete covering a wide range of values), it will be necessary to group the data into classes, in order to develop a meaningful frequency distribution.

These classes should

- (i) cover the range of the data
- (ii) not overlap
- (iii) if possible, be the same size or width.

Aim at somewhere between about 5 and 15 classes for a reasonable summary of the data.

#### **Example**

The time to breakdown (hours) of 36 generators was measured and the results are as follows:

512	126	2759	1783	507	832
1395	1860	323	898	371	109
2618	398	1461	462	1029	783
901	1880	586	1113	350	221
1387	716	621	1650	91	1565
699	1193	1218	826	264	703

Grouping the data, we obtain the following frequency distribution:

Time of Breakdown (hours)	Frequency
0 - < (500)	10
500 - <(1000)	12
1000 - <(1500)	7
1500 - <(2000)	5
2000 - <(2500)	0
2500 - <(3000)	2
TOTAL	36

The brackets denote '*up to, but not including*'. (Note that Excel uses a slightly different convention: '*up to, and including*' but this is of no real consequence.)

## **1.4. GRAPHICAL REPRESENTATION OF DATA**

It is often much easier to understand a data set by looking at a graphical representation of it rather than a list of values or a frequency distribution table. Different types of graphical display are appropriate for different types of data (i.e. Attribute, Categorical, Discrete or Continuous). The choice of chart will depend on the data it is required to represent and on the people it is aimed at. Whichever one is used, the following general rules should always be applied.

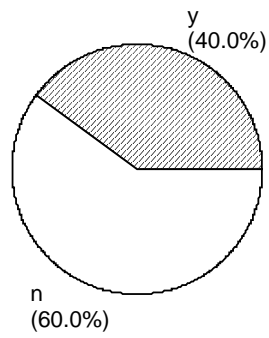
- PROVIDE A TITLE
- LABEL THE AXES CLEARLY
- INCLUDE SCALES ON THE AXES
- PRESENT THE INFORMATION CLEARLY AND UNAMBIGUOUSLY
- INCLUDE THE SOURCE OF THE DATA IF POSSIBLE

### 1.4.1. **Basic Charts**

Suitable graphical displays of the data in exercise 1 are as follows:

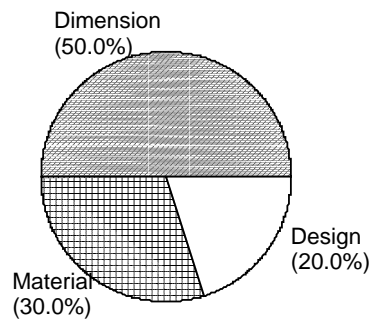
- (i) **For data on an attribute**

Possibility of re-work

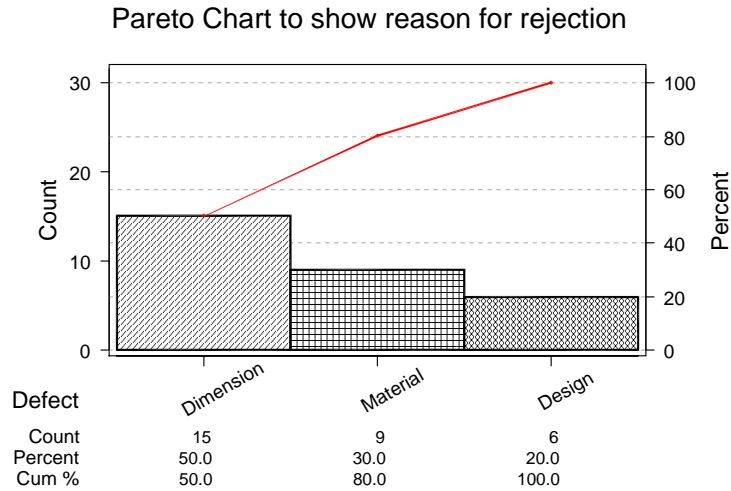


- (ii) **For data on a categorical variable**

Reason for re-work



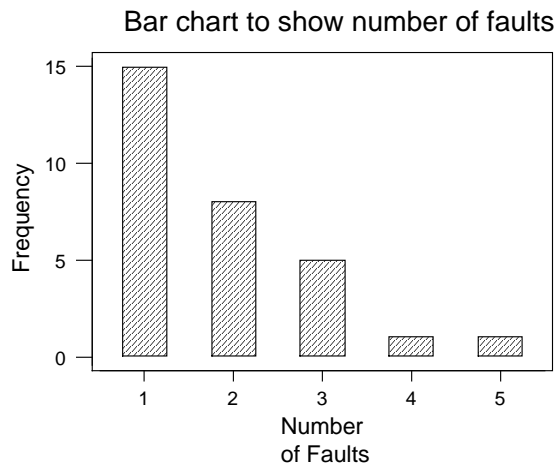
Pareto diagram (to order priorities)



Bars are arranged in decreasing order of size. Pareto charts are used extensively in industry as part of quality improvement schemes.

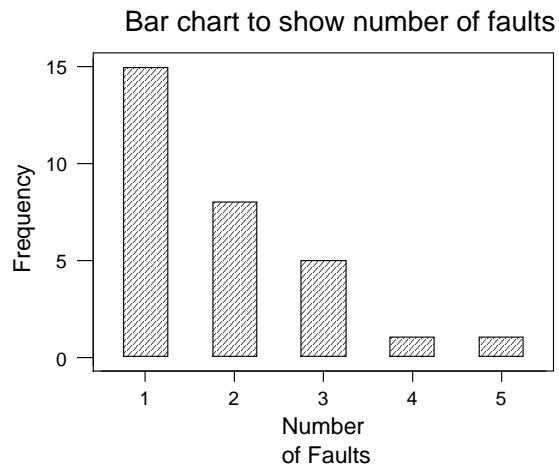
(iii) *For data on a discrete variable*

Bar Chart

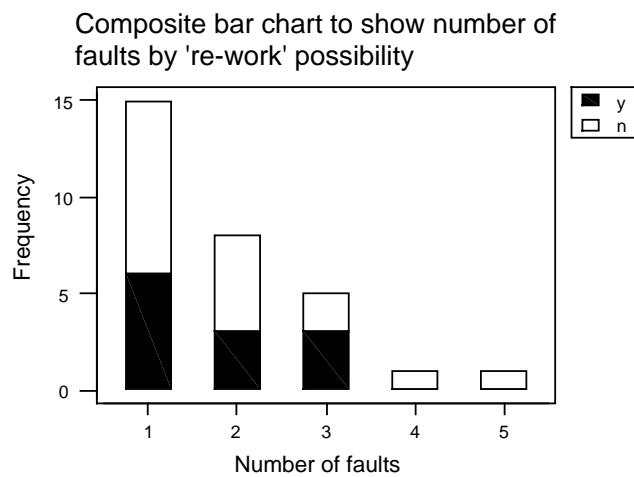


There are, in fact, a number of different bar charts which can be drawn. They are illustrated below, and should be used in the following general circumstances.

Simple bar chart : to compare the values of one quantity.

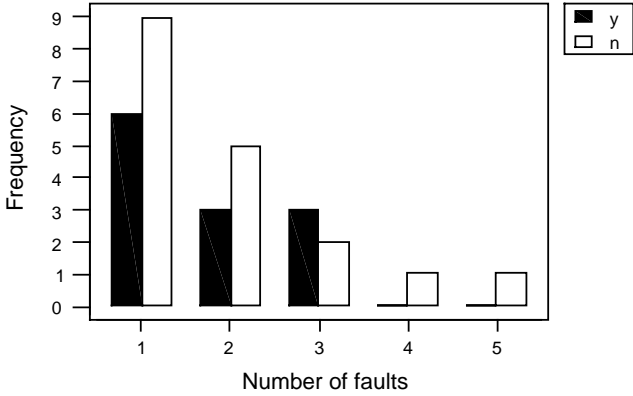


Composite bar chart : to compare the values of a total broken down into its component parts.



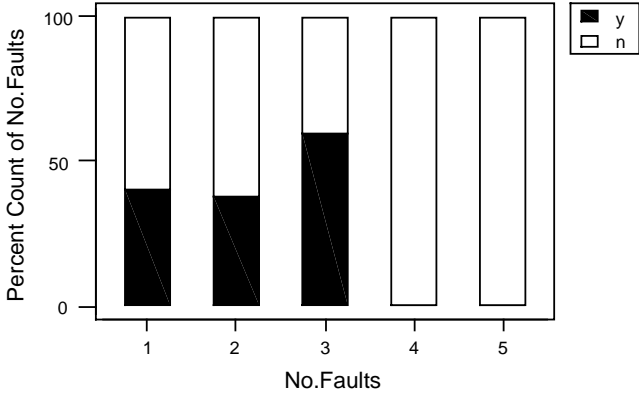
Multiple bar chart : to compare the values of more than one quantity.

Multiple bar chart to show number of faults by 're-work' possibility



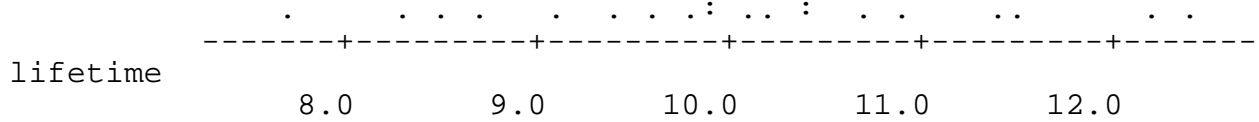
Percentage bar chart : to show how a total is broken down into its component parts on a percentage basis.

Percentage bar chart to show number of faults by 're-work' possibility



(iv) For data on a continuous variable

### Character Dotplot



Dot plots are particularly useful if there are only a few observations.

### 1.4.2. Histograms

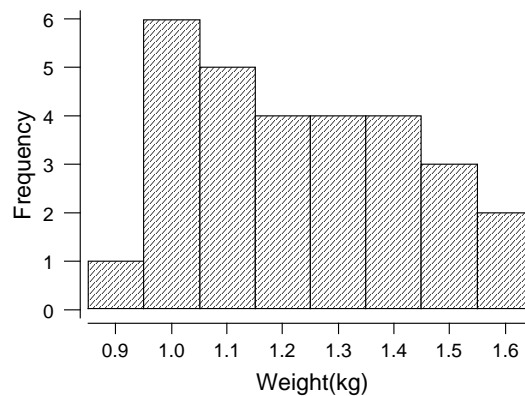
#### Equal class widths

If data are *grouped* in a frequency distribution then a histogram is preferable. Here, a continuous scale is sliced up into sections and the number of observations we have in each section is represented by a rectangle on that section. Consequently, two features of histograms which should be especially noted are:

- the area under each bar is proportional to the frequency within that group.
- the rectangles used to represent the frequencies are touching (unlike a bar chart)

If the classes in the table are all the same width then the heights of the rectangles can be used to represent the frequencies.

Histogram to show weight of rejected parts



### Unequal class widths

If the widths of the groups (class widths) into which the data has been divided are not equal, the heights of the bars must be adjusted so that the areas are in the correct proportions.

Situations where this may be necessary are as follows:

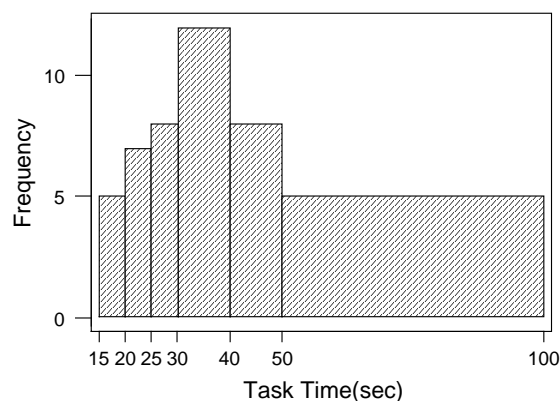
- (i) data has already been grouped;
- (ii) naturally occurring ranges (e.g. pre-school, primary, secondary school ages, etc.) effectively define the groups;
- (iii) extreme values make equal sized groups impractical e.g. house prices, salaries.

### Example

Task Time (sec)	No of Observations (frequency)
15 - (20)	5
20 - (25)	7
25 - (30)	8
30 - (40)	12
40 - (50)	8
50 - (100)	5

If we plot the data as it is, we obtain

Histogram of task times (raw data)



Comparing the data for Group 15 - (20) and Group 50 - (100), we see that in both cases the frequency is 5 (i.e. 5 people took between 15 and 20 seconds to do a particular task, and 5 people took between 50 and 100 seconds). However, when we look at the above histogram, it appears as though the frequency is much larger

in the 50 - (100) group. This is because the eye naturally compares the areas of the bars, and not their height.

So, bars of different widths must have their height adjusted accordingly. This is done by calculating the **frequency density** as follows:

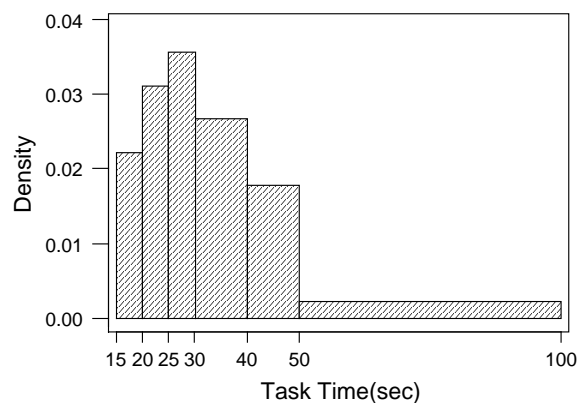
$$\text{Frequency density} = \frac{\text{Frequency}}{\text{Class width}}$$

(Note that some packages, including Minitab, use Relative Frequency/Class width).

Task Time (sec)	Class Width (sec)	Frequency	Frequency Density
15 - (20)	5	5	1.0
20 - (25)	5	7	1.4
25 - (30)	5	8	1.6
30 - (40)	10	12	1.2
40 - (50)	10	8	0.8
50 - (100)	50	5	0.1

The new (corrected) histogram is then

Histogram of task times (frequency density data)

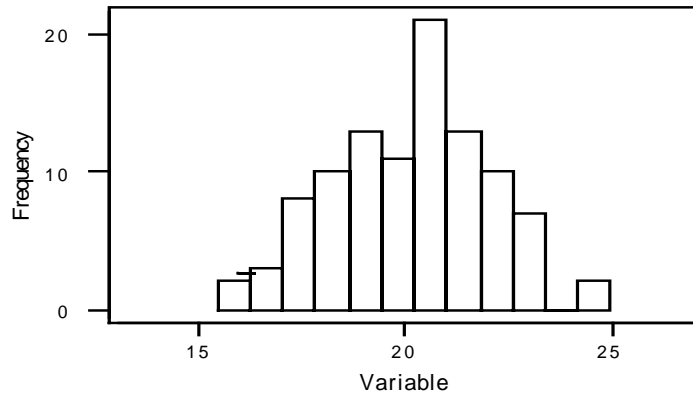


This is a fairer representation of the distribution showing, for example, that task times are most dense or concentrated in the 25-30 second range.

(Note that Minitab has plotted *relative* frequency densities i.e. the frequency densities in the above table divided by the sample size of 45. However, the shape of the histogram is the same.)

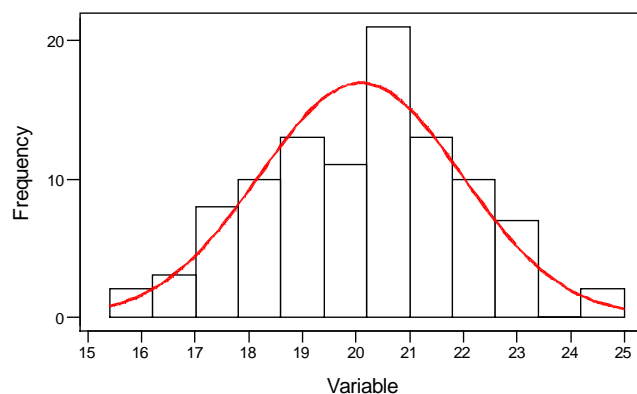
### 1.4.3. Shapes of Distributions

An alternative to a histogram for outlining the shape of a distribution is a **frequency polygon**. These may be superimposed on a histogram as follows:



Note how it is tied down to the horizontal axis by imagining a class at either end of the same width as the class next to it, with zero frequency. Frequency polygons are just line graphs of frequency (or frequency density) against class midpoint.

It should be emphasised that a frequency polygon (or histogram) is just describing the shape we have found in a *sample*. Another sample from the same population is not expected to have exactly the same polygon but it would be expected to have roughly the same characteristics - ie. cover a similar range, peak in roughly the same places, tail away in a similar fashion. Thus we can 'smooth out' a frequency polygon to give a **frequency curve** which describes the basic underlying pattern, - the 'true' shape of the distribution.



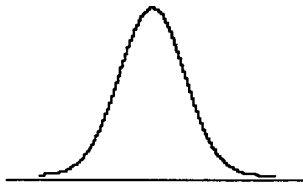
Thus:

A *frequency polygon* describes the shape we observe in a *sample*.

A *frequency curve* describes the shape we expect to find in the *population*.

## Typical Shapes

### SYMMETRIC



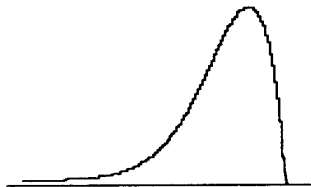
- eg. • many physical measurements (heights, weights etc.)  
• errors in measurements.

### POSITIVELY SKEW (long tail in positive direction)



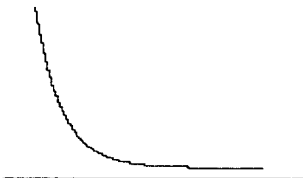
- eg. • income distributions  
• lifetimes of certain items

### NEGATIVELY SKEW



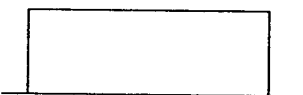
- eg. • results of an easy exam

### REVERSE J-SHAPED



- eg. • waiting times  
• times to failure

### UNIFORM



- eg. • random numbers

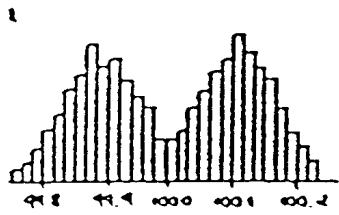
### BIMODAL



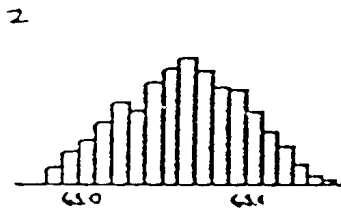
Usually denotes a mixture (eg. heights of males and females mixed in onesample). Split and analyse separately. Need big sample to detect bimodality.

**Exercise 2**

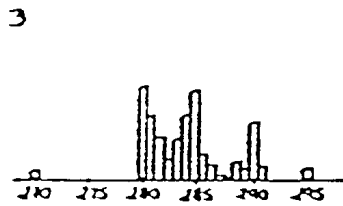
Match each of the histograms to the most likely description/explanation chosen from A-H.



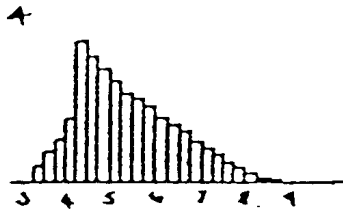
Lengths of timber beams from a sawing process (cm)



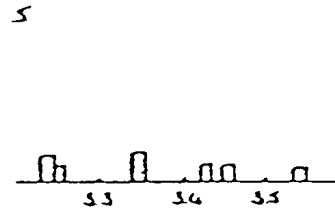
Piston diameter



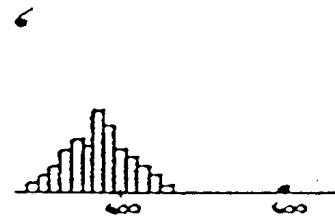
cc's in half pint bottles of Guinness



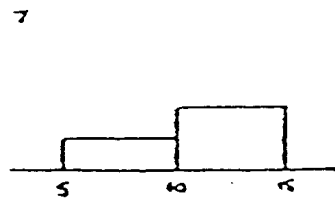
Concentrations of copper in effluent (mg/l)



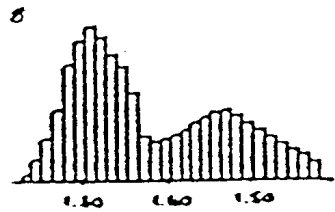
Weight of grinding wheels (kg)



Oven temperature (C)



% yield of a chemical process



Tear strengths of paper (kN)

- A Mixed production - 2 streams differing in average and variability
- B Grouping too coarse/not enough data
- C Grouping too fine/not enough data
- D Digit-preference by inspector
- E Gauge misread/copying error
- F Normal symmetric variability
- G Mixed production - 2 streams differing in average
- H Lopsided (skew) distribution of measurements

{Solution: in order, 8 7 5 3 6 2 1 4}

### 1.4.4. Frequencies and Probabilities

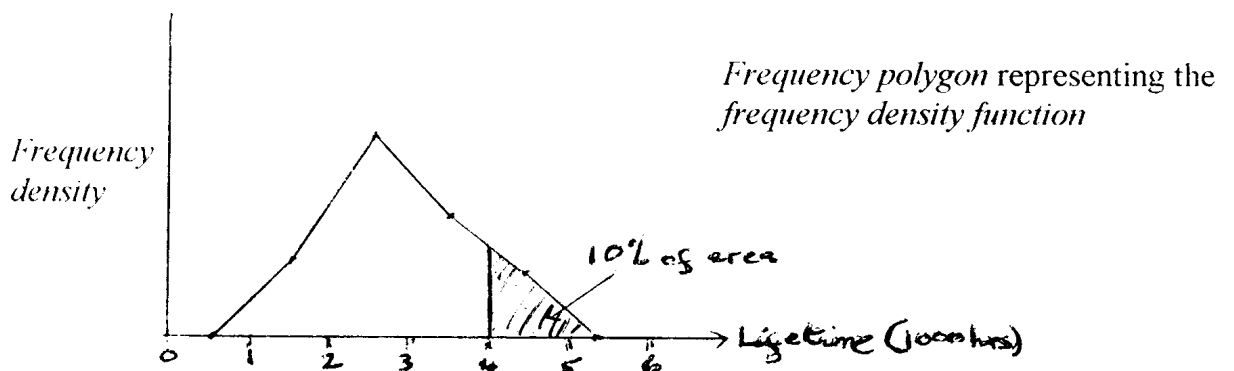
We have seen that *frequency polygons* and *histograms* describe the shape observed in a sample, while *frequency curves* describe the shape expected in the population - the true underlying shape.

On similar lines, *relative frequencies* tell us how often we have observed values in a given range (eg. 10% of a sample of batteries had lifetimes over 4000 hours). Correspondingly, *population relative frequencies* tell us how often we can expect values in a given range in the population.

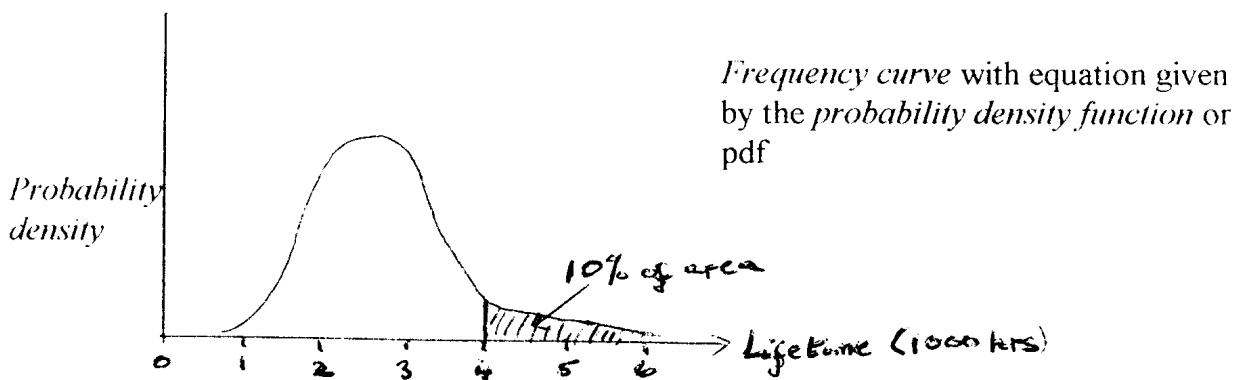
Population relative frequencies are usually referred to as **probabilities**. Then, for example, if we expect 10% of all batteries in the population to have lifetimes over 4000 hours this implies that if one battery is selected at random, it has a 10% chance of lasting more than 4000 hours, i.e. a probability of 0.1. (This probability, incidentally, is called the **reliability** at 4000 hours, considered in section 3).

Thus, *relative frequencies* in samples relate to *probabilities* in populations. Other similar correspondences which will be met in section 2 are represented in the example below.

Representation of *frequency distribution* in a sample:



Representation of *probability distribution* in a population:



### 1.4.5. Other Graphs

Other commonly used charts for representing data are:

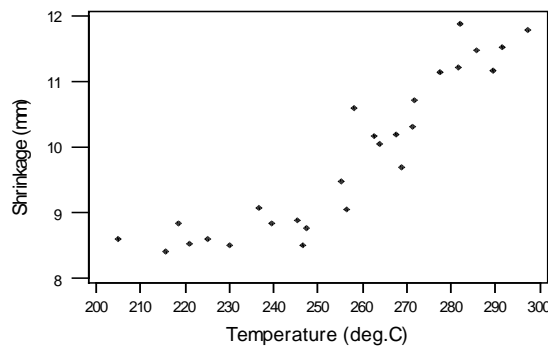
(i) Scatter Diagrams

Used to illustrate how two quantitative variables are related. The two variables are plotted on the x and y axes - we have a point (pair of co-ordinates) for each item.

Example 1

Problems have been reported in shrinkage of cable coverings around a central core. The cables are heat treated as part of the manufacturing process and the temperature at which this is done is thought to affect shrinkage (which should be minimal). A scatter diagram for a sample of cables is given below.

Effect of temperature on shrinkage for 27 cables.

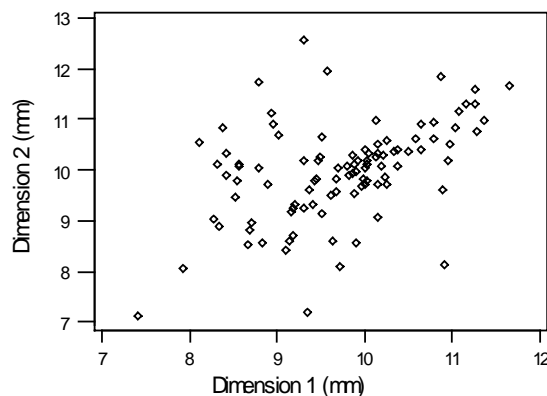


This clearly suggests that it would be unwise to treat at temperatures above 250°C.

Example 2

A component has two critical dimensions which should be related (if one dimension is relatively large then so should the other one be). A scatter plot of a sample of components suggests there is a problem here (though see section 1.6).

**Relationship between two measures dimensions**



(ii) Times Series Plot

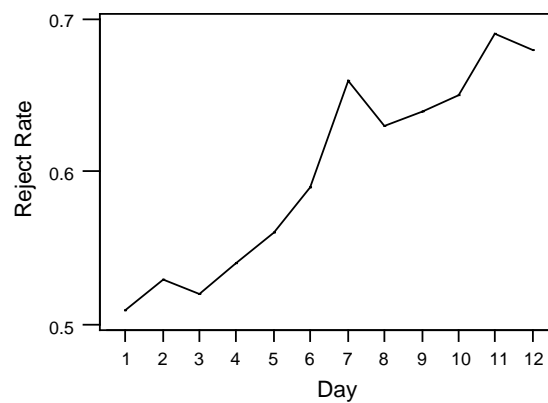
Used to show the change in a quantity over time (sometimes called a 'line graph' or 'run chart').

Example 1

The reject rates over 12 days production are recorded below. The time series plot shows a steadily rising trend over the period which would indicate that some action should be taken.

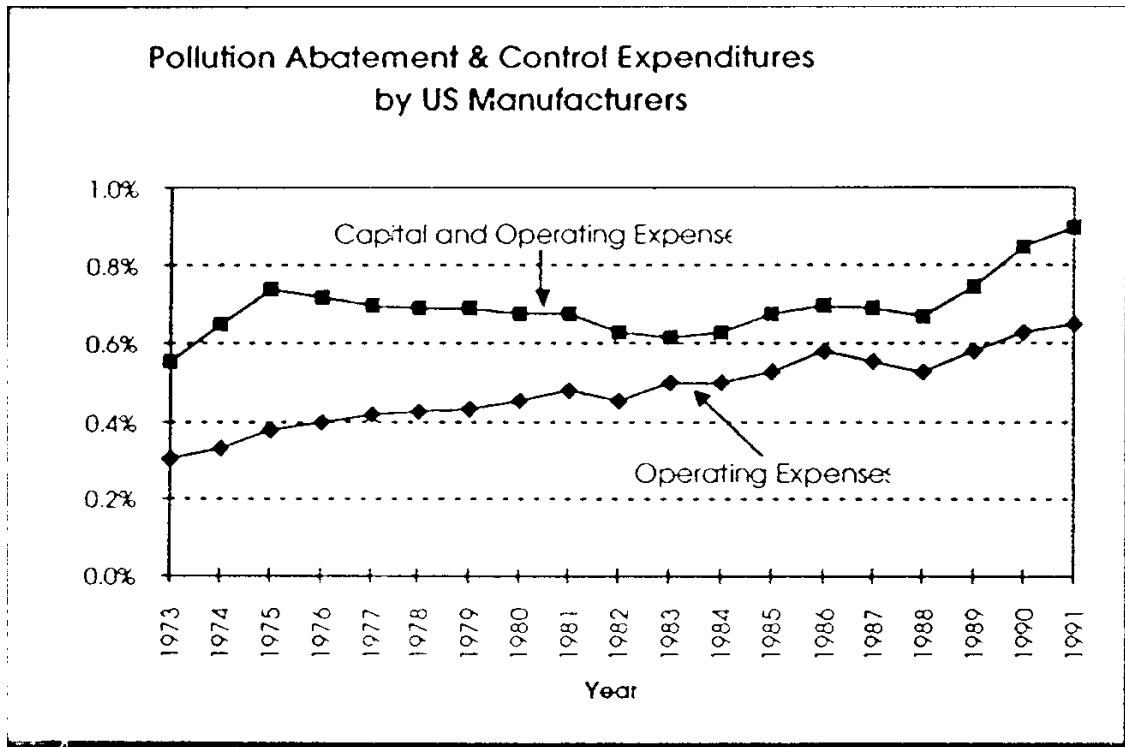
Reject Rate	Day
0.51	1
0.53	2
0.52	3
0.54	4
0.56	5
0.59	6
0.66	7
0.63	8
0.64	9
0.65	10
0.69	11
0.68	12

Time series plot to show how reject rate changes over time



**Example 2**

The following time series, taken from "Logistics Spectrum", April 1996, shows how the cost of complying with laws concerned with environmental protection have changed over the last few years.

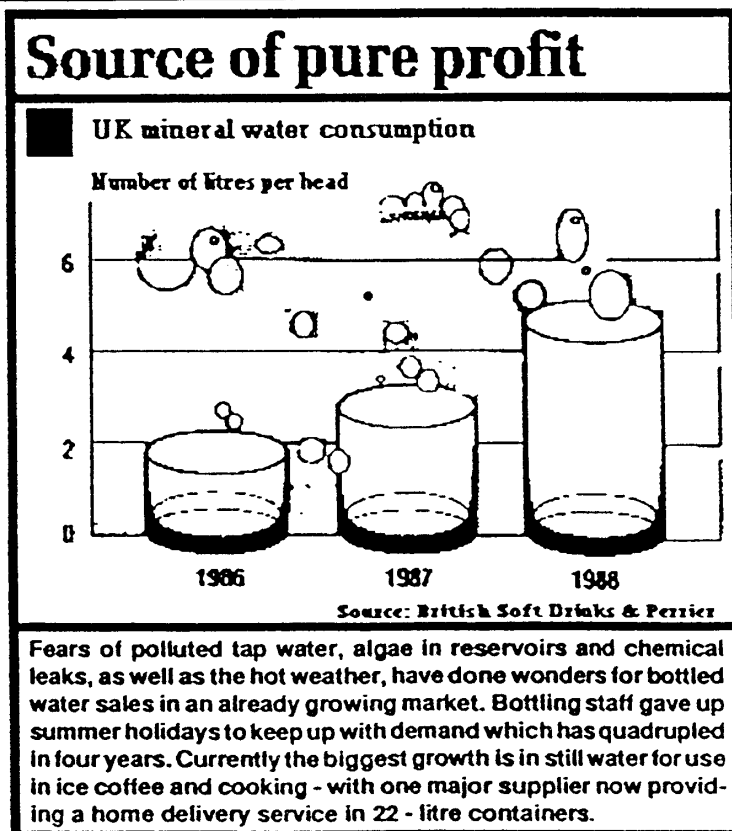
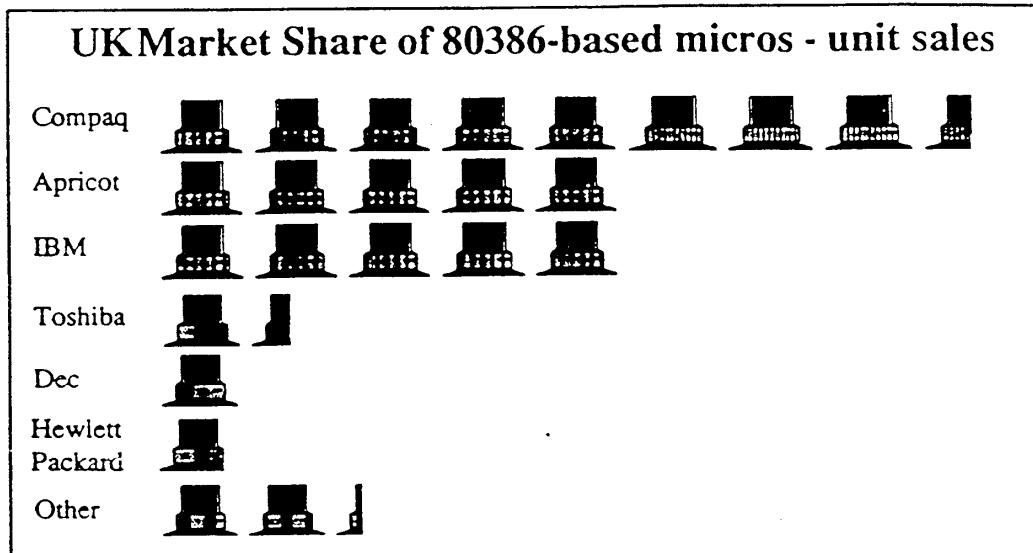


**Note**

Presenting data wrongly can be very misleading and can result in incorrect decisions being made. This is tragically illustrated in the Challenger space shuttle example given in the appendix.

(iii) Pictograms

These are similar to bar charts where the bars are replaced by symbols relevant to the quantity being charted. It is a very popular and visually stimulating way of displaying data. Typical examples are:



Pictograms are sometimes constructed by changing the size of the plotting symbol. The second example above is a good example of this type. Unfortunately, most pictograms of this type are misleading. The following is a poor example of a pictogram - the area covered by each symbol effectively represents the plotted quantity and this is clearly being mis-represented here.

**AVOID THIS TYPE OF PICTOGRAM**

