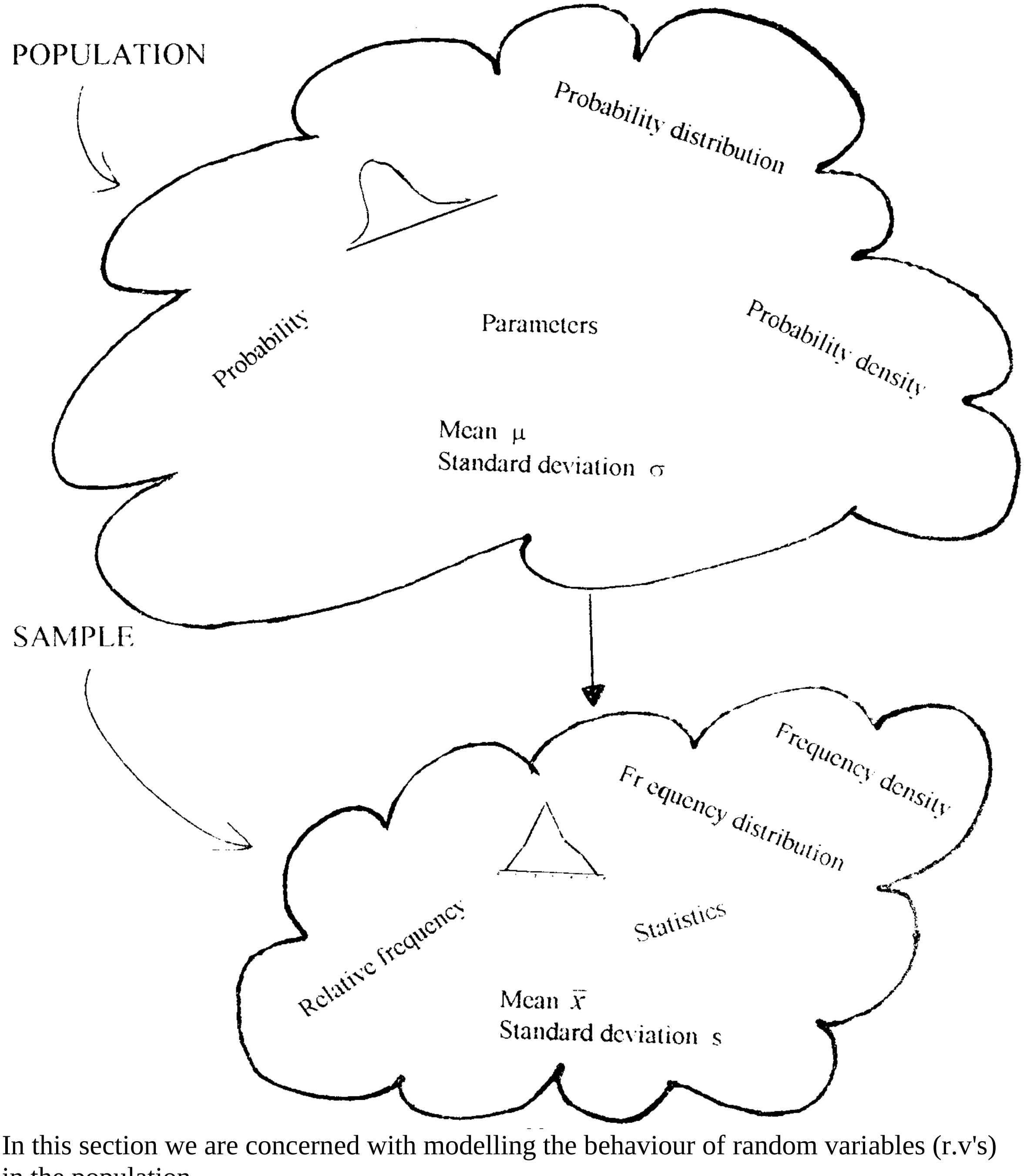
SECTION 2

PROBABILITY AND PROBABILITY DISTRIBUTIONS

2.1 INTRODUCTION

In section 1 we have highlighted the correspondences between samples and populations as summarised below.



In this section we are concerned with modelling the behaviour of random variables (r.v's) in the population.

We will consider such questions as

- what is the shape of the probability distributions?
- how can we estimate the parameters?
- what does this imply about how we expect the r.v. to behave?

This will enable us to answer such practical questions as:

- what percentage of batteries can be expected to last for more than 4000 hours?
- what is the chance that a machine will break down next month?
- what is the failure pattern of an electrical component?
- how reliable is a car engine?

First we consider how to measure and manipulate probabilities.

2.2 PROBABILITY

2.2.1 Definitions

We define **probability**, **P**, to be a measure of how likely it is that event will occur.

(i) $0 \leq P \leq 1$

(ii) $P = 0 \Rightarrow$ Event is impossible
$P = 1 \Rightarrow$ Event is certain.

There are several ways of calculating probabilities. Perhaps the simplest approach is to take what has happened in the past as an indication of what will happen in the future.

Thus, we ***use the relative frequency*** (or proportion of times an event happened in the past) ***as an estimate of probability*** (or how likely an event is to happen in the future).

Let us consider a numerical example.

Suppose that for some manufacturing process we have taken a sample of 30 items, and counted the number of faults per item, and that the resulting frequency distribution of the number of faults is:

Table 2.1

X = No. of Faults	Frequency	Relative Frequency
0	0	0
1	15	0.5
2	8	0.267
3	5	0.167
4	1	0.033
5	1	0.033
6	0	0
7	0	0
TOTAL	30	1

Then let the discrete random variable X be the *number of faults.* In theory, X can take the values 0, 1, 2, 3, …etc.

We write P(X = 3) to mean *the probability that the number of faults is 3.*

We can see that the relative frequency (or proportion of the total no. of observations) of X = 3 faults per item is 0.167.

That is, on 0.167 of the total occasions, or 16.7% of the time, 3 faults per item were observed, or

The probability of observing 3 faults per item is 0.167.

Then

$$P(X = 3) = 0.167$$

Also

$$\begin{aligned} P(X > 3) &= \text{Probability that number of faults is more than 3} \\ &= \text{Probability that number of faults is 4 or 5 or 6 or 7} \\ &= P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) \\ &= 0.033 + 0.033 + 0 + 0 \\ &= 0.066 \end{aligned}$$

Similarly, the probability that the number of faults is greater than 4 is 0.033.

2.2.2 **Probability Rules**

The last example illustrates one of the two basic rules for manipulating probabilities. These are:

(i) *Addition Rule*

If events A, B, C, . . . are mutually exclusive (i.e. cannot occur at the same time), then P(A occurs or B occurs or C occurs or . . .) is given by the sum of the individual probabilities i.e.

$$P(A \cup B \cup C \cup \ldots) = P(A \text{ or } B \text{ or } C \text{ or } \ldots) = P(A) + P(B) + P(C) + \ldots$$

(ii) *Multiplication Rule*

If events A, B, C, . . . are independent (i.e. they do not affect each other), then P(all events occur) is given by the product of the individual probabilities. i.e.

$$P(A \cap B \cap C \cap \ldots) = P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \cdot P(B) \cdot P(C) \ldots$$

Example

Suppose we select two items from the above manufacturing process. Let X_1 be the number of faults in item 1 and X_2 the number of faults in item 2.

The probability that both items have 1 fault is then

$$
\begin{aligned}
P(X_1 = 1 \text{ and } X_2 = 1) &= P(X_1 = 1) \,.\, P(X_2 = 1) \\
&= 0.5 \times 0.5 \\
&= 0.25
\end{aligned}
$$

provided item 1 having a fault does not affect whether item 2 has a fault (i.e. we have 'independent events').

2.2.3 Other Ways of Estimating Probabilities

The relative frequency approach of estimating probabilities given in 2.2.1 won't work if it's impossible to observe how often something has happened in the past. For example, what are the probabilities of:

(i) *my lottery ticket winning next Saturday*?
(ii) *Plymouth Argyle winning their next away game*?
(iii) *a new engine modification improving performance*?

Basing estimates on past data will be:

(i) *wrong (I've never won but the probability is not zero*)
(ii) *inaccurate (previous away games were under different conditions)*
(iii) *impossible (no data)*.

However, it is still possible to provide answers:

(i) *About 1 in 14 million*. This can be calculated by evaluating all possibilities and uses the **classical** definition of probability.

(ii) *0.01* In my opinion! This is a **subjective** probability.

(iii) *Unknown* but an engineer close to the project may be able to assign a subjective probability.

Subjective probabilities, such as those in (ii) and (iii), are assigned in the light of judgement and experience using any relevant information at hand. Bookmakers make their living out of them and they are the cornerstone of a relatively new branch of statistics called **Bayesian** methods. These methods are outside the scope of this course; we will concentrate on data-based estimates.

2.2.4 **<u>Conditional Probabilities</u>**

<u>Example</u>

Two machines (M_1 and M_2) are used to produce an item that has *'shaft diameter'* as a critical dimension. A total of 200 items are inspected and classified as having shaft diameters that are either too small, within specification or too large. The following table summarises the results:

	Shaft diameter			
	Too small	Within spec.	Too large	TOTALS
Machine M_1	5	40	5	50
Machine M_2	15	80	55	150
TOTALS	20	120	60	200

(a) What percentage of all items are within specification?

(b) What percentage of all items are from machine M_1?

(c) What percentage of all items are from machine M_1 and too small?

(d) What percentage of items produced by machine M_2 are too small?

(e) What percentage of items which are too large come from machine M_1?

{Answers: (a) 60% (b) 25% (c) 2.5% (d) 10% (e) 8.5%}

We could alternatively have presented these results as proportions of the overall total of 200 as in the following table.

	Shaft diameter			
	Too small (S)	Within spec. (W)	Too large (L)	TOTALS
Machine M_1	0.025	0.200	0.025	0.250
Machine M_2	0.075	0.400	0.275	0.750
TOTALS	0.100	0.600	0.300	1.000

Then example (d) above $\left(=\frac{15}{150}=0.100\right)$ could be calculated as $\frac{0.075}{0.750}=0.100$.

Equivalently, these proportions can be considered as ***probabilities*** associated with various outcomes:

	Shaft diameter			
	S	W	L	
Machine M_1	$p(M_1 \cap S)$	$p(M_1 \cap W)$	$p(M_1 \cap L)$	$p(M_1)$
Machine M_2	$p(M_2 \cap S)$	$p(M_2 \cap W)$	$p(M_2 \cap L)$	$p(M_2)$
	$p(S)$	$p(W)$	$p(L)$	1.000

Then example (e) $\left(= \frac{5}{60} = 0.083\right)$ could be calculated as $\frac{0.025}{0.300} = 0.083$.

These two are examples of **conditional probabilities**.

Example (d) is p(S given M_2) written $p(S|M_2)$.

Example (e) is p(M_1 given L) written $p(M_1|L)$.

In symbols, the two calculations above are

in (d) $$p(S|M_2) = \frac{0.075}{0.750} = \frac{p(M_2 \cap S)}{p(M_2)}$$

in (e) $$p(M_1|L) = \frac{0.025}{0.300} = \frac{p(M_1 \cap L)}{p(L)}$$

In general, for two events A and B, we define the **conditional probabilities** as:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

and $$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

Notes

(i) Rearranging these gives, for example:

$$p(A \cap B) = p(A \mid B) \times p(B)$$

This is a more general form of the multiplication rule for two events.

(ii) If two events are **independent**,

$$p(A \mid B) = p(A)$$

and $$p(B \mid A) = p(B)$$

i.e. the condition is irrelevant. The expression in (i) then reduces to the special rule for independent events suggested in section 2.2.2.

(iii) The *addition rule* for mutually exclusive events given in section 2.2.2 can also be generalised to any events, mutually exclusive or otherwise. For any 2 events A and B the general addition rule is:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

(iv) **Venn diagrams** are useful in handling problems of this type. Similarly, **tree diagrams** are useful displays for working with probabilities when events occur at different times or stages, involving conditional probabilities. The second exercise below illustrates this.

Exercise 1 *(Independence)*

Refer to the example at the beginning of this section. Is shaft diameter independent of the machine it is made on? Justify and interpret your answer.

Solution

There are several ways of checking for this. For example:

We have calculated in (e) $P(M_1 \mid L) = 0.083$ and in (b) $P(M_1) = 0.25$.
If M_1 and L are independent, these two probabilities should be the same from note (ii) above. They are *not*, so shaft diameter is not independent of machine.

If events A and B are independent, $P(A \cap B) = P(A) \times P(B)$. From the table on the bottom of page 58, it can be seen that this is not always so. For example,

$P(M_1 \cap W) = 0.200$ while $P(M_1) \times P(W) = 0.250 \times 0.600 = 0.150$.

So the events are not independent. JUST ONE FALSE RESULT IS ENOUGH.

Exercise 2 *(Tree Diagrams)*

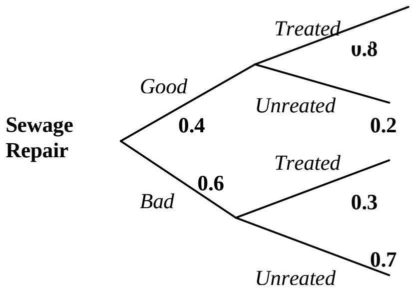
In a national survey of sewage and effluent treatment, the state of the sewage pipe system was classified either as in *good* or *not in good repair*, and the level of sewage treatment was categorised either as fully or not treated. It was found that 40% of places had sewage pipe systems in a state of good repair and of these 80% had fully treated sewage. Of the places where the sewage pipe system was not in a state of good repair only 30% of these had fully treated sewage. Display the information using a tree diagram. Find the probability of the event that:

(i) sewers are in good repair and have fully treated sewage;
(ii) sewers are in good repair but with untreated sewage;
(ii) sewers are not in good repair but have fully treated sewage;
(iii) sewers have fully treated sewage.

Is the level of sewage treatment independent of state of repair?

Solution
It is helpful to draw a tree diagram of the situation:

Overall (joint) Probabilities



$P(G \cap T) = P(G) \times P(T|G) = 0.4 \times 0.8 = \mathbf{0.32}$

$P(G \cap T') = P(G) \times P(T'|G) = 0.4 \times 0.2 = \mathbf{0.08}$

$P(G' \cap T) = P(G') \times P(T|G') = 0.6 \times 0.3 = \mathbf{0.18}$

$P(G' \cap T') = P(G') \times P(T'|G') = 0.6 \times 0.7 = \mathbf{0.42}$

(Notation: T' means *not T* and G' means *not G.)*

Notes (i) Probabilities on each branch sum to 1.
(ii) Probabilities on later branches are all conditional probabilities.
(iii) Overall, the joint probabilities sum to 1.

The answers to the above questions are then:

(i) 0.32 (ii) 0.08 (iii) 0.18 (iv) 0.32 + 0.18 = 0.50

Finally, the level of sewage treatment is NOT independent of state of repair. Note, for example that the probability of full treatment is 0.8 if in good repair but only 0.3 if in bad repair.

2.3 PROBABILITY DISTRIBUTIONS

Consider, for example the two random variables:

X : number of breakdowns in a month
and Y : lifetime of a battery

X is a **discrete** r.v. that can take values 0, 1, 2, 3, ...

Y is a **continuous** r.v. that can take values in the range 0 $\rightarrow$ 100000 hours (say)

However, not all values are equally likely and we now consider how the probabilities are distributed across the range of possible values. Discrete and continuous probability distributions are treated separately.

2.3.1 Discrete Distributions

Recall that discrete data is numerical data that can only take specific values.

e.g. Shoe size
No. of brothers and sisters.
No. of accidents

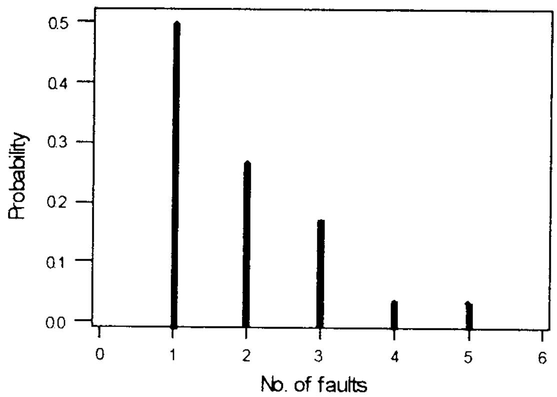
Then the "No. of faults" data shown in Table 2.1 is discrete data, and the estimated probability distribution is

No. of Faults, X	Probability
1	0.5
2	0.267
3	0.167
4	0.033
5	0.033
TOTAL	1.000

Note that the total probability is 1 since X must take some value within the range. In general, for any discrete probability distribution with probabilities
$p_1, p_2, p_3, \ldots$

$$\sum p_i = 1$$

This distribution can be shown graphically by a bar chart. The height of each bar then represents the probability of that particular value of the random variable occurring.



The ‘average’ or *mean* of the distribution is also known as the **expected value** and is defined by:

$$\mu = E(X) = \sum xp_x$$

Thus in the above, $\mu = (1\times 0.5) + (2\times 0.267) + (3\times 0.167) + (4\times 0.033) + (5\times 0.033) =$ **1.832**
This means that, on average, there are 1.832 faults per item.

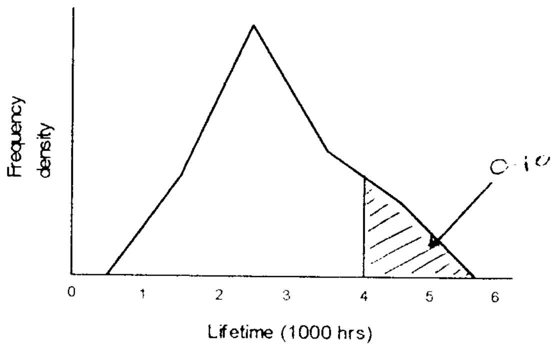
2.3.2 Continuous Distributions

Continuous data can take any value within a given range.

e.g. Time spent studying for an exam.
Person's height or weight.
Lifetimes

Such data can be represented graphically by histograms or frequency polygons.

For example the frequency polygon of component lifetimes in section 1.4.4 was:



The estimated probability of a lifetime being >4000 hours is given by the area of the polygon as shaded above.

We have said that if this histogram (or associated frequency polygon) is smoothed out to a curve then this describes the 'true' shape of the distribution - i.e. it is a graphical representation of the probability distribution of X.

For continuous data, the probability that the random variable(e.g. time spent to perform some task) takes a particular value (e.g. 2 hours) is meaningless since we need to know what is meant by 2 hours. (Does it mean 2 hr ± 30 min? 2 hr ± 5 min? 2 hr ± 1 sec?)

i.e. We cannot consider

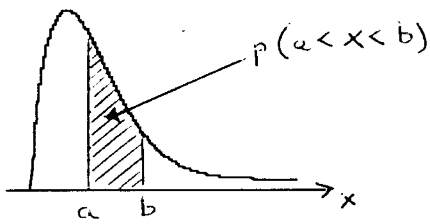
$$P(X = a)$$

for a continuous random variable.

Instead however, we calculate the probability that X lies within some interval.

i.e. $P(a < X < b)$

By analogy to the fact that the area under the histogram represents frequency, the area under the curve between 2 values represents the probability that X lies between the 2 values.



i.e. If X = time spent on some task

$P(a < X < b)$ = Probability time spent is between a and b hours

= area shown shaded.

The unshaded area to the left is $P(X < a)$ while the unshaded area to the right is $P(X > b)$.

Note: Since total probability = 1,

(i) **area under curve = 1**

(ii) **$P(X = a) = 0$**

($P(X = a)$ cannot be defined for continuous data.)

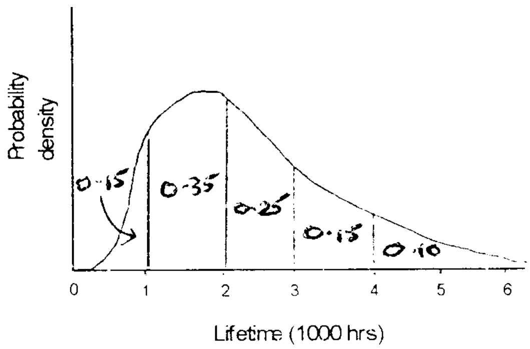
The equation of this curve is given by the **probability density function** (or **pdf)**.

Thus, to summarise:

1.	**The probability distribution of a continuous random variable is usually specified in terms of its probability density function or p.d.f.**
2.	**<u>Probabilities</u> are then given by areas under the curve of the probability density function.**

Example

The sketch below shows the p.d.f of the 'lifetime' distribution of an electronic component.



(The numbers given in sections of the curve represent area).

Find

(i) the probability that a randomly selected component will fail before 1000 hours.

$P(X \leq 1000) = 0.15.$

(This could be written as P(X < 1000) since P(X = 1000) = 0.)

(ii) the probability that it will still be working at 1000 hours.

$P(X \geq 1000) = 1 - 0.15 = 0.85$

(iii) the proportion of components with the above lifetime distribution which will have failure times between 1000 and 2000 hours.

$P(1000 \leq X \leq 2000) = 0.35.$

(iv) the probability that two randomly selected components will both fail before 1000 hours.

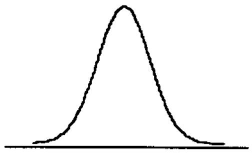
$P(X_1 \leq 1000) \times P(X_2 \leq 1000) = 0.15 \times 0.15 = 0.0225$

(Q. If what? A. *Provided components are failing **independently**.)*

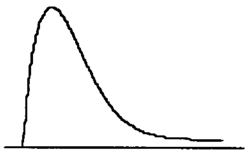
2.3.3 Common Shapes and Specific Distributions

The most common shapes of probability distributions appropriate for lifetimes or failure times are:

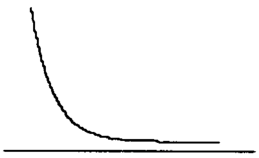
(i) SYMMETRIC



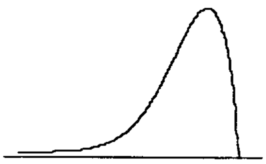
(ii) POSITIVELY SKEW



(iii) REVERSE J-SHAPED



(iv) NEGATIVELY SKEW



In this course we consider the four specific probability distributions which can be used to model data with these shapes. These are:

- *Normal distribution* (shape (i))
- *Lognormal distribution* (shape (ii))
- *Exponential distribution* (shape (iii))
- *Weibull distribution* (flexible - all 4 shapes)

Given some data on the continuous r.v. X, the fundamental questions are:

Question 1 *Which distribution generally fits the data best*?

Question 2 *Which distribution exactly (as specified by it's parameters)*?

Question 3 *What are the implications*?

These implications include evaluating probabilities and interpreting estimates of the parameters.

Question 1 is addressed in section 2.4 and questions 2 and 3 are covered in sections 2.5-2.8 which deal with the four distributions listed above.

2.4 FITTING DISTRIBUTIONS TO DATA

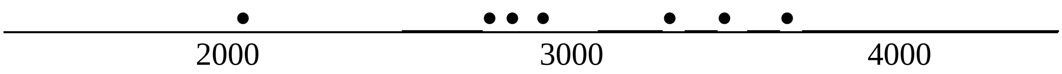
2.4.1 Probability Plots

A first step could be to plot a histogram of the data to visually assess the shape (e.g. is it symmetrical?) but to do this requires a reasonable amount of data which is not always available. An alternative approach which we adopt is to use **probability plots**.

The basic idea uses *quantiles* as introduced in section 1. ie. values which divide a set of data or a distribution into equal parts. (For this reason, these plots are sometimes called quantile-quantile or QQ plots.)

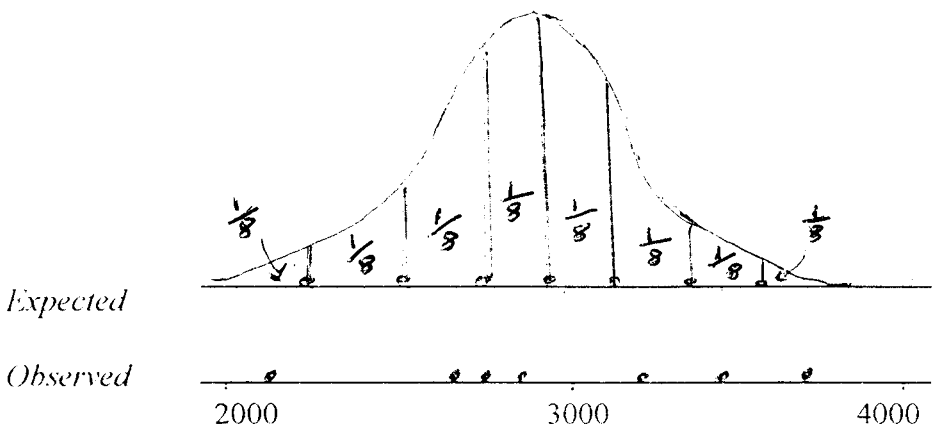
Say for example, we had the lifetimes of 7 components ranging from about 2000 to 4000 hours as follows:

2080 2730 2750 2850 3230 3400 3640



and wanted to know if the normal distribution was a reasonable model on this range.

To answer this, we ask the further question - if lifetimes really did have a normal distribution (in the population) where would we expect these 7 observations to fall? These can be found by dividing the area under the normal curve into equal slices. With 7 values, there will be 8 areas each of which should be 0.125 $\left(\frac{1}{8}\right)$.

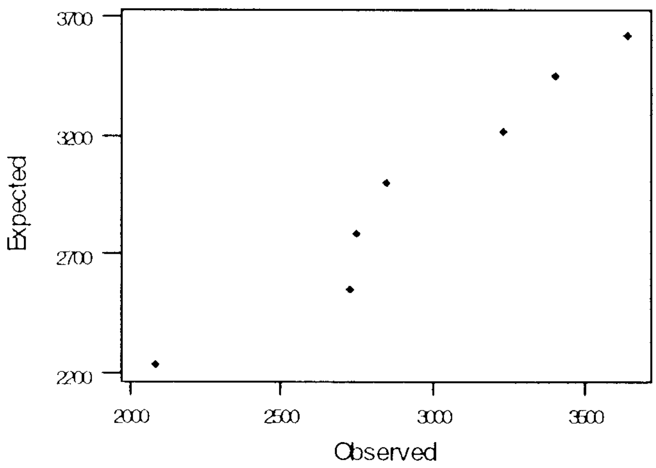


The general principle is then to compare what we have observed in the sample with what we would expect to happen if the normal distribution was used as the model for the population.

In fact, in this example, it can be shown (see section 2.5) that these expected values are:

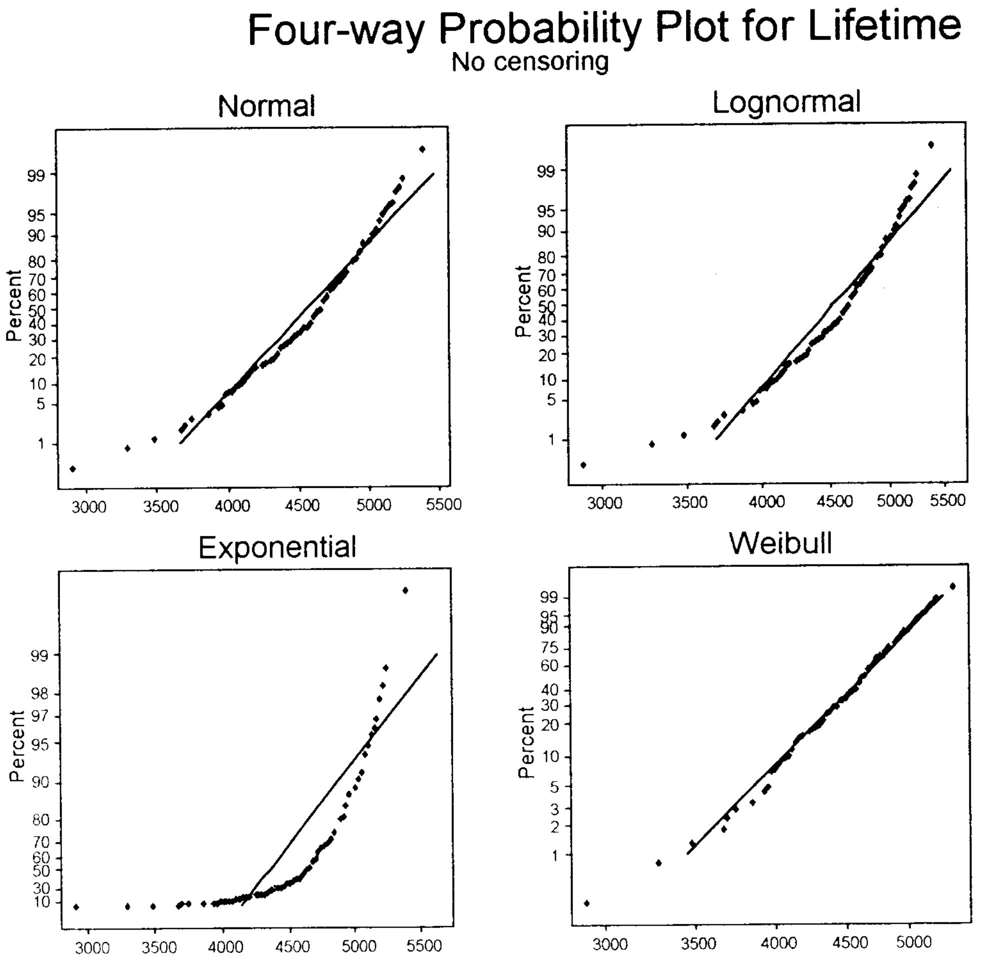
	Observed in sample	*Expected* in a normal population
1	2080	2234
2	2730	2551
3	2750	2788
4	2850	3000
5	3230	3212
6	3400	3449
7	3640	3622

If these are plotted against each other we should get more or less a straight line

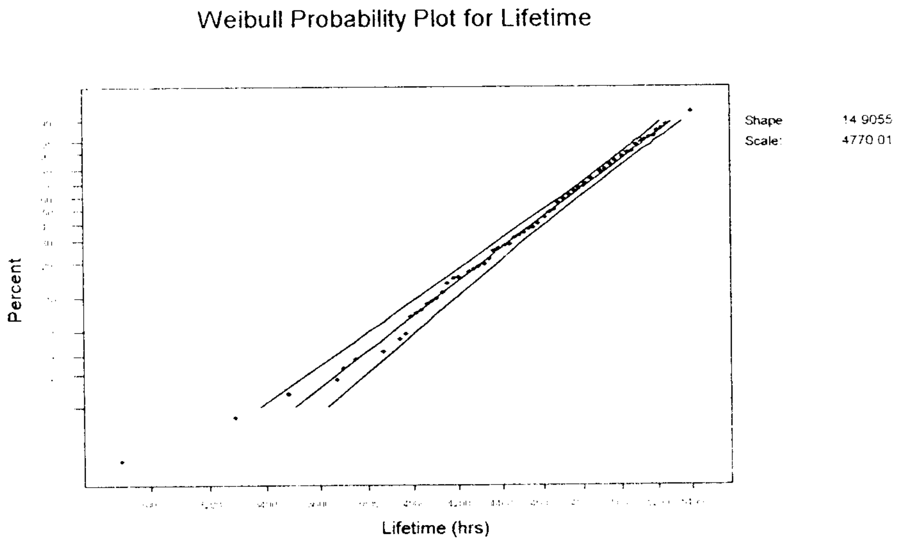


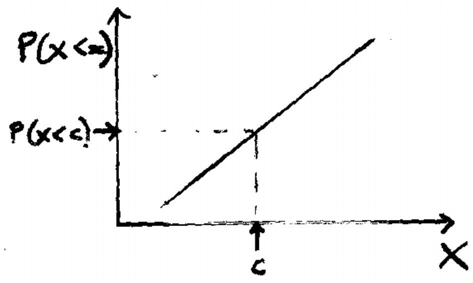
In this case, a normal distribution seems to be a reasonable fit. If it weren't, we could ask what would be expected in a lognormal or exponential or Weibull population.

A useful routine in Minitab is the ID plotting function which produces all four probability plots (on specially scaled axes). The plots on the following page, for example, suggest that the Weibull distribution is the most reasonable model for the lifetime data used. In general, look for the line that fits the data best; especially in the tails i.e. at each end of the data. The specific Weibull probability plot, which is also given includes 'error' curves around the fitted line within which most of the points should lie – especially in the tails. These are 95% confidence limits, which will be covered in section 4.


Four-way Probability Plot for Lifetime
No censoring
Normal
Lognormal
Exponential
Weibull
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
98
97
75
3
2
3000
3500
4000
4500
5000
5500

Best fit


Weibull Probability Plot for Lifetime
Percent
Lifetime (hrs)
Shape 14 9055
Scale: 4770 01



Note that:

$$p(X<c)=\begin{cases}\Pr ob.\ \ lifetime < c \\ \Pr ob.\ \ fails\ \ before\ \ time\ \ c\end{cases}$$

$$p(X>c)=1-p(X<c)=\begin{cases}\Pr ob\ \ lifetime > c \\ \Pr ob\ \ works\ \ at\ \ time\ \ c \\ \Pr ob\ \ survives\ \ to\ \ time\ \ c \\ \text{Re}\,liability\ \ at\ \ time\ \ c\end{cases}$$

2.4.2 Estimating Probabilities

The best-fitting line can be used to estimate probabilities. For example, the Weibull line in the ID plots tells us that:

(i) ____% of all components are expected to have lifetimes less than 4000 hours.

(ii) p(lifetime > 5000 hours) = 1 – ____ = ____

Or they can be used the other way round. For example:

(iii) 10% of components in the population are expected to last for more than ____ hours.

and

(iv) 50% of components will have lifetimes less than ____ hours.
(This is the half-way point - the *median* lifetime.)

More accurate estimates of these points and probabilities can be obtained mathematically from the pdf's as described in sections 2.5-2.8.

{Answers: (i)7% (ii) 1 – 0.85 = 0.15 (iii) 5050 hours – the 90th percentile (iv) 6650 hrs}

2.4.3 Testing Goodness-of-Fit

These probability plots are just a visual aid for assessing whether a particular distribution is a sensible model. It would be better to have a more formal procedure for deciding which distribution is 'best'. There are a number of **significance tests** available for this - most importantly, the Kolmogorov-Smirnoff test and the χ^2 (chi-squared) test. For now we will stick with an informal graphical assessment.