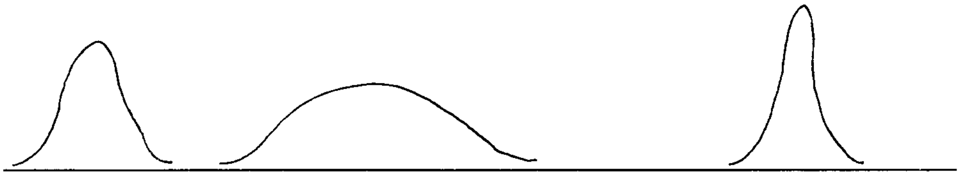
2.5 THE NORMAL DISTRIBUTION

2.5.1 Parameters and Functions



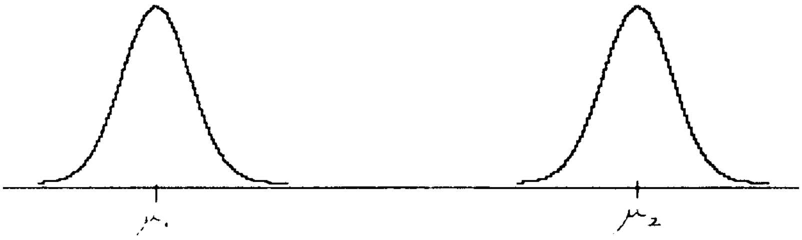
The normal distribution has a symmetric ball shaped curve. Its exact shape and size are determined by its mean (μ) and standard deviation (σ)

μ determines where the centre is located
σ determines the amount of spread about the centre.

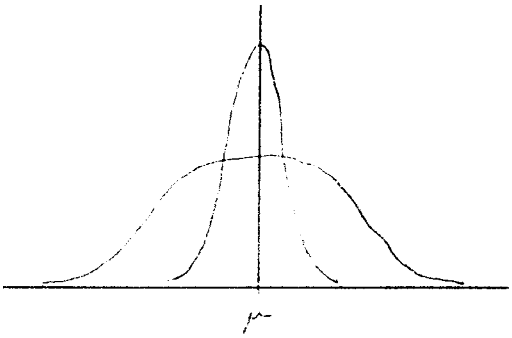
These parameters would usually be unknown and would be estimated by the sample mean $(\overline{x})$ and sample standard deviation (s).

Example

(i) These distributions have the same spread (σ) but different means (μ_1 and μ_2)



(ii) These distributions have the same mean (μ) but different spreads (σ_1 and σ_2)

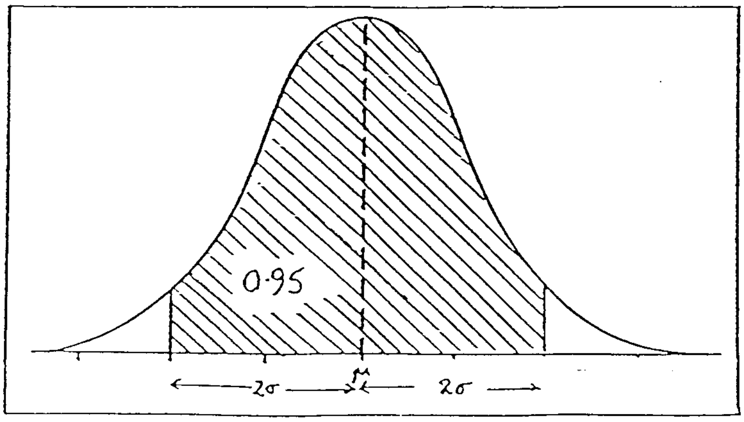


Properties of the Normal Distribution:

(i) *Symmetric about μ.*

(ii) *Bell shaped.*

(iii) *Approximately 68% of the area lies within one standard deviation of the mean, and 95% of the area lies within two standard deviations of the mean.*



(iv) *99.7% (i.e.* nearly all*) of the area lies within three standard deviations of the mean.*

(The last property is the basis for the Shewhart control charts used in Statistical Process Control or SPC. Similarly, the fact that 6σ effectively covers the whole range is the basis for the commonly used capability indexes C_p and C_{pk}.)

Example

IQ's in the population have a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$.

Then 95% of the population have IQ's in what range?

Also, it is very rare for an individual to have an IQ outside what values?

Solution

We have

$$\mu = 100$$

$$\sigma = 15$$

Then 95% of the population lie within ± 2 standard deviations of the mean, that is, within the limits

$$(\mu - 2\sigma, \quad \mu + 2\sigma)$$

or $\left(100-2(15),\quad 100+2(15)\right)$

i.e. $\left(70,\quad 130\right)$

It is rare for an individual observation to be more than 3 standard deviations from the mean, that is, outside

$$\left(100-3\sigma,\quad 100+3\sigma\right)$$

or $\left(100-3(15),\quad 100+3(15)\right)$

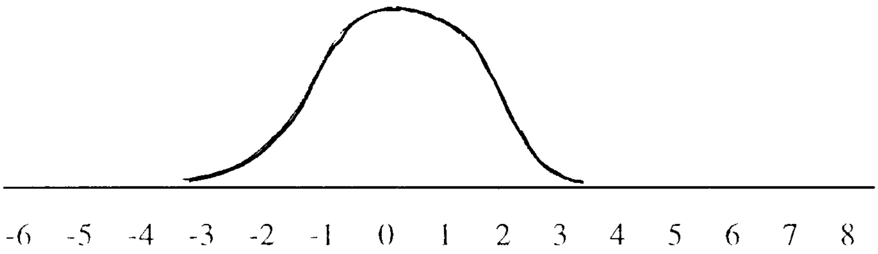
i.e. less than 55 or greater than 145.

Example

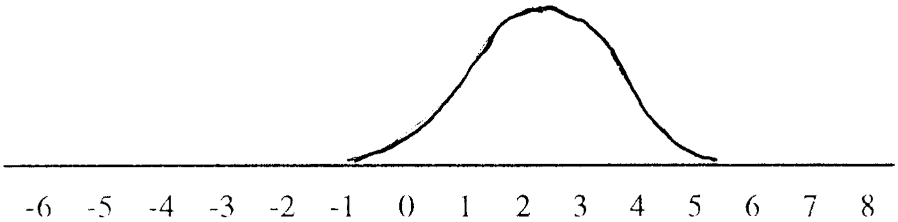
Sketch the following distributions:

(i) Normal distribution: $\mu = 0$, $\sigma = 1$

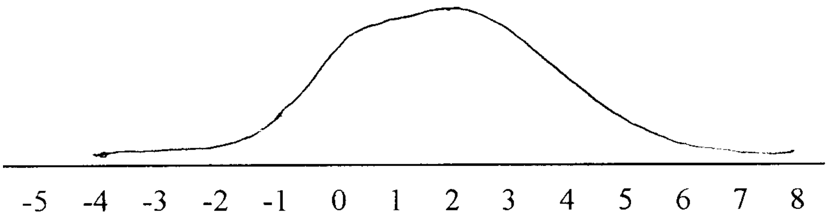
- the so-called "**standardised normal**" or N(0,1) Distribution



(ii) Normal distribution: $\mu = 2$, $\sigma = 1$



(iii) Normal distribution: $\mu = 2$, $\sigma = 2$



Probability Density Function

The equation of the normal curve is given by the pdf:

$$f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\left(x-\mu\right)^2 / 2\sigma^2\right\}$$

Finding areas under this curve to give probabilities requires integrating this function. This can't be done analytically so numerical methods are used. Results have been tabulated (see section 2.5.3).

2.5.2 Mechanism *(How does it arise?)*

The normal distribution arises naturally in many situations because of the following property of sums of random variables:

If a variable is the sum* *of other (not necessarily normally distributed) random components then it will, under fairly general conditions, be normally distributed. Thus, for example, a dimension resulting from a number of machining operations (each with its own error) may be viewed as a sum of random components and might be expected to have a normal distribution.*

(* Since the mean $\overline{X} = \sum X/n$, $\overline{X}$ will also have approximately a normal distribution (if n is 'large').

This is the **Central Limit Theorem**. It places the normal distribution in a very important position in statistical inference.

2.5.3 Applications

Finding Probabilities

We have seen how probabilities can be estimated from a probability plot. More accurate calculations require the use of tables.

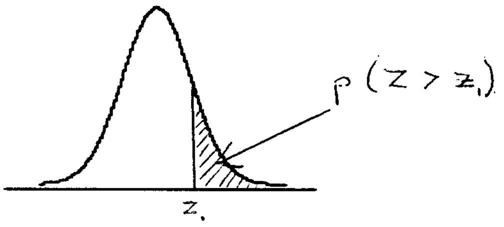
Random Variables with a Standardised Normal or N (0,1) Distribution

Probabilities associated with a Normal Distribution with mean = 0 and standard deviation = 1 (the so-called **standardised Normal Distribution** or N(0,1) distribution) are tabulated (see *Tables* at back of book).

It is actually more usual to define a normal distribution in terms of the variance (σ^2) rather than the standard deviation (σ). In this case, they are the same.

Thus for any random variable, Z, which is normally distributed with a mean of 0 and a variance of 1, we can find $P(Z > Z_1)$ by looking up Z_1 in these tables.

(i.e. The probability that Z takes a value greater than some ***known*** *value* z_1 *)*



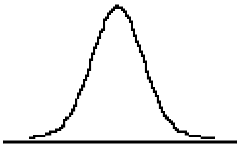
$P(Z > z_1)$ is tabulated for $z_1 \geq 0$. For other areas (i.e. probabilities) use the following properties:

- *the distribution is symmetrical*
- *the total area under the curve is 1*

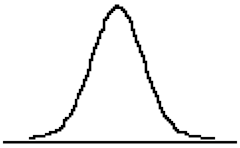
Examples

Find the following:

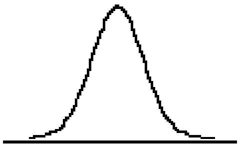
(i) $P(Z > 0.35) =$



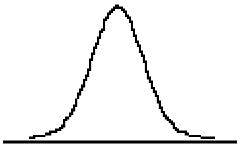
(ii) $P(Z < 1) =$



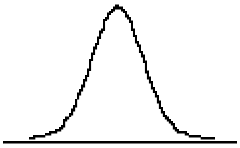
(iii) $P(Z < -1.25) =$



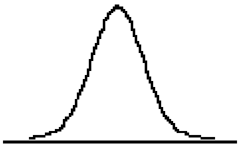
(iv) $P(1 < Z < 2) =$



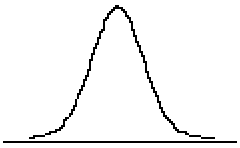
(v) $P(Z > -1.5) =$



(vi) $P(-1.81 < z < -0.23) =$



(vii) $P(-1.5 < Z < +0.95) =$



{Answers: (i) ***0.3632*** *(ii) 1-0.1587=****0.8413*** *(iii)* ***0.1056*** *(iv) 0.1587-0.02275=****0.13595***
*(v) 1-0.0668=****0.9332*** *(vi) 0.4090-0.0351=****0.3739*** *(v11) 1-(0.0668+0.1711)=****0.7621****}*

Random Variables with a N(μ,σ^2) distribution.

Probabilities associated with random variables, X, which are Normally distributed, with mean μ and variance σ^2 [i.e. X ~ N(μ,σ^2)] are obtained by

(i) ***Translating the random variable from X ~ N(μ,σ^2) to Z ~ N(0,1), using the formula***

$$Z = \frac{X-\mu}{\sigma} \qquad \text{i.e.} \qquad \frac{X - Mean}{SD}$$

Then P(X > X_1) can be rewritten as P(Z > Z_1)

(ii) ***Referring to tables of the standardised normal [N(0,1)] distribution, we can then find P(Z>Z_1)***

Z is called the **standardised variable**. It's value tells us how many standard deviations away from the mean a particular value of X is. For example an IQ of 130 from a distribution with μ = 100 and σ = 15 translates to a z value of +2 - ie. 130 is 2 standard deviations above the mean.

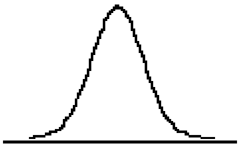
The method is illustrated by the following examples.

<u>Examples</u>

(a) Noise voltages on a sliding contact are normally distributed with mean 12 micro volts and standard deviation 1.5 micro volts. On what proportion of occasions are noise voltages:

(i) greater than 15 micro volts;

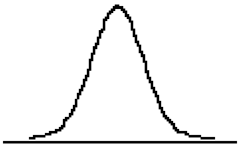
X=15 standardises to $Z = \frac{15-12}{1.5} = +2$



Then $P(X > 15) = P(Z > 2) = \mathbf{0.02275}$

(ii) less than 14.2 micro volts;

X=14.2 standardises to $Z = \frac{14.2-12}{1.5} = 1.47$

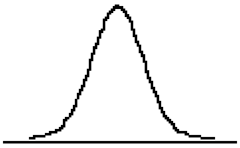


Then $P(X < 14.2) = P(Z < 1.47) = 1 - 0.0708 = \mathbf{0.9292}$

(iii) between 11 and 14 micro volts?

X = 11 standardises to $Z = \frac{11-12}{1.5} = -0.67$

X = 14 standardises to $Z = \frac{14-12}{1.5} = 1.33$



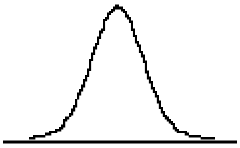
$$\begin{aligned} P(11 < X < 14) &= P(-0.67 < Z < 1.33) \\ &= 1-(0.2514+0.0918) \\ &= 1-0.3432 \\ &= 0.6568 \end{aligned}$$

(b) An electronic system is designed to have a mean response time of 1.5 secs. The times are found to be normally distributed about this mean with a standard deviation of 0.05 secs. What is the probability that on a particular activation of the system, response time is:

(i) greater than 1.4 secs;

$X = 1.4 \rightarrow Z = \frac{1.4-1.5}{0.05} = -2$

Then



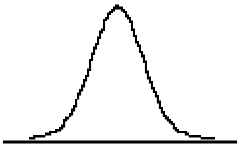
$$\begin{aligned} P(X > 1.4) &= P(Z > -2) \\ &= 1-0.02275 \\ &= 0.97725 \end{aligned}$$

(ii) within 1/10 sec of the mean?

We require $P(1.4 < X < 1.6)$

$$\begin{aligned} &= P(-2 < Z < 2) \\ &= 1-(0.02275+0.02275) \\ &= 0.9545 \end{aligned}$$

(In other words, approximately 95% are within 2 standard deviations of the mean.)



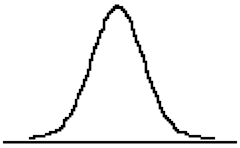
Finding Percentage Points

Values corresponding to given probabilities can also be found from tables as illustrated in the following examples.

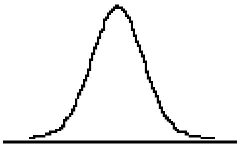
Example

(a) Find the values from the standard normal distribution corresponding to:

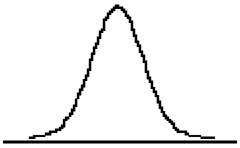
(i) an upper tail area of 0.05



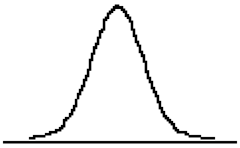
(ii) a lower tail area of 0.10



(iii) an upper tail area of 0.8



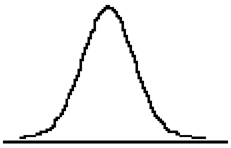
(iv) a lower tail area of 0.99



{Answers: (i) +1.6449 – the 95th percentile (ii) –1.2816 – the 10th percentile (iii) –0.8416 – the 20th percentile (iv) +2.3263 – the 99th percentile}

(b) Consider the system with mean response time of 1.5 sec and standard deviation of 0.05 sec in Example (b) on the previous page. 99% of the time, the system will have a response time below what?

Z = +2.3263. Then the required value of X (response time) is



$$\begin{aligned} & mean + (2.3263)\times(SD) \\ &= 1.5 + (2.3263)(0.05) \\ &= 1.5 + 0.12 \\ &= 1.62\,\text{sec} \end{aligned}$$

2.6 THE LOGNORMAL DISTRIBUTION

2.6.1 Parameters and Functions

A random variable X has a lognormal distribution if log(X) has a normal distribution. The pdf is given by

$$\boxed{f(x) = \left(2\pi\sigma^2 x^2\right)^{-\frac{1}{2}} \exp\left\{-\left(\log x - \mu\right)^2 / \left(2\sigma^2\right)\right\}}$$

for positive values of x. (Logs of negative numbers do not exist.)

The parameters μ and σ^2 are the mean and variance of the underlying normal distribution. As before, they measure the location and spread of the distribution. (Minitab refers to them as location and scale in the lognormal screens).

In effect σ^2 relates to the amount of skewness in the distribution of X. The larger it is, the more skew the distribution.

2.6.2 Mechanism *(How does it arise?)*

We have seen that a normal distribution will result if a variable is the SUM of random components that can themselves have any distribution.

Similarly, a lognormal distribution will result if a variable is the PRODUCT of random components that can themselves have any distribution.

In other words, if a random variable X is the result of a combination of many effects then if those effects combine in an additive fashion, a normal distribution results; if in a multiplicative fashion, a lognormal distribution results.

2.6.3 Applications

Calculating probabilities and percentage points for the lognormal distribution involves working with the normal distribution as in section 2.5.3 and transforming to the original lognormal scale as appropriate. This is illustrated in the following example.

Example

A probability plot suggests that the lifetime (X in hours) of a component has a lognormal distribution with the mean and standard deviation of $\log_e(X)$ being 10 and 0.5 respectively. (Logs to base e are used here but logs to base 10 can be used if preferred).

(i) What proportion of components have lifetimes over 50,000 hours?

$$\log_e(50000) = 10.8198$$

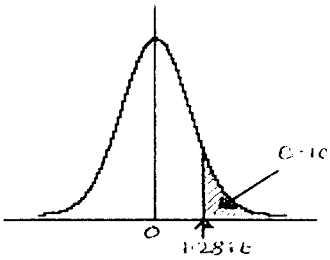
Standardizing gives $Z = \dfrac{10.8198 - 10}{0.5} = 1.64$

Then from normal tables, $p(z > 1.64) = 0.0505$

Thus approximately 5% of components will last for more than 50,000 hours.

(ii) What lifetime is exceeded by 10% of components?

From normal tables, the required standardized value is $z = 1.2816$



Then $\log_e(X)$ should be 1.2816 standard deviations above the mean

i.e. $\log_e(X) = 10 + (1.2816)\,(0.5)$

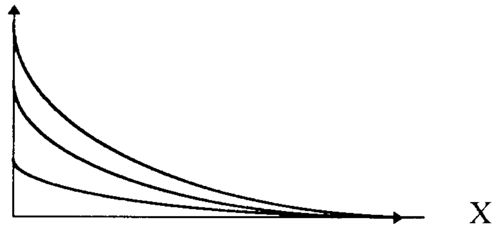
$= 10.6408$

Then $X = e^{10.6408} = 41{,}806$

So 10% of components will have lifetimes greater than 41,806 hours.

2.7 THE EXPONENTIAL DISTRIBUTION

2.7.1 Parameters and Functions



These reverse J-shaped distributions differ only in where they start on the vertical axis (λ). The probability densities then decay exponentially. The equation of the pdf is:

$$\boxed{f(x) = \lambda e^{-\lambda x}} \quad \text{for} \quad x > 0$$

Unlike the normal and lognormal distributions that have two parameters, the exponential distribution has only one parameter λ.

2.7.2 Mechanism - the Poisson Process

In many situations we are interested in the number of occurrences of some random event in time, or along a length, or in a given area or volume.

In fact, there are 2 (related) variables we may be interested in:

– the number of occurrences in a given interval.

– the 'distance' between successive occurrences.

For example, we may be interested in the occurrence of :

- system breakdowns through time
- arrivals of telephone calls through time
- accidents over a specified period of time
- component failures through time

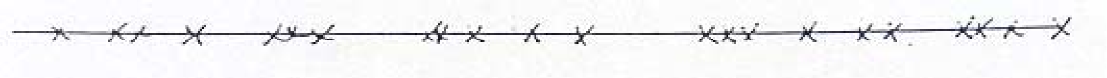
} *Random in time*

- flaws along the length of a wire
- surface defects on an area of video screen
- bacteria in a volume of liquid.

} *Random in space*

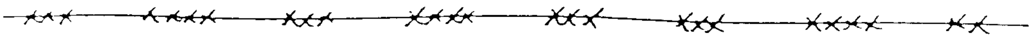
Consider the example of flaws occurring along a length of wire.

We are only interested in trying to model events which occur **randomly**. If flaws occur randomly throughout the length of the wire then we would expect a pattern something like:

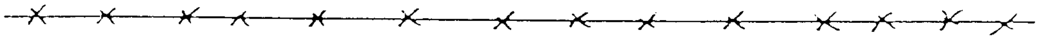


Note that the following patterns are **NOT** random:

BUNCHED
(ie. events 'attract' each other)
eg. item fails, causing other related items to fail immediately (so the system experiences several failures close together)



SPACED
(ie. events 'repel' each other)
eg. item which has lifetime of 1000 uses fails, is replaced, then system OK for another 1000 uses.



Neither of these last two situations is of interest here.

Variables of Interest

As mentioned previously, in the case of random occurrences of some event, such as those described above, there are generally two variables of interest:

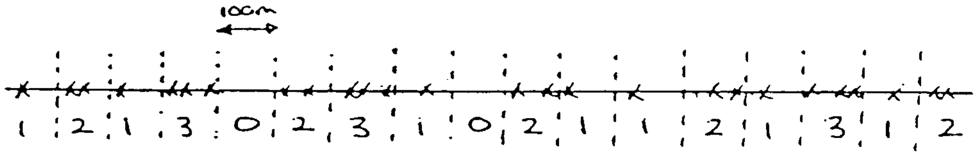
– the number of events

– the distance between events.

Let us continue with the above example, where the random event is the occurence of a flaw in a length of wire. The two variables are then

(i) *<u>the number of flaws in a given length of wire</u>*

For example:

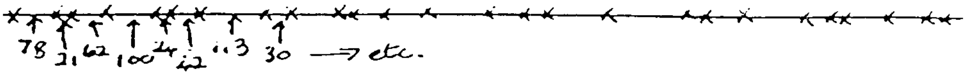


No. of flaws per 100m

Note that the number of flows in a given length of wire is a <u>discrete</u> variable (in general, a **count** of the number of times the event occurs in given a space/time).

(ii) *<u>the distance between successive flaws</u>*

For example:



Distance
between
flaws (m)

The distance between successive flaws is a <u>continuous</u> variable (in general, a **measurement** of the distance or time between successive events).

In general,

If	**(i) An event is randomly occurring in time, or space.**
and	**(ii) Event is occurring at a constant rate, λ, over the time (or space) of interest.**

then the events are said to follow a **POISSON PROCESS**.

If the above two conditions are satisfied then the probability distribution of the two variables of interest

– number of events

– distance between events

can be derived from statistical theory.

Examples of λ are:

- 2 breakdowns per month
- 4.7 calls per minute
- 5 flaws per l00 m
- 0.2 defects per 100 sq cm of screen

etc.

Thus

λ represents the number of times the event is expected to happen in given time/space

Then:

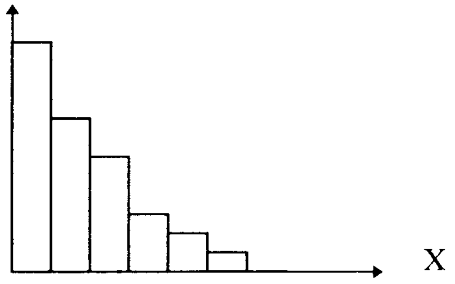
- the **number of events** in some interval follows a *Poisson distribution*.
- the **distance between events** follows an *Exponential distribution.*

The Poisson distribution will be considered in section 2.10.

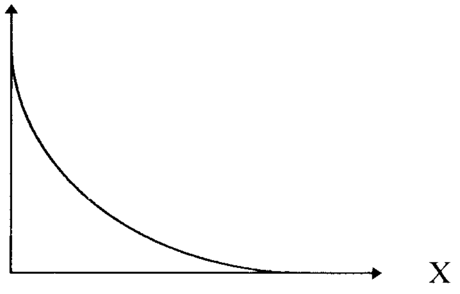
Summary

If events occur at random, at a constant rate of λ then the 'gaps' between successive events have an EXPONENTIAL DISTRIBUTION specified by λ.

A histogram of data collected on such a variable would look like:



The probability distribution (ie. the pattern expected in the population) looks like:



Thus, the probability of a small ‘gap’ is greater than the probability of a large ‘gap’.

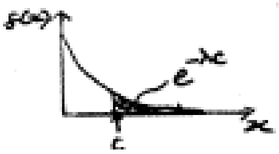
In this case, the variable is continuous, (‘distance between occurrences’ can take any numerical value within some range) and so instead of defining the probability associated with each individual value of the random variable, we talk in terms of the probability that the variable X lies between 2 values.

The probability density function (pdf) then has equation

$$\boxed{f(x) = \lambda e^{-\lambda x}}$$

2.7.3 Applications

The probability that the gap between 2 successive events, X, is less than some given value, C, denoted by $P(X \leq C)$, is then the area under the pdf curve between 0 and C. This function is easy to integrate, unlike the normal or lognormal pdf's, and gives the following results:



$$P(X \leq C) = 1 - e^{-\lambda c}$$

$$P(X > C) = e^{-\lambda c}$$

The function $P(X \leq c) = 1 - e^{-\lambda c}$ is called the **cumulative distribution function (cdf)** or sometimes just the distribution function. It is analogous to the cumulative frequency distribution of a sample described in section 1.3.

Also of interest:

The MEAN value of X (the gap between successive events) is given by $1/\lambda$

For example, (i) if breakdowns occur at a rate of $\lambda = 2$ per year then expected time between breakdowns = ½ year;

(ii) if flaws in a wire have $\lambda = 0.015$ per metre then on average we expect the distance between flaws to be 1/0.015 = 66.7 metres.

Estimating λ

Since the mean of X is $1/\lambda$ in the population then given a sample of values from an exponential distribution (with sample mean $\overline{X}$), we can estimate λ as $\frac{1}{\overline{X}}$. Packages such as Minitab which are used to fit a distribution to a set of data will provide this estimate automatically.

Example

Failures of an electronic component follow a Poisson process. The failure rate is 1 per 200 hours on average.

(i) Find the probability that the component is still working after 500 hours of operation?

(ii) What proportion of components will fail before 100 hours of operation?

Solution

If λ = 1 per 200 hours *(i.e. mean lifetime = 200 hours)* we require

P(X > 500) = $e^{-\lambda c} = e^{-\left(\frac{1}{200}\right)(500)} = e^{-2.5}$ = <u>0.0821</u> *(since 500 hours is 2.5 times the mean)*

P(X ≤ 100) = 1 - $e^{-\left(\frac{1}{200}\right)(100)}$

= 1 – $e^{-0.5}$ *(since 100 hours is 0.5 times the mean)*

= 1 – 0.6065

= <u>0.3935</u>

Note

Saying that the failures follow a Poisson process means that the electronic component is failing purely at random throughout the period of interest (500 hours or 100 hours) and that the failure rate does not change over that interval. If this is not the case, one of the other lifetime distributions may be appropriate - particularly the distribution described in the next section.

‘No Memory’

An interesting implication of a Poisson process is that no matter how big the ‘gap’ since the last event, the probability that the event occurs in the next instant remains the same. For example, the chance that the component in the above example works for at least another 500 hours is the same if the component has just started working or if it has been working for 1000 years! The system is said to have *‘no memory’*.

Applications To Reliability

The probability that a component is working at time t is called its **RELIABILITY** at time t. Reliability is thus a function of time and if the component is failing at random at a constant rate λ (ie. in a Poisson process) we have the reliability function:

$$\boxed{P(\text{Lifetime} > t) = R(t) = e^{-\lambda t}}$$

We will consider this in more detail in the next section.

2.8 THE WEIBULL DISTRIBUTION

2.8.1 Parameters and Functions

The Weibull distribution has pdf:

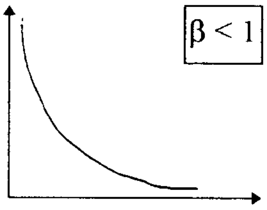
$$\boxed{f(x) = \beta\alpha^{-\beta}\, x^{\beta-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}} \quad \text{for } x > 0$$

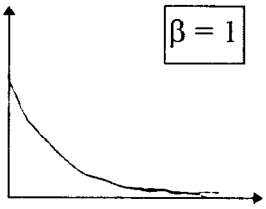
It has two positive parameters:

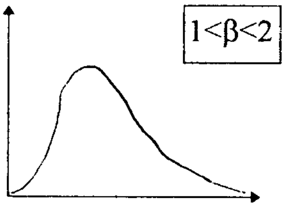
α the **scale** parameter

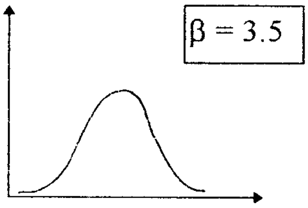
β the **shape** parameter

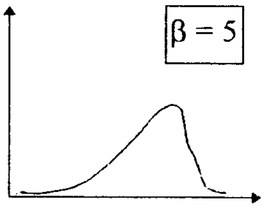
It is a very flexible distribution widely used in reliability modelling because of the variety of shapes it can take. Thus it can be used to model many different types of lifetime data. Different values for the parameter β give the following shapes:

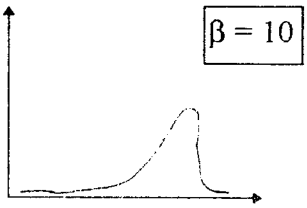




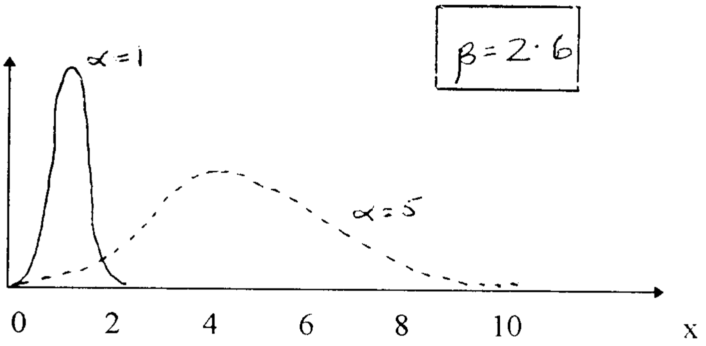








The effect of different values for the parameter α is to spread out the basic distribution shape over a different part of the X scale.



For a given set of data, estimates of α and β can be obtained using an important statistical technique called the **Method of Maximum Likelihood**. Distribution-fitting packages such as Minitab provide these maximum likelihood estimates (MLE's) automatically. When fitted to lifetime data, estimated value of β has important implications for the manner in which the item is failing (see section 3).

Special Cases

(i) When β = 1, the Weibull pdf reduces to

$$f(x) = 1.\alpha^{-1}x^{1-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{1}\right\} = \frac{1}{\alpha}\exp\left(-\frac{x}{\alpha}\right)$$

i.e. $f(x) = \alpha' e^{-\alpha' x}$ where $\alpha' = \dfrac{1}{\alpha}$

which is the pdf of an exponential distribution with rate α' (or mean $\dfrac{1}{\alpha'}$).

In other words, β = 1 implies failures are at a constant rate.

(ii) When β = 2, the Weibull pdf reduces to a form known as the RAYLEIGH distribution which will sometimes be met in the reliability literature.

2.8.2 Applications

As with the other distributions, probabilities can be estimated from the Weibull probability plot or found more accurately using the *cumulative distribution function*. The Weibull pdf can be integrated to give:

$$P(X \le c) = 1 - \exp\left\{-\left(\frac{c}{\alpha}\right)^{\beta}\right\}$$

$$P(X > c) = \exp\left\{-\left(\frac{c}{\alpha}\right)^{\beta}\right\}$$

Example

Probability plots suggest that the lifetime of a switch (in thousands of activations) has a Weibull distribution with parameters $\alpha = 2.2$ and $\beta = 1.5$.

(i) Find the probability that a switch will fail before it has been operated 2000 times.

(ii) What percentage of switches will still be working after being activated 3000 times.

Solution

(i) 2000 activations represents 2 time units. So,

$$P(X \leq 2) = 1 - \exp\left\{-\left(\frac{2}{2.2}\right)^{1.5}\right\}$$

$$= 1 - e^{-0.8668}$$

$$= 1 - 0.420$$

$$= \underline{\mathbf{0.580}}$$

(ii) 3000 activations represents 3 time units. So,

$$P(X > 3) = \exp\left\{-\left(\frac{3}{2.2}\right)^{1.5}\right\}$$

$$= e^{-1.592}$$

$$= 0.203$$

Thus **20.3%** of switches will still be working after 3000 activations.

(NB This value is defines as the RELIABILITY at 3000 activations.)

2.9 OTHER CONTINUOUS DISTRIBUTIONS

The four distributions of sections 2.5-2.8 are, of course, not the only ones used in engineering. In this section we mention a few more important distributions that may be met in the literature and introduce the general methodology for working with continuous distributions.

2.9.1 Other Named Distributions

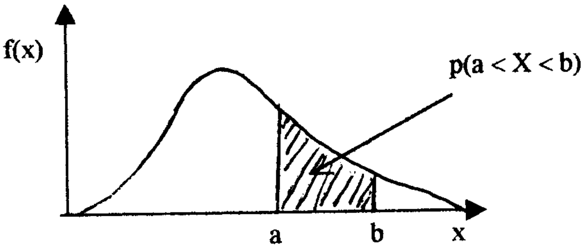
- If a random variable is equally likely to take on *any* value within some range, it is said to have a **uniform distribution**.

- We have seen that if events occur in a Poisson process (i.e. at random at a constant rate) then the length until the first occurrence of the event has an exponential distribution. Generalizing this, the length (of time or distance) until the event occurs *r* times has an **Erlang distribution**. This is actually a special case of a more general distribution called the **gamma distribution** where the parameter *r* is not restricted to being an integer.

- Random variables that involve some aspect of 'extremeness' (e.g. minimum or maximum reservoir levels, river flows or material strength) are often modelled by a **Gumbel distribution.**

2.9.2 Using Calculus

We have seen that the probability density function (pdf) is the equation of the curve that models the behaviour of a random variable (in the sense of describing how likely it is that different values will occur) and that probabilities associated with this variable correspond to areas under the curve.



Thus in general if the random variable X has pdf f(x):

$$\boxed{p(a < x < b) = \int_a^b f(x)dx}$$

The function f(x) is sometimes easy to integrate (e.g. the exponential distribution) and sometimes not (e.g. the normal distribution). For common distributions, tables or packages are available so that these areas (probabilities) can be found easily. Otherwise, the integration will have to be performed manually. The following examples illustrate this.

Notes

(i) What functions can serve as pdf's? Any function f(x) can serve as a pdf **provided:**

- f(x) is always ≥ 0 (i.e. probabilities cannot be negative.)
- $\int_{-\infty}^{+\infty} f(x) = 1$ (i.e. the total area under the curve is 1 since X must take on *some* value between $-\infty$ and $+\infty$)

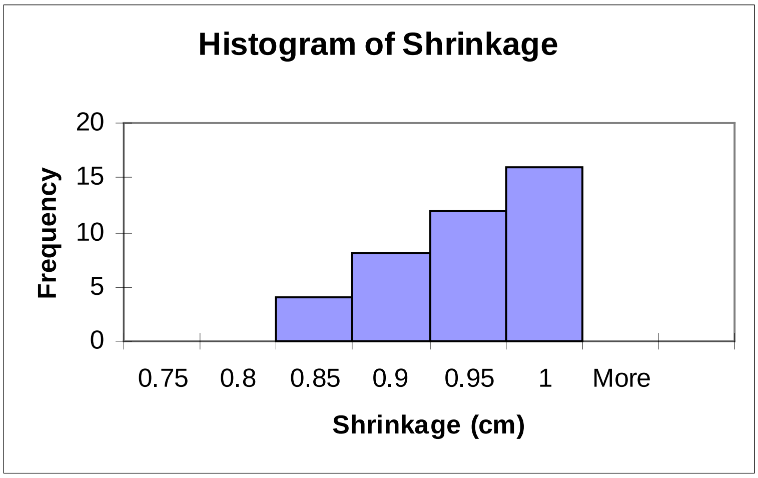
(ii) We have already met the **cumulative distribution function** (cdf) in sections 2.7.3 and 2.7.4. This is $p(X \le x)$ and, in general, it is denoted F(x). Thus:

$$\boxed{F(x) = \int_{-\infty}^{x} f(x)dx}$$

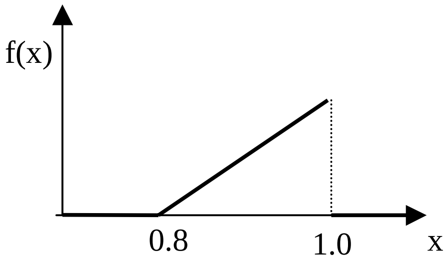
It is sometimes more convenient to work with F(x) than f(x).

Example 1

Speedometer cables are supposed to have a diameter of 1cm. However, during storage after a heat treatment process, may of them suffer from shrinkage around the core to such an extent that some of them end up with only a 0.8cm diameter. Data is collected on this shrinkage from 40 cables and the histogram looks like:



This suggests that the frequency 'curve' could be a straight line as follows:



That is, a straight line passing through (0.8, 0) with some slope k giving:

$$f(x) = \begin{cases} k(x-0.8) & 0.8 < x < 1.0 \\ 0 & otherwise \end{cases}$$

where *k* is some constant.

(i) *What is the value of* k?

For *f(x)* to be a pdf we require that:

- *f(x)* ≥ 0 for all values of X (which it is!)

and

- $\int_{-\infty}^{+\infty} f(x) = 1$

Hence we require *k* so that:

$$\int_{-\infty}^{+\infty} k(x-0.8) = 1$$

i.e. $\int_{0.8}^{1.0} k(x-0.8) = 1$ (i.e. over the range where f(x) >0)

Integrating this gives:

$$k\left[\frac{x^2}{2} - 0.8x\right]_{0.8}^{1.0} = 1$$

$$\rightarrow k\left((0.5-0.8) - \left(0.8^2/2 - 0.8^2\right)\right) = 1$$

$$\rightarrow k(0.02) = 1$$

$$\rightarrow k = 50$$

Thus the pdf we can use to model the behaviour of X (diameter after shrinkage) is:

$$\boxed{f(x) = \begin{cases} 50(x-0.8) & 0.8 < x < 1.0 \\ 0 & otherwise \end{cases}}$$

(ii) *What proportion cables will shrink to a diameter of less than 0.9cm?*

We require *p(X < 0.9)*

$$= \int_{0.8}^{0.9} 50(x - 0.8)dx$$

$$= 50\left[x^2\!\big/ 2 - 0.8x\right]_{0.8}^{0.9}$$

= **0.25**

(iii) *What, on average, do we expect the shrinkage to be?*

This is the mean value that we have denoted μ. It is often referred to as the **expected value** and written E(X).

The mean of a probability distribution can be found from:

$$\boxed{\mu \quad (or \quad E(X)) = \int_{-\infty}^{+\infty} x.f(x)dx}$$

Thus, for the shrinkage problem, this equals:

$$= 50\int_{0.8}^{1.0} x(x - 0.8)dx$$ (note that X is positive in the range 0.8 to 1.0)

$$= 50\left[x^3\!\big/ 3 - 0.8x^2\!\big/ 2\right]_{0.8}^{1.0}$$

= 0.933cm

This is the expected *diameter* after shrinkage from 1cm.

Thus, expected *shrinkage* is 1 – 0.933 = **0.067 cm.**

Example 2

The pdf of the (coded) lengths of a metal rod (X) is:

$$f(x) = \begin{cases} 1.5x^2 & -1 < x < 1 \\ 0 & otherwise \end{cases}$$

(i) Sketch the form of the pdf.
(ii) Confirm that this is a valid pdf.
(iii) Find the cumulative distribution function.
(iv) Find *p(X > 0)* and *p(–0.5 < X < 0.5)*
(v) Find the mean value (or expected value) of X.
(vi) The **variance** of X is defined as $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$

(or equivalently $\sigma^2 = \int x^2 f(x)dx - \mu^2$).

Hence find the *standard deviation (σ)* of the lengths of these metal rods.

Further Examples will be provided.

2.10 DISCRETE PROBABILITY DISTRIBUTIONS

The two distributions considered here can be used to model **counts** i.e. the number of times that something happens.

These are

- the POISSON distribution
- the BINOMIAL distribution.

2.10.1 The Poisson Distribution

Recall from section 2.7.2 that:

If events occur at random, at a rate of λ then the number of times the event occurs in a given time/space has Poisson distribution specified by the mean rate of occurrence λ.

The probability of a particular number, r, of occurrences of the event in some interval, given that the event is known to have a mean rate λ is:

$$\boxed{\begin{aligned} \text{P(event occurs r times)} &= P(X = r) \\ &= \frac{\lambda^r e^{-\lambda}}{r!} \end{aligned}}$$

Notes

(i) The variable 'number of occurrences' (X) is discrete (i.e. can only take certain, specific, values) and so only certain probabilities are defined, namely.

P(0) = P(X= 0) = Probability that there are no occurrences.

P(1) = P(X = 1) = Probability that there is one occurrence.

P(2) = P(X = 2)

P(3) = P(X = 3)

P(4) = P(X = 4) etc.

All other events, such as X = 1½ (i.e. there are 1½ events) are impossible, and, following the laws of probability outlined in Section 2.2, have the probability zero.

i.e. P(X = 1½) = 0 etc.

Thus, the Poisson distribution is an example of a discrete distribution.

(ii) We are trying to estimate the number of events in some interval, based on the known mean rate λ.

λ must be over the same interval as that for which we are trying to estimate. If it is not, we must convert it.

e.g. suppose we know that 6 flaws occur on average in 10 m of wire, and wish to estimate the probability of 0 flaws *in any given 2 m length of wire,* i.e. P(X = 0).

We can use the above formula, namely

$$P(X=0)=\frac{\lambda^{\circ}e^{-\lambda}}{0!}$$

but λ must be the mean number of flaws *in 2 m of wire*.

Then λ' = 6 flaws per 10 m.

$$=\frac{6}{10} \text{ flaws per 1 m.}$$

So $\lambda=\left(\frac{6}{10}\right)\times 2=1.2$ flaws per 2 m.

and it is this value of λ we use in the above equation.

NB In general λ is the number of times you expect the event to occur in a given time/space.

(iii) <u>*Definition of Factorials*</u>

0! = 1 (read 0 factorial)
1! = 1 (read 1 factorial)
2! = 2 × 1 = 2
3! = 3 × 2 × 1 = 6
4! = 4 × 3 × 2 × 1 = 24
etc.

(Thus, for example, P(X = 0) in part (ii) is $\frac{1.2^0 e^{-1.2}}{0!}=e^{-1.2}=0.3012$)

(iv) Ordinary rules of probability apply, as outlined in sections 2.2 and 2.3,

e.g. $P(X\le 2)=P(X=0)+P(X=1)+P(X=2)=1-P(X>2)$ etc

and

$$\sum_{r=0}^{\infty}P(X=r)=\quad 1$$ (Total probabilities sum to 1.)

The following examples illustrate how the formula may be used.

Example 1

Breakdowns in a communication system occur at random with an average of 19 per year. What is the probability that in a particular month there are:

(a) no breakdowns;
(b) less than three breakdowns;
(c) at least four breakdowns?

Solution

Since

(i) The number of breakdowns is occurring randomly.

and

(ii) the event has a mean no. of occurrences which is known and constant

(i.e. breakdowns occur in a *Poisson Process*) then the number of breakdowns follows a Poisson distribution.

We are given

$$\lambda' = 19 \text{ per year}$$

but want to estimate over the period of one month, so we must first convert the mean to breakdowns per month before we can use the formula to estimate probabilities. Then

$$\lambda = \frac{19}{12} \text{ per month}$$

i.e. $\lambda = 1.58$ per month (i.e. we expect 1.58 breakdowns per month.)

Using the above formula we have:

(a) $$P(X = 0) = \frac{1.58^0 \, e^{-1.58}}{0!}$$

$$= e^{-1.58}$$

$$= \underline{0.206}$$

(b) $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$

where $P(X = 0) = 0.206$

$$P(X = 1) = \frac{1.58^{1}\,e^{-1.58}}{1!}$$

$$= 0.325$$

$$P(X = 2) = \frac{1.58^{2}\,e^{-1.58}}{2!}$$

$$= \underline{0.257}$$

Thus $P(X < 3) = 0.206 + 0.325 + 0.257$

$$= \underline{0.788}$$

(c) $P(X \geq 4) = 1 - P(X = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3)$

$$= 1 - \{P(X < 3) + P(X = 3)\}$$

We already have $P(X < 3)$ in part (b) so we just require

$$P(X = 3) = \frac{1.58^{3}\,e^{-1.58}}{3!}$$

$$= \underline{0.135}$$

Thus $P(X \geq 4) = 1 - \{0.788 + 0.135\}$

$$= \underline{0.076}$$

Example 2

Calls to a computer plotter occur at random at a constant rate of 2 calls per minute during peak time. Find the probability that in a given 3 minute period during the peak period there are at least 3 calls.

Solution

$\lambda = 2$ per minute

So $\lambda' = 6$ per 3 minutes

Let

X = No. of calls during 3 minute period

we want

$$P(X \geq 3) = 1 - P(X \leq 2)$$

$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

Then

$$P(X = 0) = \frac{6^0 e^{-6}}{0!} = 0.0025$$

$$P(X = 1) = \frac{6^1 e^{-6}}{1!} = 0.0149$$

$$P(X = 2) = \frac{6^2 e^{-6}}{2!} = 0.0446$$

Thus

$$P(X \geq 3) = 1 - (0.0025 + 0.0149 + 0.0446)$$

$$= 1 - 0.062$$

$$= \underline{\mathbf{0.938}}$$

Mean and Standard Deviation

The mean and standard deviation of a Poisson distribution with rate λ are given by:

Mean $(\mu) = \lambda$ *Standard deviation* $(\sigma) = \sqrt{\lambda}$

Or, equivalently, the *variance* $\left(\sigma^2\right) = \lambda$.

Then, from example 2 above, the mean and standard deviation of the number of calls in a 3 minute period are 6 and $\sqrt{6} = 2.45$ respectively.

2.10.2 The Binomial Distribution

Consider each of the following questions.

(i) *A fair coin is tossed 7 times. What is the probability that 7 heads turn up?*

(ii) *An acceptance sampling plan involves inspecting a sample of 20 items from each batch and rejecting the batch if 2 or more defectives are found. What is the probability that a batch with 10% defectives will be rejected by the plan?*

(iii) *35% of digitally transmitted images contain an error of some kind. If 5 images are transmitted, what is the probability that no more than 3 of them contain an error?*

In each case:

(a) **We have a fixed number, n, of independent trials.**

For the above examples:

(i) n = *7 tosses*

(ii) n = *20 items*

(iii) n = *5 images*.

(b) **We want the probability associated with the number of times a particular event occurs.**

Let the particular event be denoted by the random variable X. Then

(i) X = *the no. of times we get a head,*

(ii) X = *the no. of times we get a defective,*

(iii) X = *the no. of times we get an error*.

Note that the maximum number of times it can occur is n and the minimum number is 0.

(c) **The event has the same probability p, from trial to trial.**

Here

(i) $p = ½$ *(The probability that one toss gives a head.)*

(ii) $p = 0.1$ *(The probability that an item is defective.)*

(iii) $p = 0.35$ *(The probability that an image has an error.)*

In such situations the number of times the event occurs in the fixed number of trials has a **binomial distribution** and the probabilities required can be obtained using a standard formula.

In general, if

(i) We have n independent trials

(ii) Each trial has p = probability of success
q = 1 – p = probability of failure.

Then X, the number of 'successes', has a **binomial distribution**.

X can take values 0, 1, 2, …, n and the probability that it takes the value r can be obtained from the formula:

$$\boxed{P(X = r) = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}}$$

where $0! = 1$ (read 0 factorial)
$1! = 1$ (read 1 factorial)
$2! = 2 \times 1 \quad = \; 2$
$3! = 3 \times 2 \times 1 \quad = \; 6$
$4! = 4 \times 3 \times 2 \times 1 = 24$

etc.

The first term gives the number of ways of getting r successes out of n. The rest gives the probability of getting any particular sequence of r successes and consequently n-r failures.

<u>Example 1</u> *(from earlier)*

A fair coin is tossed 7 times. What is the probability that 7 heads turn up?

<u>Solution</u>

Define a *'head'* as a *'success'* then let

X = number of heads obtained

then X has a binomial distribution, and we want to find P(X = 7).

In this case:

n = 7 (no. of samples)

p = ½ (the probability of a '*success*' in one trial
= the probability of a '*head*' in one toss.)

r = 7 (no. of successes)

Then

$$p(X = 7) = \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) \times \cdots \times \left(\frac{1}{2}\right) \qquad \text{(7 times)}$$

$$= 0.0078$$

Alternatively, from the formula:

$$P(X = 7) = \frac{n!}{r!(n - r)!} p^r (1 - p)^{n-r}$$

$$= \frac{7!}{7!(7 - 1)!} \left(\frac{1}{2}\right)^7 \left(1 - \frac{1}{2}\right)^{7-7}$$

$$= 1 \times \left(\frac{1}{2}\right)^7 \times \left(\frac{1}{2}\right)^0$$

$$= \left(\frac{1}{2}\right)^7 = 0.0078$$

<u>Example 2</u>

An acceptance sampling plan involves inspecting a sample of 20 items from each much larger batch and rejecting the whole batch if 2 or more defectives are found. What is the probability that a batch with 10% defectives will be rejected by the plan?

<u>Solution</u>

Here X = number of defectives
$n = 20$
r = 2 or more
$p = 0.1$ (the probability that one item is defective.)

and we want to find $P(X = 2 \text{ or more})$.

Then $P(X = 2 \text{ or more}) = 1 – P(X = 0 \text{ or } X = 1)$

where $P(X = 0 \text{ or } X = 1) = P(X = 0) + P(X = 1)$

Now $$P(X = r) = \frac{n!}{r!(n - r)!} p^r (1 - p)^{n-r}$$

i.e. $$P(X = 0) = \frac{20!}{0!20!} (0.1)^0 (0.9)^{20-0}$$

$$= (0.9)^{20}$$

$$= 0.1216$$

$$\mathrm{P(X = 1)} = \frac{20!}{1!19!}(0.1)^1(0.9)^{19}$$

$$= 0.2702$$

Then P(X = 2 or more) = 1 – P(X = 0 or X = 1)

= 1 – [0.1216 + 0.2702]

= 0.6082

This is the probability that a batch with 10% defectives will be rejected by the acceptance sampling plan. It is a typical *quality control* application.

Example 3 *from beginning of section.*

The number of images (X) with an error has a binomial distribution with *n = 5*
and *p = 0.85*.

We require $p(X \le 3) = p(0) + p(1) + p(2) + p(3)$
More conveniently, this is $1 - [p(4) + p(5)]$
where $p(4) = \frac{5!}{4! \times 1!}(0.35)^4(0.65)^1 = 0.0488$
and $p(5) = \frac{5!}{5! \times 0!}(0.35)^5(0.65)^0 = 0.0053$

Then $p(X \le 3) = 1 - (0.0488 + 0.0053) = \mathbf{0.9459}$

Mean and Standard Deviation

The mean and standard deviation of a binomial distribution with n trials and success probability p are given by:

$$\textit{Mean } (\mu) = np \qquad \textit{Standard deviation } (\sigma) = \sqrt{np(1-p)}$$

Or, equivalently, the *variance* $(\sigma^2) = np(1-p)$.

Then, from example 2 above, the mean number of defectives in a batch is
$np = 20 \times 0.1 = 2$ and the standard deviation of the number of defectives in a batch
is $\sqrt{np(1-p)} = \sqrt{20 \times 0.1 \times 0.9} = 1.34$.

2.11 REVIEW

In this section we have considered probability distributions as a means of modelling the behaviour we expect from random variables in the population. Emphasis has been placed on useful model for data on lifetimes or failure times as an introduction to Reliability ideas covered in section 3.

Distributions

We have concentrated on four continuous distributions:

- Normal
- Lognormal
- Exponential
- Weibull

and considered how probability plots can be used to decide upon an appropriate model.

Parameters

The exact distribution to use depends on the values of parameters which may be estimated from the data. These parameters and their interpretation are:

	Parameters	Interpretation
Normal	μ, σ	*Location and spread (mean and S.D)*
Lognormal	μ, σ	*Location and spread of log(X)*
Exponential	λ	*Rate of occurrence*
Weibull	α, β	*Scale and shape*

Mechanisms

These distributions may arise if the variable measured is

	Process
Normal	*the result of adding many independent effects*
Lognormal	*the result of multiplying many independent effects*
Exponential	*the 'gaps' between events occurring at random at a constant rate (i.e. a Poisson Process)*
Weibull	*the 'gaps' between events occurring with a variety of failure patterns.*

Finding Probabilities

Probabilities associated with each distribution can be estimated from probability plots or, more accurately, from:

Normal	*Normal tables*
Lognormal	*Normal tables for log(X)*
Exponential	*Cumulative distribution function* $P(X \le c) = 1 - \exp(-\lambda c)$
Weibull	*Cumulative distribution function* $P(X \le c) = 1 - \exp\left\{-\left(\frac{c}{\alpha}\right)^{\beta}\right\}$

Percentage points (or percentiles – the values associated with given probabilities) can be found by working backwards. The application of integral calculus at a more general level was also considered.

Finally, two discrete distributions for counts were introduced:

- the Poisson distribution
- the binomial distribution

The Poisson distribution, like the exponential, arises from a Poisson process when counting the number of times the event occurs. The Poisson process is often a useful first approximation when dealing with the reliability of real systems as described in section 3.