CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

The manufacturing of tools and special equipment is part of human nature. In earlier days faults and accidents were the only way of learning to make safer and more reliable equipment. Before structural design became an engineering science, the reliability of a bridge was tested with a team of elephants. If it collapsed, a stronger bridge was built and tested again! Obviously, these methods could not continue and as human skills developed a wide variety of very reliable items and structures were designed and manufactured. One example is the undersea telephone cables built by Bell Telephone Laboratories.

Man's earliest preoccupation with reliability was undoubtedly related to weaponry. Interest flowered as a result of the terrible non-reliability of electronic weapons systems used during World War II. Increasingly complex systems, such as the first missiles, also emphasized the importance of successful operation of equipment in a specific environment during a certain time period. The V-1 missile, developed in Germany with high-quality parts and careful attention, was catastrophic: the first 10 missiles either exploded on the launching pad, or landed short of their targets.

Technological developments lead to an increase in the number of complicated systems as well as an increase in the complexity of the systems themselves. With remarkable advancements made in electronics and communications, systems became more and more sophisticated. Because of their varied nature, these problems have attracted the attention of scientists from various disciplines especially the systems engineers, software engineers and the applied probabilists. An overall scientific discipline, called *reliability theory*, that deals with the methods and techniques to ensure the maximum effectiveness of systems (from

known qualities of their component parts) has developed. *'Reliability theory introduces quantitative indices of the quality of production'* (Gnedenko et al. (1969)) and these are carried through from the design and subsequent manufacturing process to the use and storage of technological devices. Engineers, Scientists and Government leaders are all concerned with increasing the reliability of manufactured goods and operating systems. As '*Unreliability has consequences in cost, time wasted, the psychological effect of inconvenience, and in certain instances personal and national security'* (Lloyd & Lipow (1962)). In 1963 the first journal on reliability, IEEE-Transactions on Reliability saw the light.

Due to the very nature of the subject, the methods of Probability theory and Mathematical statistics (information theory, queuing theory, linear and nonlinear programming, mathematical logic, the methods of statistical simulation on electronic computers, demography, manufacturing, etc.), play an important role in the problem solving of reliability theory. Other areas include contemporary medicine, reliable software systems, geoastronomy, irregularities in neuronal activity, interactions of physiological growth, fluctuations in business investments, and many more. In human behaviour mathematical models based on probability theory and stochastic processes are helpful in rendering realistic modelling for social mobility of individuals, industrial mobility of labour, educational advancements, diffusion of information and social networks. In the biological sciences stochastic models were first used by Watson and Galton (1874) in a study of extinction of families. Research on population genetics, branching process, birth and death processes, recovery, relapse, cell survival after irradiation, the flow of particles through organs, etc. then followed. In business management, analytical models evolved for

the purchasing behaviour of the individual consumer, credit risk and term structure, income determination under uncertainty and many more related subjects. Traffic flow theory is a well known field for stochastic models and studies have been developed for traffic of pedestrians, freeways, parking lots, intersections, etc.

Problems encountered in the design of highly reliable technical systems have led to the development of high-accuracy methods of reliability analysis. Two major problems can be identified, namely:

- creating classes of probability-statistical models that can be used in the description of the reliability behaviour of the system, and
- developing mathematical methods for the examination of the reliability characteristic of a class of systems.

Considering only redundant systems the classical examples are the models of Markov processes with a finite set of states (in particular birth and death processes) (Gnedenko et al. (1969)), Barlow (1984), Gertsbakh (1989) and Kovalenko et al. (1997)), the renewal process method (Cox (1962)), the semi-Markov process method and its generalizations (Cinlar (1975a, b)), generalized semi-Markov process (GSMP) method (Rubenstein (1981)), spacial models for coherent systems (Aven (1996)) and systems in random and variable environment (Ozekici (1996)) and Finkelstein (1999a, b, c)).

Depending on the nature of the research, the applicable form of reliability theory can be introduced to each. A stochastic analysis is made based on some good probability characteristics. It is, however, not simply a case of changing terminology in standard probability theory (say, “random variable” changes to “lifetime”), but reliability

distinguishes itself by providing answers and solutions to a series of new problems not solved in the "standard" probability theory framework. Gertsbakh (1989) points out that reliability,

- of a system is based on the information regarding the reliability of the system's *components*
- gives a mathematical description of the *ageing process* with the introduction of several formal notations of ageing (failure rate, etc.)
- introduces well-developed techniques of *renewal theory*
- introduces *redundancy* to achieve optimal use of standby components (an excellent introduction to redundant systems is given in Gnedenko et al. (1969)
- includes the theory of *optimal preventative maintenance* (Beichelt and Fischer (1980))
- is a study of *statistical inference* (often from censored data)

Generally, the mathematical problems of lifetime studies of technical objects (reliability theory) and of biological entities (survival analysis) are similar, differing only in the notation. The term "lifetime" therefore does not apply to lifetimes in the strictest literal sense, but can be used in the figurative sense. The idea is that the statistical analysis done in this thesis should be true in any of the applicable disciplines, although the notation is mostly as for engineering (systems, components, units, etc.). With minor modifications the discipline can be changed to biological, or financial, etc.

1.2 FAILURE

'A failure is the result of a joint action of many unpredictable, random processes going on inside the operating system as well as in the environment in which the system is operating.' (Gertsbakh (1989)). Functioning is therefore seriously impeded or completely stopped at a certain moment in time and all failures have a stochastic nature. In some cases the time of failure is easily observed. But if units deteriorate continuously, determination of the moment of failure is not an easy task. In this study we assume that failure of a unit can be obtained exactly. Failure of a system is called a *disappointment* or a *death* and failure results in the system being in the *down state*. This can also be referred to as a *breakdown* (Finkelstein (1999a)).

Zacks (1992) points out that there are two types of data to consider, namely:

- data from continuous monitoring of a unit until failure is observed
- data from observations made at discrete time points, therefore failure counts

Villemeur (1992) gives an extensive list of possible failures and inter-dependent failures. There are catastrophic failures, determined by a sharp change in the parameters and drift failures (the result of wear or fatigue), arising as a result of a gradual change in the values of the parameters.

1.3 REPAIRABLE SYSTEMS

Failed units of a system may be replaced by new ones, but this may prove to be expensive. To repair the failed units at a repair facility is usually a more cost-effective

option than replacement. A *repairable* (or *renewable*) *system* can be described as one where the system can be made operable again. If a system can be renewed, the reliability is increased, resulting in an increase in its time of service. If no repair facility is free, failed units queue up for repair. The life time of a unit while on-line, while in standby as well as the repair times, are all independent random variables. It is assumed that the distributions of these random variables are known and that they have probability density functions.

Repairable systems have been the subject of intensive investigation for a long time. Different random variables can form the basis for research, such as

- availability (or non-availability) and reliability
- time necessary for repair
- number of repairs that can be handled
- switch over time to and from the repair facility
- possibility of a vacation time for the repair facility, and many more.

Barlow (1962) considered some 'repairman' (or repair-facility) problems and they have much in common with queuing problems while Rau (1964) analyzed the problem of finding the optimum value of an *k-out-of*-n: G system for maximum reliability. Ascher (1968) has pointed out some inconsistencies in modelling of repairable systems by renewal theory. Several authors, notably Buzacott (1970), Shooman (1968) have used continuous time discrete state Markov process models for describing the behaviour of a repairable system. These models, although conceptually simple, are not practically feasible in the case of a large number of states. Gaver (1964), Gnedenko et al. (1969), Srinivasan (1966) and Osaki (1970a) have used semi-Markov processes for calculation of the reliability of a

system with exponential failures. Osaki (1969) has used signal flow graphs to discuss a two-unit system. With the use of semi-Markov processes Kumagi (1971) studied the effect of different failure distributions on the availability through numerical calculations. Branson and Shah (1971) also used semi-Markov process analysis to study repairable systems with arbitrary distributions. Srinivasan and Subramanian (1980), Venkatakrishnan (1975), Ravichandran (1979), Natarajan (1980) and Sarma (1982) have used regeneration point techniques to analyze repairable systems with arbitrary distributions. More references in this and related topics can be found in various papers by Subba Rao and Natarajan (1970), Osaki and Nakagawa (1976), Pierskalla and Voelker (1976), Lie et al. (1977), Kumar and Agarwal (1980), Birolini (1985) and Yearout et al. (1986) and Finkelstein (1993a, 1993b). Jain and Jain (1994) have considered the regulation of 'up' and 'down' times of a repairable system to improve the efficiency of the system.

1.4 REDUNDANCY AND DIFFERENT TYPES OF REDUNDANT SYSTEMS

In a *redundant system* more units are built into it than is actually necessary for proper system performance. Redundancy can be applied in more than one way and a definite distinction can be made between *parallel* and *standby* (sequential) redundancy. In parallel redundancy the redundant units form part of the system from the start, whereas in a standby system, the redundant units do not form part of the system from the start (until they are needed).

1.4.1 Parallel systems

A parallel redundant system with n units is one in which all units operate simultaneously, although system operation requires at least one unit to be in operation. Hence a system failure only occurs when all the components have failed.

Let k be a non-negative integer, such that $k \leq n$, counting the number of units in an n-unit system. It is customary to refer to such a system as *k-out-of-n* system.

1.4.2 k-out-of-n: F system

If *k-out-of-n* system fails, that is when k units fail, it is called an F-system. The functioning of a minimum number of units ensures that the system is up (Sfakianakis and Papastavridis (1993)).

1.4.3 k-out-of-n: G-system

A G-system is operational if and only if at least k units out of n units of the system are operational. Recent work related to this topic can be seen in Zhang and Lam (1998) and Liu (1998). Suppose a radar network has n radar control stations covering a certain area: the system can be operable if and only if at least k of these stations are operable. In other words, to ensure functioning of the system it is essential that a minimum number of units, k, are functioning.

Lately attention moved to load-sharing *k-out-of-n*: G systems, where

- the serving units share the load
- the failure rate of a component is affected by the magnitude of the load it shares.

1.4.4 n-out-of-n: G system

A series that consists of *n* units and when the failure of any one unit causes the system to fail. Although this type of system is not a redundant system, as all the units are in series and have to be operational, it can still be considered as a special vas of a *k-out-of-n* system. There are many papers on the reliability of these systems. Scheuer (1988) studied reliability for *shared-load k-out-of-n*: *G systems*, where there is an increasing failure rate in survivors, assuming identically distributed components with constant failure rates. Shao and Lamberson (1991) considered the same scenario, but with imperfect switching. Then Huamin (1998) published a paper on the influence of work-load sharing in non-identical, non-repairable components, each having an arbitrary failure time distribution. He assumed that the failure time distribution of the components can be represented by the accelerated failure time model, which is also a proportional hazards model when base-line reliability is Weibull.

1.4.5 Standby redundancy

Standby redundancy consists in attaching to an operating unit one or more redundant (*standby*) units, which can, on failure of the operating unit, be switched *on-line* (if operable). Gnedenko et al. (1969) classifies standby units as *cold, warm or hot*.

1. A *cold standby* is completely inactive and because it is not hooked up, it cannot (in theory) fail until it is replacing the primary unit. Also assume that, having been in a non-operating state its reliability will not change when it is put into an operating state.

2. A *warm standby* has a diminished load because it is only partially energized. The standby unit is not subject to the same loading conditions as the on-line unit and failure is generally due to some extraneous random influence. So, although such warm standby can fail, the probability of it failing is smaller than the probability of the unit on-line failing. This the most general type of standby because of hot standby's failure rate and cold standby's possible time lapse before it is operable.
3. A *hot standby* is fully active in the system (although redundant) and the probability of loss of operational ability of a hot standby is the same as that of an operating unit in the standby state. The reliability of a hot standby is independent of the instant at which it takes the place of the operable unit.

1.4.6 Priority redundant systems

A priority system consists of n (≥ 2) units in which some of the units are given *priority* (*p*-units) and the other units are termed as *ordinary* units (*o*-units). The operating on-line unit must be the *p*-unit and this *p*-unit is never used in the status of a standby and, in the event of a failure, it is immediately taken up for repair – if the repair facility is available. On the other hand, the *o*-unit only operates on-line when the *p*-unit has failed and is under repair. Different policies can be adopted (Jaiswal (1968)) if the *p*-unit fails during the repair of an *o*-unit, namely pre-emptive and non-pre-emptive priorities.

1.4.6.1 Pre-emptive priority

The repair of the *o*-unit will be interrupted by the *p*-unit if the *p*-unit fails when the

repair for *o*-units is on. After completion of the repair of the *p*-unit, the repair of the *o*-unit is continued in one of two ways:

(i) *pre-emptive resume*, where the repair of the *o*-unit continues from the previous point of interruption

(ii) *pre-emptive repeat*, where repair of the *o*-unit is started afresh after completion of the previous interruption. This implies that the time spent by the I-unit before it was pre-empted from the repair has no influence on the re-started repair time.

1.4.6.2 Non-pre-emptive priority

The repair of the o-unit continues and the repair of the *p*-unit is entertained only after completion of the repair of the *o*-unit.

1.5 INTERMITTENTLY USED SYSTEMS

When a system is turned on and off intermittently for the purpose of performing a certain function it is referred to as an *intermittently used system*. It is obvious that for such a system continuous failure free performance is not so absolutely necessary. In such cases consideration has to be given to the fact that the system can be in the down state during certain time intervals without any real consequence. The probability that the system is in the up state is not an important measure; what is really important is the probability that the system is available when needed. *Operational reliability* is thus a function of the readiness and the probability of continuous functioning over a specified period of time and it can grow or decline with age, depending on the nature of the system.

Gaver (1964) pointed out that it is pessimistic to evaluate the performance of an intermittently used system solely on the basis of the distribution of the time to failure. Srinivasan (1966), Nakagawa et al. (1976), Srinivasan and Bhaskar (1979a, 1979b, 1979c), Kapur and Kapoor (1978, 1980), Sarma (1982) and Yadavalli and Hines (1991) extended Gaver's results for two-unit and *n*-unit systems, and, obtained various system measures.

1.6 MEASURES OF SYSTEM PERFORMANCE

In the previous sections a brief discussion was given of the various types of redundant systems as discussed in the literature. In this section the discussion is about measures of system performance as applicable in different contexts (Barlow & Proschan (1965) and also Gnedenko et al. (1969)).

1.6.1 Reliability

Reliability engineering has developed, and advanced substantially during the past 50 years, mainly due to the use of high risk and complex systems (Beichelt (1997)). Reliability is a quantitative measure to ensure operational efficiency. *'The reliability of a product is the measure of its ability to perform its function, when required, for a specific time, in a particular environment. It is measured as a probability.'* (Leitch (1995)). This implies that reliability contains four parts, namely

- the *expected function* of a system
- the *environment* of a system (climate, packaging, transportation, storage, installation, pollution, etc.)
- *time*, which is often negatively correlated with reliability

- *probability*, which is time-dependent, thus causing the need for a statistical analysis.

One can distinguish between *mission reliability*, when a device is constructed for the performance of one mission only and *operational reliability*, when a system is turned on and off intermittently for the purpose of performing a certain function. In the latter case we refer to an *intermittently used system*.

Ordinarily the period of time intended is $(0, t]$.

Let $\{\phi(t), t \geq 0\}$ be the performance process of the system.

For fixed t this $\phi(t)$ is a binary variable, defined as follows:

$$\phi(t) = \begin{cases} \text{if the system is functioning at time t} \\ \text{if the system is in a failed state at time t.} \end{cases}$$

1.6.1.1 The reliability function

The *reliability function*, $R(t)$ gives the probability that the system does not fail up to t, that is

$$R(t) = P[\text{system is functioning in } (0, t]]$$
$$= P[\phi(u) = 0 \ \forall \ u \text{ such that } 0 < u \leq t].$$

1.6.1.2 Interval reliability

If the number of system failures in the interval $(t, t + x]$ is considered, the performance measure

$$R(t, x) = P[\phi(u) = 0 \ \forall \ u \text{ such that } 0 < u \leq t + x]$$

is referred to as the *interval reliability*.

If $t = 0$ the interval reliability becomes the reliability $R(x)$.

1.6.1.3 Limiting interval reliability

Limiting interval reliability is defined as the limit of $R(t, x)$ as $t \rightarrow \infty$, and is denoted $R_{\infty}(x)$.

1.6.1.4 Mean time to system failure

The expectation of the random variable representing the duration of time, measured from the point the system starts operating, till the instant it fails for the *first* time is called *mean time to system failure* (MTSF). This is obtained from the relation

$$\text{MTSF} = \int_0^{\infty} R(u)du\,.$$

1.6.2 Availability

This measure of system performance '*...denotes the probability that the system is available for use (in operable condition) at any arbitrary instant t'*. Availability is therefore the probability that, at the given time t, the system will be operational. It combines aspects of reliability, maintainability and maintenance support and implies that the system is either in active operation or is able to operate if required.

Availability pertains only to systems which undergo repair and are restored after failure, or to intermittently used systems. As such, it is eminently reasonable to introduce an *availability function* $A(t)$. In theory $A(0)$ should be 100%, but even equipment coming directly out of storage may be defective. A high availability can be obtained either by increasing the average operational time until the next failure, or by improving the

maintainability of the system. Gnedenko and Usnakov (1995) defines different coefficients of availability for one-unit systems.

1.6.2.1 Instantaneous or pointwise availability

This is a point function which describes the probability that a system will be able to operate at a given instant of time (Klaassen and Van Peppen (1989) and Beasley (1991)). In symbols:

$$A(t) = P[\phi(t) = 0].$$

1.6.2.2 Interval availability

Given an interval of time (and with given tolerances), interval availability is the expected fraction of this time that the system will be able to operate.

1.6.2.3 Average availability

If a failed unit is repaired and is then 'as good as new', the average availability is defined as

$$\text{Average Availability} = \frac{MTSF}{MTSF + MTSR}$$

where MTSF and MTSR are the Mean Time to System Failure and Mean Time to System Repair respectively.

1.6.2.4 Asymptotic or steady-state or limiting availability

The limiting availability, A_∞, is the expected fraction of time that the system operates satisfactorily in the long run (Barlow and Proschan (1965)): it is the probability that the system will be in an operational state at time t, when t is considered to be infinitely large

$$A_\infty = \lim_{t\to\infty} A(t).$$

1.6.3 Time to first disappointment

The system is said to be in a state of *disappointment* if the number of operable units at any time is less than the number of units required for the satisfactory performance of the system at that instant of time. For an intermittently used system, Gaver (1964) pointed out that a disappointment realizes in one of two possible ways: the system enters the down state during a need period, or a need for the system arises and at that time the system is in the down state. The event 'disappointment' is very useful as it renders the distribution of the time to the first disappointment, the mean number of disappointments over an arbitrary interval and also the mean duration of the disappointments.

1.6.4 Mean number of events in (0, *t*]

Let $N(a, t)$ denote the number of a particular type of a event (e.g. a disappointment, system recovery, system down, etc.) in (0, t]. The *mean number of events* in (0, t] is then given by

$$E[N(a,t)] = \frac{1}{t}\int_0^t m_1(u)du$$

where $h_1(u)$ is the first order product density of the events (product densities are defined in a subsequent section of this chapter).

The *mean stationary rate of occurrence* of these events is given by

$$E[N(a)] = \lim_{t\to\infty} \frac{\mathrm{E[N(a,\ t)]}}{\mathrm{t}}$$

1.6.5 Confidence limits for the steady state availability

A $100(1 - \alpha)\%$ *confidence interval* for A_∞ is defined by

$$P[a < A_\infty < b] = 1 - \alpha$$

where the numbers a and b ($a < b$) are determined using the appropriate statistical tables. It may be noted that A_∞ is a function of the parameters of operating time distribution, repair time, need and no-need period distributions, etc.

1.7 STOCHASTIC PROCESSES USED IN THE ANALYSIS OF REDUNDANT SYSTEMS

Previous sections briefly looked at different types of redundant systems and the various measures of system performance. In this section the techniques used in the analysis of redundant repairable systems will be summarized.

1.7.1 Renewal theory

In renewal theory there exists times, usually random, from which onward the future of the process is a probabilistic replica of the original process and interest is in the lifetime (a stochastic variable) of a unit. At time $t = 0$ a repairable unit is put into operation and is functioning. At each failure the unit is replaced by a new one of the same type, or subjected

to maintenance that completely restores it to an 'as good as new' condition. This process is repeated and replacement time is taken as negligible. The result is a sequence of lifetimes, and the study is restricted to these *renewal points*. The probability object in these sums of non-negative independently identically distributed random variables lies in the number of renewals N_t up to some time t.

Renewal processes are extensively used by many researchers to study specific reliability problems. The homogeneous Poisson process is the simplest renewal process and has received considerable attention. As in all other processes, the time parameter can be considered as either discrete or continuous. Feller (1950) gave a proper lead for the discrete and this was followed by the very lucid account of Cox (1962) for the continuous case (he provided an introduction to renewal theory in the case of a repair facility not being available and failed units queuing up for repair). Barlow (1962) applied queuing theory in his research on repairable systems. Srinivasan (1971) studied some operating characteristics of a one unit system, Gnedenko et al. (1969) obtained the mean life time to system failure of a two-unit standby system, Buzacott (1970) studied some priority redundant systems, etc.

Although renewals can take on different forms, the system starts a new cycle after each renewal (which is independent of the previous ones). If repair time is not negligible, each cycle consists of a lifetime and a repair time and both are random variables with individual distributions (repair time can also be considered as a fixed time). The process is called

- an *ordinary renewal process* if the time origin is the initial installation of the system and the repair time is considered negligibly small in comparison with the lifetime of the unit – renewal is taken as instantaneous, or

- *a general renewal process* if the time origin is some point subsequent to the initial installation of the system (Cox (1962)). Høyland and Rausand (1994) calls this a *modified* renewal process, while Feller (1957) refers to such a process considering the residual life time of a system at an arbitrary chosen time origin as a *delayed renewal* process.

1.7.1.1 Ordinary renewal process: instantaneous renewal

Consider a basic model of continuous operation where a unit begins operating at instant $t = 0$ and stays operational for a random time T_1 and then fails. At this instant it is replaced by a new and statistically identical unit, which operates for a length of time T_2, then fails and is again replaced etc. These random component life lengths T_1, T_2, …, T_r … of the identical units are independent, non-negative and identically distributed random variables that constitute a random flow or *ordinary renewal process*.

Let $P[T_i \leq t] = F(t)$; $t > 0$, $i = 1, 2, \ldots$ be the underlying distribution of the renewal process.

The time until the rth renewal is given by

$$t_r = T_1 + T_2 + \ldots + T_r = \sum_{i=1}^{r} T_i .$$

Let the random variable $N(t) = \max \{r; R_r \leq t\}$ indicate the number of times a renewal takes place in the interval $(0; t]$, then the number of renewals in an arbitrary time interval $(t_1, t_2]$ is equal to $N(t_2) - N(t_1)$.

A *renewal function* $H(t)$, which is the expected value of $N(t)$ in the time interval $(0; t]$, can be defined as

$$H(t) = E[N(t)] = \sum_{r=1}^{\infty} F^{(r)}(t)$$

where $F^{(r)}(\cdot)$ is the r-fold convolution of F.

Furthermore, (Cox (1962)),

$$H(t) = F(t) + \int_0^t H(t-x)dF(x)\,.$$

The renewal density function h(t) satisfies the equation

$$h(t) = \sum_{n=1}^{\infty} f^{(r)}(t)$$

and the renewal density function $h(t)$ satisfies the equation

$$h(t) = f(t) + \int_0^t h(t-x)f(x)dx\,.$$

Seeing that $h(t)\Delta t = P$ [exactly one renewal point in $(t, t+\Delta]$],

which implies that the renewal density $h(t)$ basically differs from the hazard rate $h^0(t)$, as

$$h^0(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{(1-F(t))}\,.$$

1.7.1.2 Random renewal time

Suppose the time for renewal is not instantaneous but considered as a random variable that is included in the consecutive time-periods, or cycles, of the systems' performance. Each cycle then consists of a *time to failure* and a *time to repair* and both are stochastic variables. Instants of failure and cycles of renewal can be identified.

Let $F(t)$ be the life distribution and $G(x)$ be the repair length function with respective probability density functions $f(t)$ and $g(x)$, then the density function of the cycles C of the life time and repair time, say $k(t)$ is obtained by the convolution formula

$$k(t) = \int_0^t f(x)g(t-x)dx\,.$$

If $N_F(t)$ counts the number of failures and $N_R(t)$ the number of repairs in (0; t], define

$$W(t) = E[N_F(t)]$$

and

$$V(t) = E[N_R(t)]$$

and let $Q(t) = W(t) - V(t)$; $\forall\, t$, assuming that $w(t) = W'(t)$ and $v(t) = V'(t)$.

The *failure and repair intensities* can be then respectively be defined as

$$\lambda(t) = \frac{w(t)}{A(t)}$$

where $A(t)$ is the availability function

$$\mu(t) = \frac{v(t)}{Q(t)}; \qquad Q(t) \neq 0.$$

1.7.1.3 Alternating renewal processes

Alternating renewal processes were first studied in detail by Takács (1957) and are discussed in many textbooks (Birolini (1994) and Ross (1970)). A generalization of the ordinary renewal process discussed previously where the state of the unit is given by the binary variable

$X(t) = \begin{cases} 1 & \text{if the system is functioning at time t} \\ 0 & \text{otherwise.} \end{cases}$

The two alternating states may be 'system up' and 'system down'. If these alternating independent renewal processes are distributed according to $F(x)$ and $G(x)$, there are two renewal processes embedded in them for the different transitions from 'system up' to 'system down'.

One-item repairable structures are generally described by alternating renewal processes with the assumption that after each repair the item is like new.

1.7.1.4 The age and remaining lifetime of a unit

In the notation of 1.7.1(a), let t_r indicate the random component life lengths, that is

$$t_r = \sum_{i=1}^{r} T_i .$$

Let R_r, $r \in N$, represent the length of the rth repair time, then the sequence $T_1, R_1, T_2, R_2, \ldots$ forms an alternating renewal process. Define

$$t_n = \sum_{r=1}^{n-1} (R_r + T_{r+1}); \qquad n \in \mathrm{N}$$

and set $t_0 = t_0^o = 0$.

This sequence t_n generates a delayed renewal process.

If $B_1(t)$ denotes the *forward recurrence time at time t*, then

$$B_1(t) = t_{N_t+1} - t \qquad \text{or} \qquad B_1(t) = t_{N_t^o+1} - t$$

Hence,

- $B_1(t)$ equals the *time to the next failure time* if the system is up at time t, or
- $B_1(t)$ equals the *time to complete the repair* if the system is down at time t.

Hence,

- $B_2(t)$ equals the *age* of the unit if the system is up at time t, or
- $B_2(t)$ equals the *duration of the repair* if the system is down at time t.

Returning to the renewal function $H(t)$, define the elementary renewal theorem (Feller (1949)), stating that, for an ordinary renewal process with underlying exponential distribution (parameter λ and $H(t) = \lambda t$)

$$\lim_{t\to\infty} \frac{\mathrm{H(t)}}{\mathrm{t}} = \frac{1}{\mu}$$

with $\mu = E(T_i) = 1/\lambda$, the mean lifetime.

If the renewals correspond to component failures, the mean number of failures in $(0, t]$ is approximately (for t large)

$$H(t) = E[N(t)] \approx \frac{1}{\mu} = \frac{1}{MTSF}.$$

1.7.2 Semi-Markov and Markov renewal processes

Consider a general description of a process where a system

- moves from one state to another with random sojourn times in between
- the successive states visited form a Markov chain
- the sojourn times have a distribution which depend on the present state as well as the next state to be entered.

This describes a Markov chain if all sojourn states are equal to one, a Markov process if the distribution of the sojourn times are all exponential and independent of the next state and a renewal process if there is only one state (then allowing an arbitrary distribution of the sojourn times).

Denote the state space by the set of non-negative integers {0, 1, 2 ...} and the transition probabilities by p_{ij}, $i, j = 0, 1, 2 \ldots$ If $F_{ij}(t)$, $t > 0$ is the conditional distribution function of the sojourn time in state i, given that the next transition will be into state j, let

$$Q_{ij}(t) = p_{ij} F_{ij}(t), \quad i, j = 0, 1, 2 \ldots$$

denote the probability that the process makes a transition into state j in an amount of time less than or equal to t, given that it just entered state i at $t = 0$. The functions $Q_{ij}(t)$ satisfy the following conditions

$$Q_{ij}(0) = 0, \qquad Q_{ij}(\infty) = p_{ij}$$

$$Q_{ij}(t) \geq 0, \qquad i, j = 0, 1, 2 \ldots$$

$$\sum_{j=0}^{\infty} Q_{ij}(t) = 1$$

Let J_0 and J_n respectively denote the initial state and the state after the nth transition occurred. The embedded Markov chain $\{J_n, n = 0, 1, 2 \ldots\}$ then describes a Markov chain with transition probabilities p_{ij}.

Let $N_i(t)$ denote the number of transitions into state i in $(0, t]$ and

$$N(t) = \sum_{i=0}^{\infty} N_i(t)$$

The stochastic process $\{X(t), t \geq 0\}$ with $X(t) = i$ denoting the process is in state i at time t is called a semi-Markov process (SMP) and it is clear that $X(t) = J_{N(t)}$. A SMP is a pure jump process and all states are regeneration states. The consecutive states form a time-homogeneous Markov chain, but it is a process without memory at the transition point from one state to the next.

The vector stochastic process $\{N_1(t), N_2(t) ...\}$ for $t \geq 0$ is called a Markov renewal process (MRP). This implies that the SMP records the state of the process at each time point, while the MRP is a counting process keeping track of the number of visits to each state.

Assuming that the time-intervals in which the random variables X(t) continues to remain in the n-point state are independently distributed such that

$$\lim_{t\to\infty} P[X(t + x) = j, X(t + u) = i: \forall u \leq x \,|\, X(t) = i, X(t - \Delta) \neq i]$$

$$= f_{ij}(x) ; i, j = 0, 1, 2 ...$$

If the transition of $X(t)$ is characterized by a change of state, then the quantities $f_{ii}(\cdot)$ are zero functions. Such a process which is a Markov chain with a randomly transformed time scale is called a MRP.

To remove the consequence that $f_{ii}(\cdot) = 0$, another function of a MRP can be given, namely defining it as a regenerative stochastic process $\{X(t)\}$ in which the epoch*s* at which $X(t)$ visits any member of a certain countable set of states are regeneration points; the visits being regenerative events.

In a combination of a Markov chain and a renewal process to form a SMP, the purpose is to create a tool that is more powerful than what either could provide individually. SMP were independently introduced by Lévy (1954) and Smith (1955). Detailed use of SMP and MRP can be found in Pyke (1961a, 1961b), Cinlar (1975) and Ross (1970). Barlow and Proschan (1965) used these processes to determine the MTSF in a two-unit system. Cinlar (1975), Osaki (1970a, 1970b), Arora (1976), Nakagawa & Osaki (1974, 1976) and Nakagawa (1974) have used the theory of SMP to discuss certain reliability problems.

1.7.3 Regenerative processes

In a regenerative stochastic process $X(t)$ there exists a sequence t_0, t_1, ... of stopping times such that $t = \{t_n;\ n \in N\}$ is a renewal process. If a point of regeneration happens at $t = t_1$, then the knowledge of the history of the process prior to t_1 loses its predictive value; the future of the process is totally independent of the past. Thus $X(t)$ regenerates itself repeatedly at these stopping times and the times between consecutive renewals are called *regeneration times*. The application of renewal theory to regenerative processes makes renewal theory such an important tool in elementary probability theory.

The *delayed renewal process* is defined as follows: if $\hat{t} = \{t_n - t_{0;} n \in N\}$ is a renewal process such that $t_0 \geq 0$ is independent of $\hat{t}$, (implying that the time t_0 of the first renewal is not necessarily the time origin) it is called a delayed renewal process. A delayed regenerative process is a process with a sequence $t = \{t_n; n \in N\}$ of stopping times which form a delayed renewal process. As an example: for any initial state i, the times of successive entrances to a fixed state j in a Markov process form a delayed renewal process.

In some cases non-exponentially distributed repair times and/or failure free operating times may lead to semi-Markov processes, but in general it leads to processes with only a few states (or even to non-regenerative processes). Recent research in this field is concerned with Brownian motion with the interest on the random set of all regeneration times and on the excursions of the process between generations.

1.7.4 Stochastic point processes

Among discrete stochastic processes, point processes are widely used in reliability theory to describe the appearance of events in time. A *renewal process* is a well known type of point process, used as a mathematical model to describe the flow of failures in time. It is a point process with restricted memory and each event is a regeneration point. In practical reliability problems, the interest is often in the behaviour of a renewal process in a stationary regime, i.e., when $t \rightarrow \infty$, as repairable systems enter an 'almost stationary' regime very quickly. A generalization of a renewal process is the so-called *alternating renewal process*, which consists of two types of independently identically distributed random variables alternating with each other in turn.

This theory of recurrent events has a huge variety of applications ranging from classical physics, biology, management sciences, cybernetics and many other areas. The result is that point processes have been defined differently by individuals in the different fields of application. The properties of stationary point processes were first studied by Wold (1948) and Bartlett (1954), to whom we owe the current terminology. Moyal (1962) gave a formal and well-knit theory of the subject that even provides an extension to cover non-Euclidean

spaces. Srinivasan (1974), Srinivasan and Subramanian (1980) and Finkelstein (1998, 1999c) extensively used point processes in reliability theory and applications.

Our interest in point processes lies in those applications which, in general, lead to the development of multivariate point processes. For this purpose we can define a point process as a stochastic process '*whose realizations are related to a series of point events occurring in a continuous one-dimensional parameter space (such as time, etc)*'. The sequence of times $\{t_n\}$ are the "renewal" epochs which generates the point process and the two random variables of interest are

- the number of points that fall in the interval $(t; t + x]$
- the time that has lapsed since the nth point after (or before) t.

The characterization property of *stationarity* applies to certain point processes, namely that the density function of observed events in a time interval does not depend on its position on the time axis, but only on the length of the interval. There are different types of stationarity that can be defined, namely simply stationary, weakly stationary and completely stationary (Srinivasan and Subramanian (1980)).

Furthermore, define $p(n; t, x) = P[N(t, x) = n]$ and if $\sum_{n\geq 2} P(n; t, t+\Delta) = o(\Delta)$ for small Δ, the point process is said to be *orderly* or *regular* (there are no multiple events, or clusters of events with probability one).

1.7.4.1 Multivariate point processes

Applications for multivariate stationary point processes can be found in many fields and the properties of these processes have been studied in depth by Cox and Lewis (1970).

If the constraint of independence of the intervals in a stationary renewal process is relaxed, a *stationary point process* is obtained; if the same constraint is removed in the case of a Markov renewal process a *multivariate point process* is obtained.

1.7.4.2 Product densities

Ramakrishnan (1954) developed, analyzed and perfected the *product density* technique as a sophisticated tool for the study of point processes. A point process is described by the triplet (Φ, **B,** P), where P is a probability distribution on some σ-field **B** of subsets of the space Φ of all states. Describe the state of a set of objects by a point x of a fixed set of points X. Assume for this discussion that X is the real number line. Define A_k as intervals and $N(.)$ as a counting measure which is uniquely associated with a sequence of points $\{t_i\}$ such that:

$N(A)$ = the number of points of the sequence $\{ t_i : t_i \in A\}$

$N(t, x)$ = the number of points (events) in the interval $(t; t + x]$

$N'(t, x)$ = the number of points (events) in $(t + x; t + x + \Delta]$.

The central quantity of interest in the product density technique is $N'(t, x)$, denoting the number of entities with parametric values between x and $x + \Delta$ at time t.

From the factorial moment distribution the product density of order , which represents the probability of an event in each of the intervals

$$(x_1, x_1 + \Delta_1), (x_2, x_2 + \Delta_2), ..., (x_n, x_n + \Delta_n),$$

can be defined. It is expressed as the product of the density of expectation measures at different points, namely

$$h_n(x_1, x_2, \ldots, x_n) = \lim_{\Delta_1, \Delta_2 \ldots, \Delta_n \to 0} \frac{E[\prod_{i=1}^{n} N(x_i, \Delta_i)]}{\Delta_1 \Delta_2 \ldots \Delta_n} \; ; \; x_1 \neq x_2 \neq \ldots \neq x_n$$

or equivalently

$$h_n(x_1, x_2, \ldots, x_n) = \lim_{\Delta_1, \Delta_2 \ldots, \Delta_n \to 0} \frac{P[N(x_i, \Delta_i) \geq 1, i = 1,2 \ldots, n]}{\Delta_1 \Delta_2 \ldots \Delta_n} \; ; \; x_1 \neq x_2 \neq \ldots \neq x_n$$

Since $h_n(...)$ is a product of the density of expectation measures at different points, the density is aptly called the *product density*.

Considering the *ordinary renewal process* as defined in 1.7.1(a), the renewal function $H(t)$ is the expected number of random points in the interval $(0; t]$. Modify the process by allocation of all integral values to $\{ t_i \}$ and consider a corresponding sequence of points on the real line. In the point process then generated by the random variables $\{ t_i \}$, the counting process $N(t, x)$ represents the number of points in the interval $(t, t + x]$ and the product density is

$$h_m(t, t_1, t_2, \ldots, t_m) = E[N'(t, t_1)\, N'(t, t_2) \ldots N'(t, t_m)]$$

The product density of degree m is

$$h_m(t, t_1, t_2, \ldots, t_m) = h_1(t, t_1)\, h(t_2 - t_1)\, h(t_3 - t_2) \ldots h(t_m - t_{m-1});$$

$$(t_1 < t_2 < \ldots < t_m).$$

1.8 SCOPE OF THE WORK

A stochastic model of an urea decomposition system in the fertilizer industry is studied in Chapter 2. A set of difference-differential equations for the state probabilities are formulated under suitable conditions. The state probabilities are obtained explicitly and the

steady state availability of the system is obtained analytically as well as illustrated numerically. Confidence limits for the steady state availability are also obtained.

In Chapter 3, a dissimilar unit system with different modes of failure is studied. The system is a priority system in which one of the units is a priority unit and the other unit one is an ordinary unit. The concept of 'dead time' is introduced with the assumption that the 'dead time' is an arbitrarily distributed random variable. The operating characteristics like MTSF, Expected up time, Expected down time, and the busy period analysis, as well as the cost benefit analysis is studied.

A two unit priority redundant system is studied in Chapter 4. The main aim of this chapter is to consider the physical conditions of the repair facility since the repair time distribution is affected by such conditions. Various system measures are studied, and the confidence limits for the availability and busy period are obtained in the steady state case.

In most of the available literature on *n*-unit standby systems, many of the associated distributions are taken to be exponential, one of the main reasons for this assumption is the number of built-in difficulties otherwise faced while analysing such systems. In Chapter 5 this exponential nature of the distributions is relaxed and a general model of a three unit cold standby redundant system, where the failure and repair time distributions are arbitrary, is studied.

In Chapter 6, a stochastic model of a reliability system which is operated by a human operator is studied. The system fails due to the failure of the human operator. Once again, it is assumed that the human operator can be in any one of the three states; namely, normal stress, moderate stress or extreme stress. Different operating characteristics like availability, mean number of visits to a particular state and the expected profit are obtained.

Results are illustrated numerically at the end of the chapters.

1.9 GENERAL NOTATION

$X(\cdot)$	A stochastic process describing the state of a system
p.d.f.	Probability density function
r. v.	Random variable
$f(\cdot)$	The p.d.f. of the lifetime of a unit while on-line
$g(\cdot)$	The p.d.f. of the repair time of a unit
©	Convolution symbol
$f^{(n)}(\cdot)$	n-fold convolution of a function $f(\cdot)$ with itself, where $f(\cdot)$ is arbitrary
$f^{*}(s)$	Laplace transform of the function $f(t)$
$F(t)$	Cumulative distribution function: $\int_0^t f(u)du$
$\overline{F}(t)$	Survivor function: $1 - F(t)$
E_i	Regenerative event of type i
A	Availability
$A_i(t)$	P(system is up at t / E_i at $t = 0$)
A_∞	Steady state availability
R	Reliability
$R_i(t)$	P(system is up in $(0, t]$ / E_i at $t = 0$)
MLE	Maximum likelihood estimator
MTSF	Mean time to system failure (also MTTF)

MTSR	Mean time to first appointment
SMP	Semi-Markov process
MRP	Markov renewal process