

CHAPTER 2

BASIC RELIABILITY AND AVAILABILITY CONCEPTS

2-1. Probability and statistics

This section provides the reader with an overview of the mathematics of reliability theory. It is not presented as a complete (or mathematically rigorous) discussion of probability theory and statistics but should give the reader a reasonable understanding of how reliability is calculated. Before beginning the discussion, a key point must be made. Reliability is a design characteristic indicating a system's ability to perform its mission over time without failure or to operate without logistics support. In the first case, a failure can be defined as any incident that prevents the mission from being accomplished; in the second case, a failure is any incident requiring unscheduled maintenance. Reliability is achieved through sound design, the proper application of parts, and an understanding of failure mechanisms. It is not achieved by estimating or calculating it. Estimation and calculation are, however, necessary to help determine feasibility, assess progress, and provide failure probabilities and frequencies to spares calculations and other analyses. With that in mind, let's first look at the theory of probability.

a. Uncertainty - at the heart of probability. The mathematics of reliability is based on probability theory. Probability theory, in turn, deals with uncertainty. The theory of probability had its origins in gambling.

(1) Simple examples of probability in gambling are the odds against rolling a six on a die, of drawing a deuce from a deck of 52 cards, or of having a tossed coin come up heads. In each case, probability can be thought of as the relative frequency with which an event will occur *in the long run*.

(a) When we assert that tossing an honest coin will result in heads (or tails) 50% of the time, we do not mean that we will necessarily toss five heads in 10 trials. We only mean that in the long run, we would expect to see 50% heads and 50% tails. Another way to look at this example is to imagine a very large number of coins being tossed simultaneously; again, we would expect 50% heads and 50% tails.

(b) When we have an honest die, we expect that the chance of rolling any possible outcome (1, 2, 3, 4, 5, or 6) is 1 in 6. Again, it is possible to roll a given number, say a 6, several times in a row. However, in a large number of rolls, we would expect to roll a 6 (or a 1, or a 2, or a 3, or a 4, or a 5) only 1/6 or 16.7% of the time.

(c) If we draw from an honest deck of 52 cards, the chance of drawing a specific card (an ace, for example) is not as easily calculated as rolling a 6 with a die or tossing a heads with a coin. We must first recognize that there are 4 suits, each with a deuce through ace (ace being high). Therefore, there are four deuces, four tens, four kings, etc. So, if asked to draw an ace, we know that there are four aces and so the chance of drawing any ace is 4 in 52. We instinctively know that the chance of drawing the ace of spades, for example, is less than 4 in 52. Indeed, it is 1 in 52 (only one ace of spades in a deck of 52 cards).

(2) Why is there a 50% chance of tossing a head on a given toss of a coin? It is because there are two results, or events, which can occur (assume that it is very unlikely for the coin to land on its edge) and for a balanced, honest coin, there is no reason for either event to be favored. Thus, we say the outcome is random and each event is equally likely to occur. Hence, the probability of tossing a head (or

tail) is one of two equally probable events occurring = $1/2 = 0.5$. On the other hand, one of six equally probable events can result from rolling a die: we can roll a one, two, three, four, five, or six. The result of any roll of a die (or of a toss of a coin) is called a discrete random variable. The probability that on any roll this random variable will assume a certain value, call it x , can be written as a function, $f(x)$. We refer to the probabilities $f(x)$, specified for all values of x , as values of the probability function of x . For the die and coin, the function is constant. For the coin, the function is $f(x) = 0.5$, where x is either a head or tail. For the die, $f(x) = 1/6$, where x can be any of the six values on a die.

b. Probability functions. All random events have either an underlying probability function (for discrete random variables) or an underlying probability density function (for a continuous random variable).

(1) The results of a toss of a coin or roll of a die are discrete random variables because only a finite number of outcomes are possible; hence these events have an underlying probability function. When the probability of each event is equal, underlying probability function is said to be uniform.

(2) The number of possible heights for American males is infinite (between 5' - 8" and 6', for example, there are an infinite number of possible heights) and is an example of a continuous random variable. The familiar bell-shaped curve describes most natural events, such as the height of a person, intelligence quotient of a person, errors of measurement, etc. The underlying probability density function represented by the bell-shaped curve is called normal or Gaussian. Figure 2-1 shows a typical normal distribution. Note that the event corresponding to the midpoint of the curve is called the mean value. The mean value, also called the expected value, is an important property of a distribution. It is similar to an average and can be compared with the center of mass of an object. For the normal distribution, half the events lie below the mean value and half above. Thus, if the mean height of a sample of 100 Americans is 5' -9", we would expect that half the sample would be less than 69" inches tall and half would be taller. We would also expect that most people would be close to the average with only a few at the extremes (very short or very tall). In other words, the probability of a certain height decreases at each extreme and is "weighted" toward the center, hence, the shape of the curve for the normal distribution is bell-shaped.

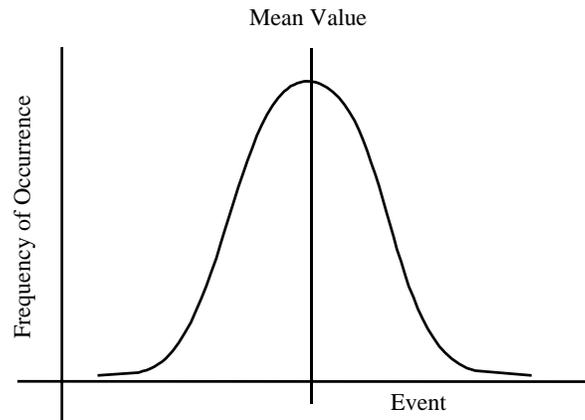


Figure 2-1. Typical normal distribution curve.

(3) The probability of an event can be absolutely certain (the probability of tossing either a head or a tail with an honest coin), absolutely impossible (the probability of throwing a seven with one die), or somewhere in between. Thus, a probability always can be described with equation 2-1.

$$0 \leq \text{Probability} \leq 1 \qquad \text{(Equation 2-1)}$$

(4) Determining which distribution best describes the pattern of failures for an item is extremely important, since the choice of distributions greatly affects the calculated value of reliability. Two of the continuous distributions commonly used in reliability are shown in table 2-1. Note that $f(t)$ is called the probability density function. It is also referred to as the pdf. For reliability, we are usually concerned with the probability of an unwelcome event (failure) occurring.

Table 2-1. Commonly used continuous distributions

Distribution	Probability Density Function	Most Applicable to
Exponential	$f(t) = \lambda \exp(-\lambda t)$	Electronic parts and complex systems
Weibull (2-parameter)	$f(t) = \frac{\beta}{\theta^\beta} t^{\beta-1} \exp\left[-\left(\frac{t}{\theta}\right)^\beta\right]$	Mechanical parts

(a) The underlying statistical distribution of the time to failure for parts is often assumed to be exponential. A glance at the equation of the probability density function explains why. It is easy to work with and has a constant mean, λ . Rather than assuming a distribution, one should determine the most appropriate one using various techniques discussed in chapter 3 for analyzing time-to-failure data.

(b) When the exponential distribution is applicable, the rate at which failures occur is constant and equal to λ . For other distributions, the rate at which failures occur varies with time. For these distributions, we cannot talk of a failure rate. Instead, we use the term Hazard Function, which is a function that describes how the rate of failures varies over time.

(c) Note that different types of parts (i.e., items that fail once and then are discarded and replaced with a new item) may have different underlying statistical distributions of the time to failure. The times to failure of electronic parts, for example, often follow the exponential distribution. The times to failure for mechanical parts, such as gears and bearings, often follow the Weibull distribution. Of course, the parameters for the Weibull for a gear most likely will be different from the parameters for a ball bearing. The applicability of a given distribution to a given part type and the parameters of that distribution are determined, in part, by the modes of failure for the part.

(d) By their very nature, systems consist of many, sometimes thousands, of parts. Since systems, unlike parts, are repairable, they may have some parts that very old, some that are new, and many with ages in between these extremes. In addition, each part type will have a specific distribution of times to failure associated with it. The consequence of these part characteristics together within a system is that systems tend to exhibit a constant failure rate. That is, the underlying statistical distribution of the time to failure for most systems is exponential. This consequence is extremely significant because many reliability prediction models, statistical demonstration tests, and other system analysis are predicated on the exponential distribution.

c. Determining failure rate or Hazard Function. How do we determine the failure rate (or Hazard Function) of a specific system or component? Two methods are used.

(1) In the first method, we use failure data for a comparable system or component already in use. This method assumes that the system in use is comparable to the new system and that the principle of transferability applies - this principle states that failure data from one system can be used to predict the reliability of a comparable system.

(2) The other method of determining failure rate or the Hazard Function is through testing of the system or its components. Although, theoretically, this method should be the "best" one, it has two disadvantages. First, predictions are needed long before prototypes or pre-production versions of the system are available for testing. Second, the reliability of some components is so high that the cost of testing to measure the reliability in a statistically valid manner would be prohibitive. Usually, failure data from comparable systems are used in the early development phases of a new system and supplemented with test data when available.

2-2. Calculating reliability

If the time, t , over which a system must operate and the underlying distributions of failures for its constituent elements are known, then the system reliability can be calculated by taking the integral (essentially the area under the curve defined by the pdf) of the pdf from t to infinity, as shown in equation 2-2.

$$R(t) = \int_t^{\infty} f(t) dt \tag{Equation 2-2}$$

a. Exponential distribution. If the underlying failure distribution is exponential, equation 2-2 becomes equation 2-3.

$$R(t) = e^{-\lambda t} \tag{Equation 2-3}$$

where:

- λ is the failure rate (inverse of MTBF)
- t is the length of time the system must function
- e is the base of natural logarithms
- $R(t)$ is reliability over time t

(1) Figure 2-2 shows the curve of equation 2-3. The mean is not the "50-50" point, as was true for the normal distribution. Instead, it is approximately the 37-63 point. In other words, if the mean time between failures of a type of equipment is 100 hours, we expect only 37% (if $t = \text{MTBF} = 1/\lambda$, then $e^{-\lambda t} = e^{-1} = 0.367879$) of the population of equipment to still be operating after 100 hours of operation. Put another way, when the time of operation equals the MTBF, the reliability is 37%.

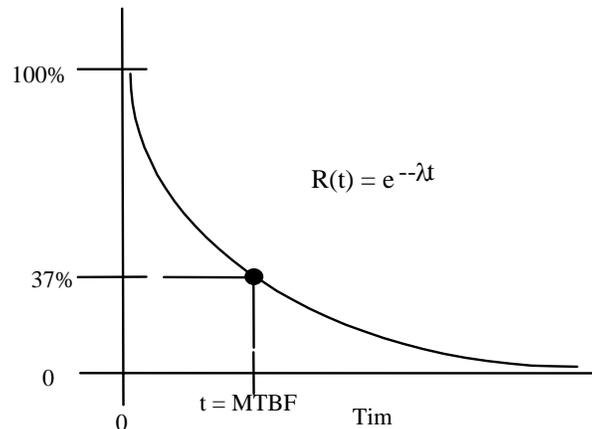


Figure 2-2. Exponential curve relating reliability and time.

(2) If the underlying distribution for each element is exponential and the failure rates, λ_i , for each element are known, then the reliability of the system can be calculated using equation 2-3.

b. *Series Reliability.* Consider the system represented by the reliability block diagram (RBD) in figure 2-3

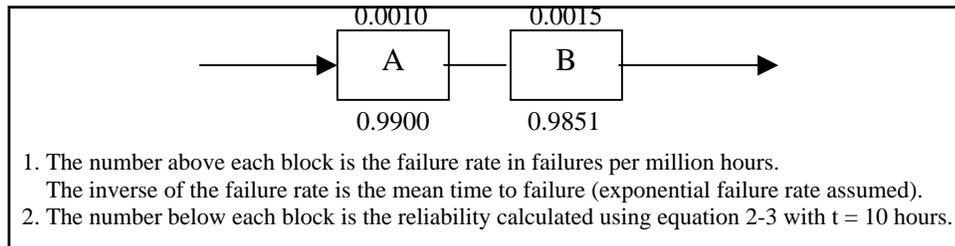


Figure 2-3. Example reliability block diagram.

(1) Components A and B in figure 2-3 are said to be in series, which means all must operate for the system to operate. Since the system can be no more reliable than the least reliable component, this configuration is often referred to as the weakest link configuration. An analogy would be a chain; the strength of the chain is determined by its weakest link.

(2) Since the components are in series, the system reliability can be found by adding together the failure rates of the components and substituting the result in equation 2-4. The system failure rate is $0.001000 + 0.001500 = 0.002500$. The reliability is:

$$R(t) = e^{-0.0025 \times 10} = 0.9753 \quad \text{(Equation 2-4)}$$

(3) Alternatively, we could find the system reliability by multiplying the reliabilities of the two components as follows: $0.9900 \times 0.9851 = 0.9753$.

c. *Reliability with Redundancy.* Now consider the RBD shown in figure 2-4.

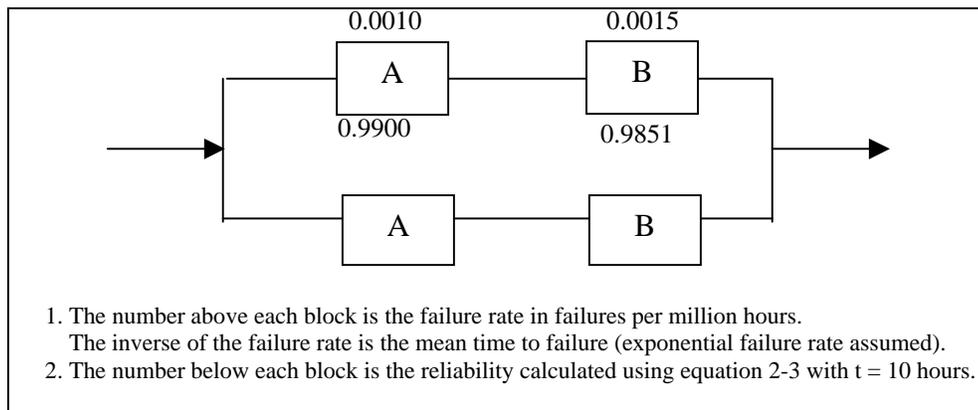


Figure 2-4. RBD of a system with redundant components.

(1) The system represented by the RBD in figure 2-4 has the same components (A and B) used in figure 2-3, but two of each component are used in a configuration referred to as redundant or parallel.

Two paths of operation are possible. The paths are: top A-B and bottom A-B. If either of two paths is intact, the system can operate. The reliability of the system is most easily calculated by finding the probability of failure ($1 - R(t)$) for each path, multiplying the probabilities of failure (which gives the probability of both paths failing), and then subtracting the result from 1. The reliability of each path was found in the previous example. Next, the probability of a path failing is found by subtracting its reliability from 1. Thus, the probability of either path failing is $1 - 0.9753 = 0.0247$. The probability that both paths will fail is $0.0247 \times 0.0247 = 0.0006$. Finally, the reliability of the system is $1 - 0.0006 = 0.9994$, about a 2.5% improvement over the series-configured system.

(2) Two components in parallel (redundant) may always be on and in operation (active redundancy) or one may be off or not in the "circuit" (standby redundancy). In the latter case, failure of the primary component must be sensed and the standby component turned on or switched into the circuit. Standby redundancy may be necessary to avoid interference between the redundant components and, if the redundant component is normally off, reduces the time over which the redundant component will be used (it's only used from the time when the primary component fails to the end of the mission). Of course, more than two components can be in parallel. Chapter 3 (3-1) discusses the various types of redundancy and how it can be used to improve the availability of current C4ISR facilities.

(3) Adding a component in parallel, i.e., redundancy, improves the system's ability to perform its function. This aspect of reliability is called functional or mission reliability. Note, however, that in figure 2-4, we have added another set of components that has its own failure rate. If we want to calculate the total failure rate for all components, we add them. The result is 5000 failures per million operating hours (0.005000). The failure rate for the series-configured system in figure 2-3 was 2500 failures per million operating hours. Although the functional reliability of the system improved, the total failure rate for all components **increased**. This perspective of reliability is called basic or logistics reliability. When standby redundancy is used, the sensing and switching components add to the total failure rate.

d. Logistics reliability. Whereas functional reliability only considers failures of the function(s), logistics reliability considers all failures *because some maintenance action will be required*. Logistics reliability can be considered as either the lack of demand placed on the logistics system by failures or the ability to operate without logistics. If standby redundancy is used with the redundant component not on, the apparent failure rate of that component will be less than that of its counterpart (because the probability it will be used is less than 1 and the time it will operate less than 10 hours), but the failure rate of the switching circuits must now be considered.

2-3. Availability

For a system such as an electrical power facility, availability is a key measure of performance. An electrical power facility must operate for very long periods of time, providing power to other systems, such as C4ISR, that perform critical functions. Even with the best technology and most robust design, it is economically impractical, if not technically impossible, to design power facilities that never fail over weeks or months of operation. Although forced outages (FAs) are never welcome and power facilities are designed to minimize the number of FAs, they still occur. When they do, restoring the system to operation as quickly and economically as possible is paramount. The maintainability characteristics of the system limit how quickly and economically system operation can be restored.

a. Reliability, maintainability, and availability. Consequently, reliability and maintainability (R&M) are considered complementary characteristics. Looking at a graph of constant curves of inherent availability (A_i), one can see this complementary relationship. A_i is defined by the following equation and reflects the percent of time a system would be available if delays due to maintenance, supply, etc. are ignored.

$$A_i = \frac{MTBF}{MTBF + MTTR} \times 100\% \quad \text{(Equation 2-6)}$$

where MTBF is mean time between failure and MTTR is mean time to repair

There are R&M trades. If the system never failed, the MTBF would be infinite and A_i would be 100%. Or, if it took no time at all to repair the system, MTTR would be zero and again the availability would be 100%. Figure 2-5 is a graph showing availability as a function of reliability and maintainability (availability is calculated using equation 1). Note that you can achieve the same availability with different values of R&M. With higher reliability (MTBF), lower levels of maintainability are needed to achieve the same availability and vice versa. It is very common to limit MTBF, MTTR, or both. For example, the availability requirement might be 95% with an MTBF of at least 600 hours and a MTTR of no more than 3.5 hours.

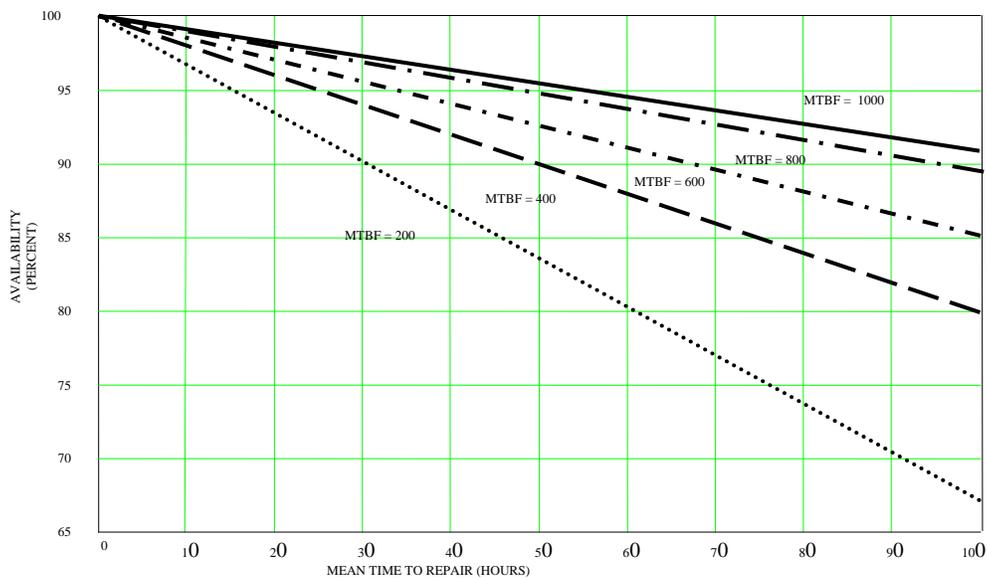


Figure 2-5. Different combinations of MTBF and MTTR yield the same availability.

b. *Other measures of availability.* Other measures of availability include operational availability, A_o , and measured availability, A .

(1) Operational availability includes maintenance and logistics delays and is defined using equation 2-7:

$$A_0 = \frac{MTBM}{MTBM + MDT} \quad \text{(Equation 2-7)}$$

where MTBM is the mean time between all maintenance and MDT is the mean downtime for each maintenance action.

(2) Measured availability is defined in equation 2-8.

$$A = \frac{\text{Uptime}}{\text{Uptime} + \text{Downtime} = \text{Total Time}} \quad (\text{Equation 2-8})$$

where Uptime is the time during which the system is available for use and Downtime is the time during which the system is not available for use.

(3) Note that A_o and A_i are probabilistic measures, while A is a deterministic measure. MTBF and MTBM and MTTR and MDT are measures of reliability and maintainability (R&M), respectively, and are random variables. By designing for appropriate levels of R&M and conducting adequate statistically based tests, a high confidence in the availability can be obtained. That confidence can never be 100%. Measuring A is done by actually measuring the amount of time in a given time interval during which the system is "up" and then calculating the observed availability. For this measure of availability, the time interval for the measurement is extremely important. Its importance can be understood by considering an availability requirement of 95% with a maximum downtime of 10 hours. Table 2-2 shows the effect of varying intervals of time for measuring A .

Table 2-2. Effect of measurement interval on observed availability

Time Interval	Actual Downtime	Measured Availability	Maximum Downtime to Meet Requirement
1 hour	0.5 hour	50%	0.05 hour (3 minutes)
8 hours	1 hour	87.5%	0.4 hour (24 minutes)
24 hours	2 hours	91.67%	1.2 hours
240 hours	10 hours	95.83%	10 hours
7200 hours	10 hours	99.86%	10 hours

(a) Very short intervals make it increasingly difficult, if not impossible, to meet an availability requirement. It is very possible that a failure could occur in the first 8 hours of operation. If that were the case, the system would pass the 95% availability test only if the repair could be made in 3 minutes or less. For many systems, it may be impossible to correct any failure in 3 minutes or less. So even if it is unlikely that a failure will occur in the first hour of operation (i.e., the system is highly reliable), the probability of such a failure is not zero. If a failure occurs in the first hour and requires more than 3 minutes to repair, the system will have failed to meet an availability requirement of 95%. Yet, if the system is truly reliable, it may experience no more failures (and no more downtime) in the next 24 hours of operation, in which case the measured availability will be greater than the requirement.

(b) Since A_o , A_i , and A are not measured in the same way, it is extremely important in contracts to state, a priori, (e.g., in a step-by-step, deductive manner) how availability will be measured during acceptance or qualification testing.

2-4. Predictions and assessments

Predictions and assessments refer to the process of evaluating the reliability of a system, its weaknesses, and areas offering opportunities for improvement. Quantitative numbers are a usual byproduct of a prediction or assessment, and such numbers are necessary for calculating spares requirements, probability of success, and for other purposes. However, another very important result of a prediction or assessment is in identifying ways to improve the system.

a. *Reliability Predictions.* In a new development program, reliability predictions are a means of determining the feasibility of requirements, assessing progress toward achieving those requirements and

comparing the reliability impact of design alternatives. Predictions can be made through any appropriate combination of reliability models, historical data, test data, and engineering judgment. The choice of which prediction method to use depends on the availability of information, which in turn is a function of the point of the system life cycle at which the prediction is performed. Considerations in performing predictions are that correct environmental stresses are used, the reliability model is correct, the correct part qualities are assumed and that all operational and dormancy modes are reflected. Chapter 3 addresses the types of models commonly used.

b. Reliability Assessment. Predictions are one method of assessing the reliability of an item. At the onset of a new development program, the prediction is usually purely analytical. As the program progresses, other methods become available to improve or augment the analytical prediction. These methods include testing, design reviews, and other methods. For existing systems, reliability assessments include analyzing field data to determine the level of reliability being achieved and identify weaknesses in the design (i.e., opportunities for improvement).

(1) Table 2-3 lists some common techniques that can be used for assessing reliability and guidance for their use. Methods especially useful for existing systems are shown in bold. Some of these methods provide a numerical value that is representative of the system reliability at a point in time; all provide a valuable means of better understanding the design's strengths and weaknesses so that it can be changed accordingly.

(2) The assessment methods chosen should be appropriate for the system and require only a reasonable level of investment given the value of the results. The failure of some components, for example, may have little impact on either system function, or on its operating and repair costs. A relatively costly analysis may not be justified. For other systems, a thermal analysis may not be needed, given the nature of the system and its operating environment. When the consequences of failure are catastrophic, every possible effort should be made to make the system fail-safe or fault tolerant.

Table 2-3. Methods for assessing reliability

Method	Application
Accelerated Life Testing	Effective on parts, components or assemblies to identify failure mechanisms and life limiting critical components.
Critical Item Control	Apply when safety margins, process procedures and new technology present risk to the production of the system.
Design of Experiments (DOE)	Use when process physical properties are known and parameter interactions are understood. Usually done in early design phases, it can assess the progress made in improving system or process reliability.
Design Reviews	Continuing evaluation process to ensure details are not overlooked. Should include hardware and software.
Dormancy Analysis	Use for products that have "extended" periods of non-operating time or unusual non-operating environmental conditions or high cycle on and off periods.
Durability Analysis	Use to determine cycles to failure or determine wearout characteristics. Especially important for mechanical products.
Failure Modes, Effects and Criticality Analysis (FMECA)	Applicable to equipment performing critical functions (e.g., control systems) when the need to know consequences of lower level failures is important.
Failure Reporting Analysis and Corrective Action (FRACAS)	Use when iterative tests or demonstrations are conducted on breadboard, or prototype products to identify mechanisms and trends for corrective action. Use for existing systems to monitor performance.
Fault Tree Analysis (FTA)	Use for complex systems evaluation of safety and system reliability. Apply when the need to know what caused a hypothesized catastrophic event is important.
Finite Element Analysis (FEA)	Use for designs that are unproven with little prior experience/test data, use advanced/unique packaging/design concepts, or will encounter severe environmental loads.
Life Cycle Planning	Use if life limiting materials, parts or components are identified and not controlled.
Parts Obsolescence	Use to determine need for and risks of application of specific parts and lifetime buys
Prediction	Use as a general means to develop goals, choose design approaches, select components, and evaluate stresses. Equally useful when redesigning or adding redundancy to an existing system.
Reliability Growth Test (RGT)/Test Analyze and Fix (TAAF)	Use when technology or risk of failure is critical to the success of the system. These tests are costly in comparison to alternative analytical techniques.
Sneak Circuit Analysis (SCA)	Apply to operating and safety critical functions. Important for space systems and others of extreme complexity. May be costly to apply.
Supplier Control	Apply when high volume or new technologies for parts, materials or components are expected
Test Strategy	Use when critical technologies result in high risks of failure.
Thermal Analysis (TA)	Use for products with high power dissipation, or thermally sensitive aspects of design. Typical for modern electronics, especially of densely packaged products.
Worst Case Circuit Analysis (WCCA)	Use when the need exists to determine critical component parameters variation and environmental effects on circuit performance.