1

INTRODUCTION

Alvin W. Strong

1.1 BOOK PHILOSOPHY

This CMOS technology reliability book has been written at a beginning graduate level or senior undergraduate level and assumes some solid state physics background.

The book is divided into seven relatively independent chapters consisting of an introduction, gate dielectric characterization, gate dielectric physics and breakdown, negative bias temperature instability or just NBTI reliability, hot carrier injection or hot electron reliability, stress-induced voiding or stress migration reliability, and electromigration reliability. The chapters describe the reliability mechanisms and the physics associated with them. They then take that understanding as the framework to build the bridge between the accelerated mechanism and the product mechanism.

For a CMOS reliability course or understanding focused only on one of the mechanisms, the authors expect that the material covered would include most of the first chapter and that focus chapter.

Several mechanisms are occasionally considered with reliability mechanisms, but these are not included here. Examples of these include latch-up [1], electrostatic discharge (ESD) [2, 3], and the radiation-induced soft-error rate (SER) [4].

Reliability Wearout Mechanisms in Advanced CMOS Technologies. By Strong, Wu, Vollertsen, Suñé, LaRosa, Rauch, and Sullivan

1.2 LIFETIME AND ACCELERATION CONCEPTS

It is a fact of life that every human-devised system has a finite lifetime before the catastrophic failure of the system occurs. However, most systems have a reasonably well-defined lifetime, and the catastrophic failure, or wearout, occurs well past that expected lifetime. That system has met our expectation and the customer is satisfied. Wearout is best thought of in terms of all of the systems or subsystems failing within one or two orders of magnitude in time. For example, a computer system with an expected lifetime of 10 years should experience no significant wearout before 10 years. However, all of the systems could be expected to wear out sometime between 20-plus years and 200-plus years.

1.2.1 Reliability Purpose

The purpose then of reliability is to ensure that the life of the system will be longer than the target life and that the failure rate during the normal operating life of the system will be below the target failure rate. The reliability of the product must be known when the product is sold so that the operating-life warranty costs can be quantified and customer satisfaction protected. Ensuring these objectives are met means that each failure mechanism must be quantified so that its impact during normal operating life can be predicted and the time at which it starts to cause the system to wear out can be predicted as well.

The length of time one has to do the reliability stressing and make the predictions is dependent on the state of the program. As a new technology is being developed, reliability engineers should be generating reliability data to help guide the program in the appropriate design, cost, and reliability tradeoffs. This work may occur over the course of several months to a few years. However, feedback on any given experiment needs to be given as quickly as possible. Once the technology is ready for implementation, it would typically undergo a ''qualification'' of no more than three months in duration. If a problem is discovered after qualification, that is, during manufacturing, it is all the more crucial to give feedback quickly.

The concept of an accelerated life is necessary for reliability stressing to have meaning. That is, it must be possible to find some condition or conditions that will allow one to shrink a 10-year product life down to a three month period, or less, so that the reliability of the system can be investigated and guaranteed in that three months. The conditions used to accelerate a given mechanism usually cannot be applied to the whole system (in our case, the semiconductor product chip). In this case, a test structure must typically be built that will replicate the behavior of the element in the product chip, but allow one to apply an accelerating condition. Hence, with the concept of accelerated life, we also need to posit the concept of a representative test structure.

It must be noted that in all of the above discussion, the product is assumed to operate perfectly when it is first turned on, for example, at time zero.

Once the reliability of each element has been investigated, understood, and modeled, an additional step should be feedback for the next design pass so that the

product team can design in reliability. A simple example of this design for reliability would be to use minimal groundrules only where necessary.

1.2.2 Accelerated Life

An accelerated life concept must include several features to be useful. In addition to the requirement that it can be used to accelerate a particular reliability mechanism, it must be possible to quantify how much that condition actually accelerates the reliability mechanism. It must be possible to build a bridge between the accelerated stress conditions and the use condition so that it is possible to quantify the degree or amount of acceleration. This quantification is also necessary to ensure that no mechanism is introduced with the accelerating condition that does not exist at the use condition of the product. One must understand if the behavior of the mechanism is uniform and consistent from the use condition to the accelerating conditions.

Once an appropriate accelerating condition has been determined, the acceleration between the stress conditions and the use condition can be determined. First, data from at least two different values of accelerating conditions are measured. The cumulative fails from those conditions are then plotted on the y axis, versus time on the x axis. This plot is done on a set of axes that have been transformed in such a way that the resulting plot is a straight line. The methodology for doing this for the most common distributions used in semiconductor reliability is discussed in detail later in this chapter. The two distributions that are most commonly used in semiconductor technology reliability are the two-parameter lognormal and Weibull distributions because each of these is very flexible and can be used to describe many different types of behaviors. These distributions provide a functional form with which a distribution can be characterized so that it can then be treated analytically. The details of these distributions, and their axes transformations, are the topics of Section 1.4. A simple, although somewhat unrealistic, example would be a distribution whose cumulative failures were linear with the log of time. If the cumulative fails were to be plotted against time, a nonlinear curve would result. However, a transformation of the x axis by taking the log of the time and then plotting the cumulative failures against that log of time would result in a straight line. These transformations are necessary so that the distributions and the slopes remain invariant across all accelerating conditions and down to the use conditions. All transformations have been made for the example in Figure 1.1 so that the plot is linear. Two sets of voltage data are shown plotted on the top left of Figure 1.1. At least three different values of each accelerating condition are preferred but only two are shown on Figure 1.1 for simplicity. The data from these accelerated conditions are used to calculate an acceleration factor, which is in turn used to calculate the acceleration time between the lowest accelerated condition and the use condition.

One uses the lowest accelerated condition to minimize projection error. Obviously if any new mechanism is introduced due to the accelerated condition, or a nonlinearity in the expected mechanism is introduced, this acceleration time

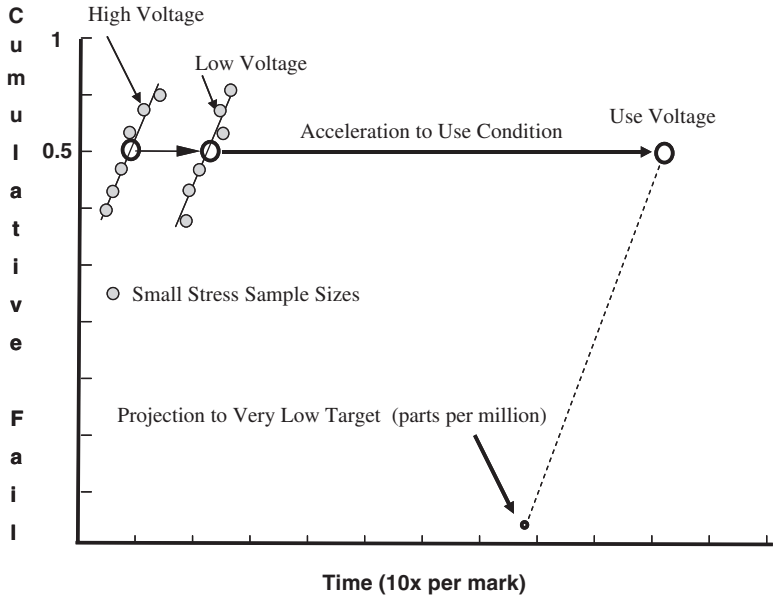

Figure 1.1. Acceleration example with two voltage conditions with the first projection to use condition at a 50% life and the final projection to low (ppm) failure rate target.

would not be valid. The transformation equations for the x and y axes of the plot will depend on the particular reliability mechanism in question and the probability distribution function that is most appropriate for that mechanism. The sample sizes for the stresses are typically very small, whereas the wearout target is typically expressed in terms of parts per million (ppm) or less. This means that once the acceleration between the stress conditions and the use condition is calculated, a second projection must then be made from that value, which typically is at about the 50% fallout point of each of the accelerated curves, to a very small percent fallout for that second projection curve. Note that the 50% value is used only as a convenient example here. The exact value at which the acceleration calculations will be made will depend on the distribution that is to be used and is discussed in detail later in this chapter. The intent here is to give a broad overview and avoid losing the reader in the detail. This extrapolation is done using the same slope found during at the stress conditions, since the axes have been transformed so that they remain invariant across all of the conditions of interest. Thus, an error in the acceleration factor causes the use condition to be incorrectly located in time, and any errors in determining the correct slope causes the projection to the small fraction fail target to have an additional error. Often this last error, the error due to an incorrect slope determination, can cause the largest error in the resultant projection. It should be highlighted that we are not speaking of graphing errors here since the calculations can all be performed using

computer software. If done graphically, those errors would be in addition to the errors mentioned previously.

Three plots are shown in Figure 1.1. The two plots on the left show the two different accelerated voltage conditions with all other conditions held constant. The dotted line plot on the right shows the projection from the 50% failure time for the use condition to the failure time associated with the target failure rate for that mechanism. One other potential factor that is not shown in Figure 1.1 is a test-structure scaling factor. This factor will be discussed in each of the chapters for which it is applicable. Although there are many accelerating conditions as shown in Section 1.2.3, by far the two most common accelerating conditions are voltage and temperature. The minimum experimental design that can yield a voltage acceleration factor is the two conditions as shown in Figure 1.1. For example, assume that the stress voltages in Figure 1.1 are 4.37 V and 4 V. The lines on the left side of Figure 1.1 represent fits to data taken at these accelerated voltage conditions. The slopes of the two stress conditions are shown as equal. The slope fit would normally be accomplished by a fitting program, which could force the best fit to all of the accelerated curves simultaneously. For this case we will assume an Eyring acceleration model [5] applies that has the form $A\text{cc} = t_2/t_1 = \exp\{(\Delta H/k)\{(1/T_2) - (1/T_1)\}\} \exp\{-\beta_V (V_2 - V_1)\}$. For the acceleration due to voltage, $A\text{cc}_{\text{VOLTstress}} = t_2/t_1 = \exp\{-\beta_V(V_2 - V_1)\}$, where V is voltage, t is time, and β_V is the voltage acceleration factor for this Eyring model. The temperature acceleration model by itself has the form $A\text{cc}_{\text{TEMPstress}} = t_2/t_1 = \exp\{(\Delta H/k)\{(1/T_2) - (1/T_1)\}\}$ where k is Boltzmann's constant and is also known as an Arrhenius model. These models will be discussed in more detail throughout this book but are introduced here to give the reader an early qualitative introduction to the acceleration concepts. Observation of the first two curves will reveal a time difference or acceleration of about 30 ×. The large circles in Figure 1.1 represent a mean life of the hardware under stress and as mentioned above are used as a convenient example. As will be discussed later, the points at which the acceleration calculations will be made are the most accurate values for the distribution under consideration. If the voltage used for the first curve on the left is $V = 4.37$ and the voltage for the second curve is $V = 4$, then β_V may be calculated, given $A\text{cc}_{\text{VOLTstress}} = 30$, as $\beta_V = (\ln A\text{cc}_{\text{VOLTstress}})/(V_1 - V_2) = 9.2$. This value for β_V is then used to project from $V = 4$ to the $V = 2$ use condition as $A\text{cc}_{\text{VOLTuse}} = \exp\{9.2 \times (4-2)\} = 10^8$. Having made this calculation, one then needs to consider whether or not the value calculated is reasonable based on comparable data both from the reliability analyst's prior work, as well as literature values. A similar procedure would be used to calculate any acceleration including, for example, a temperature acceleration. This example should give the reader a better understanding of the actual process of stressing and then projecting to use conditions using acceleration concepts. Obviously the stress conditions must be appropriately chosen and the experiment appropriately designed to achieve useful results. Note that if too small of difference is used between two accelerating conditions then the experimental error and the statistical variation in the two sets of data may cause enough overlap of data such that the acceleration factor between the two sets of data cannot be calculated. On the other hand if the difference between

the two sets of data is too large, the failure times of the lower condition may be longer than the time designated for the stress. The discussion of the extrapolation to very small failing percentage targets will commence in Section 1.4.5.

We now return to a more general discussion of acceleration. The mean life from Figure 1.1 is plotted against one of the accelerating conditions in Figure 1.2. Figure 1.2 presents a picture of the progress of the state of the art of reliability stressing across the last 20-plus years. Each circle represents a change of approximately 40 $\times$ in time.

More than 20 years ago, all reliability stressing was done with the reliability test structures wire-bonded onto a die carrier or module. This structure, which was contained within the package, was then put into a stress apparatus, which typically applied stress temperature and voltage between weeks and months, depending on the mechanism under investigation. The readouts were made at preset values, typically on the order of two or three times per decade. The test structures had to be removed from the stress apparatus and physically transported to a tester for each readout. This stressing is represented in Figure 1.2 by the second circle from the left. Note that the left most circle represents the useful life of the structure, typically 10 years. For mechanisms like ionic contamination, which will relax unless the voltage is continuously applied, it was necessary to have large batteries connected to keep the hardware at stress voltage while transporting the hardware

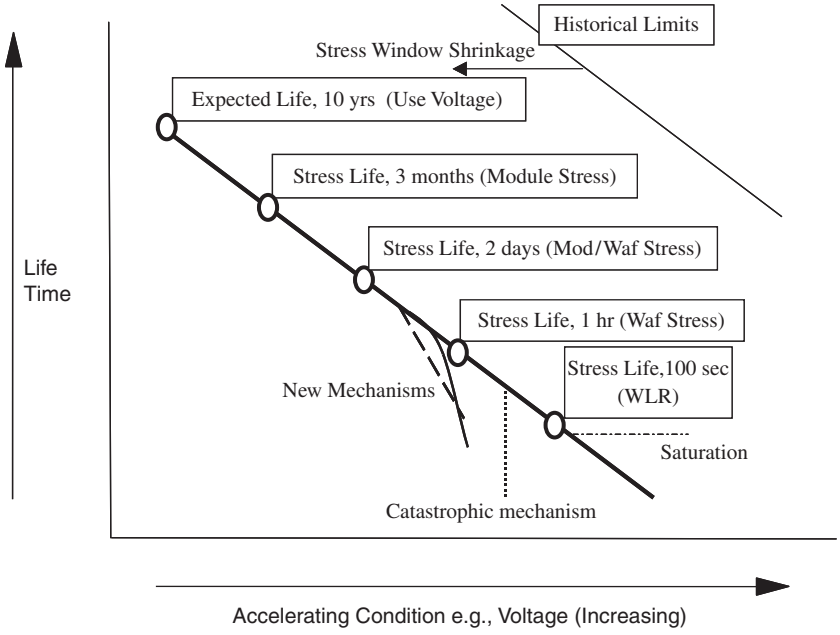

Figure 1.2. Lifetime projection curve with each large circle representing the 50% life for the accelerating or use condition and showing various nonlinearities that can compromise a highly accelerated stress.

to the tester and awaiting test. Even for mechanisms that do not relax when not under bias, this method has the disadvantage that no data can be obtained for the first few weeks after the hardware becomes available because the wafers are being built onto the chip carriers or modules and this build typically takes several weeks. The advantage of this method, even today, is that the stress equipment is relatively inexpensive and the test equipment is general enough to be used for readouts for several mechanisms. This amortizes the test equipment across all of these stresses and further decreases the cost of ownership.

However, for accuracy and simplicity, it is desirable to use the same equipment for the stress application and the readouts. This also minimizes handling damage and human error. Data acquisition improvements during the past 20 years have allowed the detection of exact times-to-fail even for the long three-month stresses. Now whether the stress is a long, three-month stress or a very short stress, the exact times-to-fail can be obtained with the same equipment.

In addition, advances in the state of the art in hot carrier stressing, in dielectric stressing, and in electromigration stressing have moved the leading practice to the far right two points on Figure 1.2, that is, to stresses of hours and minutes. Typically, stressing with seconds of duration is used in conjunction with at least one additional stress of longer duration. For example, in the case of dielectric stressing, optimally three voltages are stressed with the shortest stress duration having a median fallout on the order of 10 to 100 sec and the longest (lowest voltage) having a duration of 1000 to 10000 sec. This has been practiced for the last 5 to 10 years for dielectric stressing but as the state of the art thickness approaches 1 nm, it may actually become necessary to return to the relatively long stresses of several months. In the case of electromigration, the quantitative bridge for stressing on the order of seconds was only demonstrated a few years ago [6–8].

One of the obvious points that should be explicitly made is that for a three-month stress, the extrapolation to a 10-year use life is only a factor of 40. While for a 100 sec stress, the extrapolation is a factor of more than 1E6. Much more care concerning the projection error must be exercised when one is extrapolating six orders of magnitude, than when one is extrapolating only a little more than one order of magnitude.

Also one has to investigate very carefully whether any change in the accelerating condition of the mechanism in question has occurred or can occur under any reasonable set of conditions. This is depicted graphically in Figure 1.2 as new mechanisms, which may occur above a certain stress level. If any such mechanisms exist, they may be either linear or nonlinear as shown, and they would preclude exceeding that stress level since no straightforward model or bridge to use conditions would be possible in that case. Another possibility is that the accelerating or stress condition saturates above a certain level. That is, a further increase in the accelerating condition causes no resultant decrease in the lifetime. Again, a mechanism of this nature would limit the accelerating condition to a value no higher than just below its saturation value, and even there, the physics would need to be well understood.

The final concept that Figure 1.2 attempts to depict is that the window available for stressing is shrinking as the technology features continue to shrink. In the past, a very significant margin existed in many of the mechanisms. Often a calculated acceleration would demonstrate that the stress had gone well beyond a 10-year life but no wearout for that mechanism had been observed. Dielectric stressing is an excellent example of this. For 12 nm oxides, dielectric stressing on the order of three months could not detect any indication of wear out, and the stress focus was just on the extrinsic or defect part of the curve. Today, models are constructed to understand whether fractions of nanometers can be shaved from the oxide thickness and still meet the end of life targets.

1.2.3 Accelerating Condition

Acceleration concepts have been discussed and we now turn our attention to accelerating conditions. What types of external forces can be applied to the semiconductor test structure in such a way as to cause the end of life, say 10 years, to be reached in a time period that is significantly shorter than that 10-year life. In principle, the shorter the stress time the better, as long as one can still bridge to the use conditions. Examples of accelerating conditions for semiconductors are shown below. This extensive list includes all of the common accelerating conditions and a list of pertinent mechanisms with chapter references where applicable.

- Voltage (DC)
  - Dielectric breakdown (3.4)
  - Electromigration {indirectly} (7.3)
  - Hot carrier (5.2)
  - Temperature bias stability (4.3–4.5)
  - Interconnect opens and shorts (7.3)
  - Ionic contamination
  - Energetic particle-induced soft error mechanisms
  - Variable retention time mechanisms
  - Leakage mechanisms
- Voltage Change (AC)
  - Conducting hot carrier mechanisms (5.2)
- Temperature
  - Dielectric breakdown (3.4)
  - Electromigration {indirectly} (7.3)
  - Stress migration (6.2–6.5)
  - Interconnect shorts and opens (6.2–6.5, 7.3)
  - Hot carrier (5.2)
  - Temperature bias stability (4.3–4.5)

- Ionic contamination
- Variable retention time mechanisms
- Leakage mechanisms

- Temperature Change
  - Interconnect opens
- Temperature Change Rate
  - Interconnect opens
- Current Density
  - Dielectric breakdown (3.2–3.4)
  - Electromigration (7.3)
- Humidity
  - Corrosion
- Humidity and Pressure
  - Corrosion
- Harsh environment
  - Corrosion
- Mechanical pull tests
  - Mechanical strength of interconnects and adhesives
- Radiation
  - Some dielectric breakdown concerns
  - Soft error rate (SER) for certain flash memory

Note: The SER effect does not get worse with time for most CMOS devices.

1.3 MECHANISM TYPES

1.3.1 Parametric or Deterministic Mechanisms

A parametric or deterministic mechanism is defined, albeit somewhat arbitrarily, as any mechanism that impacts all identical structures nearly equally. A stress for this type of mechanism will always cause the parameter under question to shift. And, even if many samples are stressed, the shifts will all be very close to the same value assuming all of the stressed structures are identical. For this reason, very small sample sizes can successfully be used to characterize a parametric mechanism. Most of the variation of the shifts observed for parametric mechanisms is caused by variations of the controlling parameters and not by random statistical variation.

The hot carrier (HC) mechanism is one example of a parametric mechanism. While a field effect transistor (FET) is turning on or off, the gate current has a peak value resulting from channel hot electron injection. These electrons gain enough energy to surmount the Si/SiO_2 interface without suffering energy-losing

collisions in the channel. The electrons are trapped and result in FET performance degradation. This mechanism is uniform and parametric in the sense that for a set of FETs that are all structurally identical, the shifts resulting from the above stress will be almost identical across all of the devices stressed; that is, the shifts will be determined by their parameter values, not by the random variation. In practice, if chips from several wafers or lots are stressed, variation will be seen but that variation will be a function of slight differences in the structures of the FETs across the wafers and lots.

Electromigration is an example of a mechanism that has aspects of a parametric mechanism. A current flowing through a line will cause atomic motion in that line. If that line is aluminum, significant atomic motion will occur at higher current densities and will cause the line resistance to increase and ultimately open. This is fundamental to the structure and the metallurgy. For high enough current densities, electromigration will always occur for that aluminum line. It is not caused by a defect although it can be exacerbated by a defect. Although the physics cannot be changed, sometimes it is possible to mitigate the problem. For example, if redundant layers of certain other metals are used in conjunction with aluminum, the sandwich line structure will increase in time-zero resistance if the overall cross-sectional area of the line remains constant, but electromigration typically will only cause a resistance increase and not an open under the same high current-density stress. For some metallurgies, no electromigration will occur even at higher current densities. However, it must be pointed out that in the case of electromigration, there are also aspects of a random mechanism because the grain structure of the line is random. And this randomness is true for metallurgy that is identical in processing. Typically, larger sample sizes are necessary when stressing mechanisms that have a greater degree of randomness.

Obviously it is crucial to understand the fundamental physics for a parametric mechanism. Once the physics is understood, strategies can be put into place to mitigate the effect or to eliminate the problem by structural or operating-point changes. Sometimes mitigation is possible and sometimes it is not. For the electromigration example, tungsten is sometimes used for the lower levels of wiring where the distances are small and the higher time-zero line resistance is tolerable. For the longer wiring levels, the resistivity of tungsten is too large and aluminum or copper must be used and other strategies invoked to decrease the impact of electromigration.

Once the physics is understood, so that all of the controlling parameters are identified and each of their impacts quantified, it is possible to address elimination and mitigation strategies. To be able to quantify the impact of a given parameter, it is usually necessary to characterize the impact of that parameter on a test structure where individual control of all of the terminals is possible. If, for example, the physics of the mechanism is related to a parasitic edge transistor in parallel to the bulk transistor, the decision must be made as to whether to change the process to eliminate the parasitic transistor, or to simply mitigate its impact on the circuit. A problem may occur only at one extreme of the normal processing window or set of biases and tolerances. HC is one example since it is worst at the shortest channel

lengths for a given set of stress conditions. In some cases the strategy may be to run the process to a tighter manufacturing limit. Because this type of mechanism equally affects all structures with the identical process, only a few structures need to be investigated to reasonably well characterize a parametric mechanism. However, these devices under tests (DUTs) must all be structurally identical.

From this previous discussion it should be obvious that it is necessary to investigate parametric mechanisms at all salient process window extremes to ensure that no undesired effects occur. Again, each process window investigation point requires only a small sample size.

Below are some examples of parametric mechanisms and one or two strategies for controlling or eliminating the effect. In most of the cases, there are other strategies that could also be invoked. Applicable chapter references are shown.

- *Hot carrier*: design point change, e.g., lower operating voltage the device experiences
- *Bias temperature stress or (negative bias temperature instability)*: design point change, e.g., decrease operating voltage
- *Ionic contamination*: discovery and removal of contamination source
- *Stress induced leakage current*: design point change, e.g., decrease operating voltage or thicken gate oxide
- *Electromigration*: design point change, decrease current density
- *Soft error rate (radiation induced)*: design point change, increase critical charge of pertinent cells or decrease charge collection efficiency. This is not discussed further in this book

1.3.2 Structural Mechanisms

Structural mechanisms are those mechanisms for which the fails physically occur in the same place. The distinction here from the structurally induced parametric fails is that these fails are only a function of a structural artifact. Although these definitions are all somewhat arbitrary, they help in understanding the sample size differences recommended in the later chapters. Usually significant failure analysis is required to determine that a particular failure type has a structural, systematic signature. Often this signature only occurs at one of the process extremes so that it does not occur on every wafer or lot. Sometimes it is even more difficult to identify because not only does it only occur at one process extreme, it may also require a certain set of process biases and/or tolerances to align in just a ''right'' way for the failure to occur. This may take the form of one part of the wafer having an acute susceptibility, or it may be tool dependent. In some of these cases, it may appear random, while in fact, the fail is part of a manufacturing defect or process window tail. This can usually be avoided if a large enough sample is investigated and if at least part of that sample comes from the salient process extremes. If the failure analysis then identifies a particular feature failing more than once, that feature should undergo very careful scrutiny.

Sampling is very important since the problem may not impact all lots or wafers or die equally. The sampling must gauge all process variations unless one process extreme can be identified as the worst case for the given mechanism. A minimum of three manufacturing lots is recommended with one produced at the identified critical extreme. For this type of mechanism, sampling for random statistical variation is less important than sampling for the pertinent process extremes.

Once this type of mechanism is understood, it can often be mitigated with a strict application of statistical process control (SPC). However, no structural fails should be acceptable within the normal process limits. Otherwise this would represent a technology weakness that, if accepted, would likely result in an inordinate number of customer failures even with tight SPC. It is always better in the long term, to fix a problem rather than to try control it. Fixing the problem can take the form of structural modifications or a redefinition of the process limits. Especially for hardware made later in a program, the normal exercise of SPC, once the process line is full of hardware, may eliminate the possibility of a problem.

Process improvements made during manufacturing can inadvertently introduce new structural mechanisms. An effective method to avoid this is to sample a large number of chips and wafers looking for changes even in time-zero characteristics. Changes in the time-zero characteristics will not always flag a change in a reliability mechanism, but a change in the time-zero characteristic should be carefully investigated especially if a significant database exists for the normal properties of the parameter. Wafer-level reliability (WLR) is an even better gauge as to the impact of process improvements on reliability. And in fact, occasionally the time-zero properties have been changed through process changes only to make the reliability worse in a direct tradeoff between yield and reliability. All of the examples for this type of mechanism are very technology/process dependent.

1.3.3 Statistical Mechanisms

Statistical mechanisms are defined as those mechanisms that are primarily random. Thus the more susceptible area to a particular statistical mechanism, the more likely that mechanism will cause a chip fail. The occurrence of the fail will be totally random within that susceptible area. This is in contrast to a structural fail, which will always occur at a given feature within a structure. It is also in contrast to the parametric or deterministic mechanism for which the process variation or process extreme will have a larger impact on the result than does the random statistical variation within a given process point. One must be careful at this point because the statistical mechanisms are also caused by fundamental physics unless the discussion is limited to defects. The distinction is more focused on the impact that the random statistical variation has on the investigation of the mechanism.

For the statistical mechanisms, it is critical that a significant sample be stressed to understand the failure behavior. Hardware made at the process extremes is only important in as much as it has an impact on the occurrence of the mechanism. For example, a thinner oxide will have a higher failure rate for a given set of conditions than the thicker oxide. In this case a significant sample should be stressed at the process minimum thickness. The fails will be randomly distributed throughout the area in both cases.

Aspects of dielectric breakdown and electromigration are two examples of statistical mechanisms. The aspect of electromigration that is statistical in nature is the grain structure of the line. Although the fabrication conditions very clearly impact the grain structure, that grain structure will still have random variation even when the fabrication conditions are identical. And first dielectric breakdown is generally accepted to be random and has been shown to follow a weakest-link behavior resulting in a Weibull distribution for its cumulative fail distribution. This will be discussed in Section 1.4.7.

For a statistical mechanism, a much larger sample is required than for a parametric mechanism. The reason is that there is a given randomness in samples that have identical structures, even for identical processing in as much as possible. Thus for a given stress on parts that are identical as far as can be known, there will be a distribution of results instead of a single-valued result. The sample size must be large enough to ensure that some mean or characteristic life is truly representative of the population from which the sample is drawn and not just caused by random statistical variations. To further elucidate this point: if five processes were being compared and five small samples were chosen for each case, very different results could be obtained just based on the random statistical variation, and it could be possible, and in fact likely, to not choose the best process due to this random variation. Obviously it must also be assumed in this case, as well as the other cases, that the samples being investigated are representative not only of the population from which they were drawn but also from the entire production population.

A discussion of statistical mechanisms must include a general discussion of defects. There are four classifications of random defects depending on the time of occurrence. The first class of defects is screened out at time zero. These are obviously the yield fails. The next class of defects is the infant defects. These pass all time-zero testing but fail very early during stressing. These first two categories of random defects typically reflect the manufacturing defect level. The third class of defects is the operating life defects. These defects reflect the ultimate manufacturing and process capability and cause failure throughout the useful life of the product. The final class of defects is wearout. This category of random defects reflects the technology capability and was the subject of the preceding paragraphs.

Figure 1.3 shows the instantaneous failure rate, or hazard function, versus time for the last three reliability classifications depicted as the so-called bathtub curve for reliability. The first stage on the left is the early or infant mortality region. The instantaneous failure rate is here characterized as rapidly falling, as the parts are in the very early stage of use and the weaker parts are still failing at a

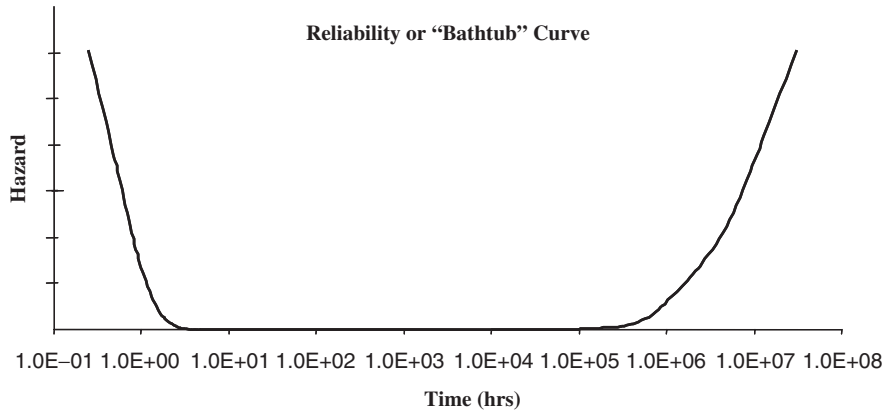

Figure 1.3. Reliability bathtub curve, a hazard plot or instantaneous failure rate plot in units of fails per time.

relatively rapid rate. A short, highly accelerated stress, called a burn-in stress, can often protect the customer because that stress has the possibility of moving the hardware past the infant region before any of it is shipped to the customer. Thus the customer only receives parts that are already in the operating life region where the failure rate is very low. Although shown constant for convenience above, typically the failure rate is slightly decreasing during the operating life for semiconductors. Finally, ultimately the hardware will start to wear out. The depiction here is of a wearout target that is 10 years or about 100 K hours.

1.3.4 Infant Defects

Infant defects are those that pass time-zero tests without fail, but fail shortly thereafter. Often failure analysis of the infant fails reveals structures that were extremely marginal but somehow survived the time-zero or yield tests. These defects should be entirely due to manufacturing defects. As the process matures for a new technology, the level of these defects should decrease to a low number bounded by the capability of that manufacturing facility.

There should be no structural fails contained within the infant defect population. The existence of structural fails could indicate a technology problem that still needs to be addressed or a process window problem that needs to be eliminated.

Burn-in is the primary method of removing infant defects once the product is completed. Burn-in for semiconductors is almost always done with a short temperature and voltage stress. Sometimes both a wafer burn-in and a module burn-in are performed. The wafer burn-in will typically occur at elevated temperatures and very high voltages compared to the use conditions, and will last only a few seconds. The module burn-in conditions will include an elevated

temperature and a voltage that is higher than use condition, but lower than the wafer burn-in voltage, and will be applied for several hours. As the process matures and the manufacturing line defects decrease, the burn-in times and/or conditions will also typically be decreased.

There are several types of module burn-in. The most effective burn-in is also the most costly. The simplest but least effective burn-in is a DC stress where the parts are not exercised during burn-in and hence, neither are they tested while under stress. They are only tested before and after the burn-in stress. This method suffers from the fact that not all circuits are being exercised, and therefore the burn-in coverage is less than 100% and possibly very significantly less than 100%. It also suffers from the fact one cannot be sure that all of the parts really received the voltage stress. This could happen for several reasons but the parts that do not receive burn-in are called escapes and will be the most likely to fail in the field. The next level of burn-in is for the parts to be exercised but not tested during burn-in. Again they are only tested before and after the burn-in stress. This method still suffers in that parts may be incorrectly inserted during stress, hence one cannot be sure that all of the parts really received the voltage stress. The most complex, costly, and effective burn-in is for the parts to be both exercised and tested during burn-in as well as before and after the burn-in stress. This last option ensures that all of the parts that are put into the burn-in chamber will be flagged as fails at both the stress and measurement conditions. This last option also minimizes escapes since the chip responds in the oven to the testing while it is at the stress conditions. However, even for this last option a few escapes can still occur.

Infant defects can be minimized by attention to general line cleanliness and particulate control and monitoring. Again, the defect level typically decreases as the technology maturity increases on a given fabrication line.

Good design-for-reliability practices also improve the apparent defect learning. One of the more common practices is to use the minimum groundrules only when absolutely necessary. Another practice is to use special features only for those circuits where there is great leverage.

Infant defects can impact many features and their nature depends on the details of the technology. If the infant defects are not adequately eliminated either through strict process control and line maturity, by burn-in, or both, then the product will have high very early fallout rate when the customer starts using it.

1.3.5 Operating Life Defects

Operating life defects are those defects that occur, as the name suggests, during the operational life of the product. The instantaneous failure rate should be small and must be contained within the target to ensure that the product does not fail at greater than the expected rate. To meet a given specification, accelerated life modeling is used to predict the product defect level and to bridge that to the given specification. The sample size used to determine the level of operating life defects must be relatively large since this is a random mechanism and the operating-life defect level is very small.

1.3.6 Wearout

Ultimately, some part of any system will cause wearout. The time to wearout is dependent on many factors. The first factor is the technology itself. In addition to the technology, examples of other factors include the manufacturing process, the temperature, and the voltage conditions. The frequency of use or duty cycle is also very important. For example, a redundancy check circuit that is only used when the chip is powered up can tolerate much more degradation per cycle than can a circuit that is always operating when the chip in use. Other factors also have an impact on the onset of wearout including the circuit itself.

The objective of modeling in this case is to ensure that the wearout does not start until well after whatever the lifetime specification is for the product. Again, because this is a random statistical mechanism, sample size is important.

One good example of a random mechanism that causes wearout is the intrinsic dielectric mechanism of relatively thick oxides.

1.4 RELIABILITY STATISTICS

1.4.1 Introduction

The following treatment of reliability statistics is very abbreviated since the primary focus of this book is the physics of the CMOS reliability mechanisms, and there are many excellent texts on the subject of reliability statistics [9–13]. Ideally, the reader will already have some background in reliability statistics. However, since that will not be the case for everyone, this abbreviated treatment is included.

The author's experience, from teaching a course based on this material, is that some students raise objections to any treatment of statistics because they do not understand the necessity of even a minimal background. By the end of the class, however, the students have an appreciation for this background. This material is necessary to understand the following chapters, and in fact, further treatment of these concepts is given in the remaining chapters of this book.

1.4.2 Assumptions

Many assumptions, and indeed compromises, must be made in the exercise of semiconductor reliability. The first assumption is that the variation in the stress and test results is due to just the random statistical variation and/or random process variations. This assumption can usually be met, but care must be taken in the design of the experiment to ensure that the test site is appropriately designed, that appropriate equipment is being used for the stress and test, and that the equipment is in calibration.

The second assumption is that the sample stressed is representative of the population. Clearly, if the sample is not representative of the population, one can question what the results really mean and how pertinent they are to what will

happen to the product in the field. This assumption is typically only weakly satisfied in the case of technology qualifications. Fortunately, technology qualifications offer special opportunities that provide relief from only weakly meeting this assumption. Quality control (QC) sampling is very different and can typically satisfy this assumption quite well.

A third assumption is that each fail is independent of all other fails and therefore has no impact on those other fails. This assumption significantly simplifies model generation and can generally be considered true for reliability defects. However this is typically not true for yield defects and sometimes there is a direct relationship between yield and reliability fails [14–16]. For some special fail cases one can decrease yield fails at the expense of increasing reliability fails and vice versa, with appropriate process changes. Please note that we are talking about process changes and not burn-in tradeoffs here. It is well known that yield fails are not usually randomly distributed on a wafer. Stapper [17, 18] and others have demonstrated that defects often follow a negative binomial distribution. Simplistically this means that defects are clustered about areas on the wafer instead of being independently distributed about the wafer area. Thus one is more likely to find a second defect in the general area that already has a defect than near an area on the wafer that has no defects. Yield issues and the negative binomial distribution are both beyond the scope of this book, but they are mentioned because of the occasional relationship between yield and reliability fails.

A fourth assumption is that it is possible to represent the discrete results obtained during reliability stressing with continuous functions. Again, this assumption is necessary in order to be able to generate models with predictive capability. Much debate and agonizing can surround this assumption depending on how well the data are behaved and how well they fit the continuous function and which continuous function fits best. Often the time range of data is limited to such a narrow window so as to allow the use of more than one continuous function to characterize the data. Thick dielectric data are a classic example of this case and the appropriate model was debated for many years. Sometimes very early or late fails cloud the fit, but much of the time these fails can be explained due to known phenomena, albeit often only after the stress. Care must be taken to ensure no new regime or mechanism has been introduced during the stressing if some fails behave significantly differently than the majority of the population of fails. If this is the case, the stress conditions may have been too extreme and the fails are no longer representative of the use conditions. Even if all of the fails follow the same distribution, care must be taken not to stress at an extreme that is not representative of the use conditions as described previously.

Two or more modes can be present with two different characteristic life times and two different characteristic variation parameters. Here a bimodal fit can often be achieved. An example would be dielectric breakdown for a relatively thick dielectric of 5–10 nm. In this case both the characteristic life and shape parameter are very different for the intrinsic or wearout population and the extrinsic or defect population. There is a clean break in the curve between the two populations and nearly all of the fails can be clearly attributed to one population or the other.

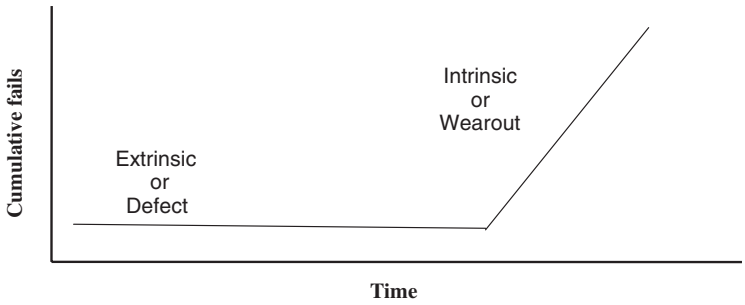

Figure 1.4. Example showing the extrinsic, or defect failure typically having a shallow slope, and the very late intrinsic, or wearout failure having a steep slope.

This example is shown in Figure 1.4 for the case of a relatively thick dielectric. (The case where the characteristic life is the only variable for two dielectric distributions will be discussed further in the dielectric chapter in relationship to the Weibull distribution.) This has direct applications to other distributions as well by inference. One would hope and expect to be able to identify two different failing types through failure analysis when two distributions are seen. This is usually the case, although it is not always true possibly due to the obvious limitations of failure analysis resource and the total number of fails. In the end, some continuous function must be determined to be able to make the reliability predictions.

A final comment about sample size is that there are some applications where the attempt is so futile that other considerations must be used. For example, the number of parts required for space applications is typically small. In these cases, the qualification would need to bridge to similar hardware, use very long-term product stressing, or possibly some other method.

1.4.3 Sampling and Variability

One can have variability in results due to raw material variations, process variations, test equipment variations, stress equipment variations, or random statistical variations. The process norm and its variations as well as the random statistical variations are the subjects of interest. The other variations should be minimized as much as possible.

Test and stress equipment variations can be minimized by timely and scheduled calibration, and detected by using standards during the testing and stressing. The standards are parts that have not undergone stress and therefore should show no variation from readout to readout. For *in situ* test equipment, that is, stress equipment that not only applies the stress but also performs the readouts, standards cannot be as readily used, so more care should be taken to ensure that equipment is calibrated. It is important to establish the variation of each piece of stress equipment. For example, ovens should have a temperature profile done for

each oven and that profile should be done when the oven is fully loaded with DUTs. Ideally, a thermocouple would be monitored in each (DUT) position. For most stresses, one would want to know the temperature to within a few degrees or better. A fully loaded oven will change the airflow characteristics and could result in hot or cold spots during the stress. Each stress should be considered in a similar fashion. Note that here we are not talking about variation that would necessarily be random in these cases. If the loaded oven interior is running five degrees high, the projections will all have an additional error caused by that source. If the projections are far enough from the norm so that the ovens are checked after the fact, appropriate corrections can be made then; however, it is much more satisfying to the customer and the responsible engineer to have the equipment calibrated before the stress. A more extreme but yet typical example is the case where several tiers of DUTs are put into an oven. Depending on the airflow, each tier may have a temperature offset and even the DUTs in one tier could be different. Here is a case where temperature profiling is crucial because normally one would not keep track of the DUT position within the chamber. Thus it would not be possible to apply correct temperatures after a stress even if the oven were to then be profiled. The ovens provide a clear example of the importance of using calibrated stress and test equipment to avoid nonrandom variations. They also provide a clear example of a nonrandom variation that could be interpreted as random variation if it were unknown as in the case of temperature variation within each tier and even DUT. Another example of nonrandom variation would be dielectric stresses where one needs to ensure that the interconnect wiring has a low enough resistance, such that when leakage current is present, either due to tunneling currents or stress induced leakage currents, that the current does not cause voltage drops along the wiring and especially that the current does not cause nonuniform voltage drops along the wires. In that worst case scenario, each measurement and stress would have its own set of variations that would have to first be understood and then considered when doing the projection. It is obviously highly desirable to minimize these nonrandom variations everywhere possible. One must consciously think through each experiment to ensure that the experiment is correctly designed using appropriate and calibrated equipment. One last example of the use of appropriate equipment includes the detailed consideration of the how the measurements will be taken. If the equipment has been calibrated but it is to be used in an auto-range mode, care must be taken that there is enough time in that auto-range mode for the measurement to be returned to the computer. Even just a step up to the next higher range might cause an additional delay time. If the time required for the measurement was set at one scale without thought to these cases, errors could occur that would be nonrandom and very confusing. If this type of variation is present and not detected, it will add to the overall error. This error could cause the projections to be skewed due a number of factors including, for example, a constant offset or additional data scatter. In some cases the projections would be pessimistic and in others the projection would be optimistic. However, without recognizing the error in the system, the reliability analyst would not even know to doubt the projection. Sections in each of the

following major chapters will address how to avoid many of these pitfalls for each of the mechanisms in terms of test site design, stress, and test.

One also has variation due to the process differences within each chip, wafer, and lot. Ideally, the sample chosen for stress would contain all process iterations possible in the manufacturing line.

Note that the challenge here is in the early qualification of new technologies and not in the ongoing inspection, or quality control, of the manufacturing line. These stresses are done at a much lower stress level and a shorter time than the qualification stresses since the hardware must still be shippable after these stresses. That is, a significant portion of the intrinsic life of the product cannot be used during the QC stressing. It is also true of burn-in (BI) that a significant portion of the intrinsic life of the product cannot be used during BI. For this QC issue, one has a continuous flow of parts through a short stress and a test. The good news is that a very large sample is typically stressed so one has a representative sample of the population for that time period. This stressing, taken in conjunction with the qualification stressing, does give a good overall picture of the population. Thus for QC, as yield learning occurs and the appropriate process changes are made, hardware with those changes are available for the QC-type stressing and testing. Typically in this case, no special process window lots are fabricated but the hardware currently being made on the line can be stressed to the QC times and conditions.

Now let us return to the variation within each chip, wafer, and lot due to the process differences. Our focus here is now the early qualification of new technologies. Care needs to be taken to have samples that are as representative of the standard process as possible. As discussed previously for random statistical mechanisms, the larger the representative sample from the population, the better the statistics of the result. This is true for lots, wafers, and test structures. However, typically three to five lots of hardware with several wafers from each lot is about as much hardware that the reliability engineer can expect to obtain for a process qualification. Most specifications from JEDEC, a standards governing body used by most of the semiconductor industry, require a minimum of three lots. Recall that the assumption is that the sample is randomly selected from the total population. For the ongoing quality inspections this can be true; however, for the early qualification work, this is not possible. The first set of the lots that have the final approved process will most often be used for the qualification work. This hardware is also necessarily from a single snapshot in time. Because it is a technology qualification, the hardware must be sampled very early in the manufacturing cycle, and hence, before significant additional process learning has had time to occur. This hardware should be reasonably representative of that early vintage hardware, but without due diligence during the continuing process learning, the process may change in such a way that the reliability becomes unacceptable.

The sampling should always focus on the final expected process since the processing order of two processes can be critical even when they are ostensibly independent of each other [16]. This is not usually a problem but the engineer

should be aware that it could possibly make a difference. A careful design of the experiment is always a requisite to avoid confusing or misleading results.

The point of the above brief discussion is to point out that even with the best one can do to ensure that the sample is representative of the entire population, realistically one can only weakly satisfy this assumption in the semiconductor reliability world during the early qualification phases. Other strategies must be instituted to mitigate this weakness and they are discussed below. However, it is possible to do a much better job at satisfying this assumption in the ongoing quality control during the life of the product.

Returning to a focus on the types of mechanisms, if the lots are chosen wisely and if the experiments are carefully designed including the test structures, then one can investigate and qualify a new technology with a very high degree of confidence even though the sampling is extremely limited. The stress and test equipment to be used in the experiment must be carefully considered and calibrated, so that the variability in the experiment can be limited to just the process and statistically random variability that one is trying to understand.

It is at this point that previously given mechanism definitions, that is, parametric, structural, and statistical mechanisms, should be applied. The design of the experiment and sampling plan must consider each type of mechanism uniquely and separately. For the parametric mechanism, if only three lots are available for stressing, at least one of the three lots must be manufactured in such a way so as to guarantee the pertinent extreme of the process variation. For example in the case of gate dielectric, the thinnest gate permitted by the process definition must be investigated. The thickest gate permitted does not typically need to be investigated from a technology reliability perspective since the lifetime of the thicker gate will usually be greater, and depending on the variation, possibly much greater than the lifetime of the thinnest dielectric. That is not to say that no one cares how thick this gate dielectric gets. At its thickest extreme, device performance may be compromised so that extreme must be investigated for product performance related issues but not for reliability issues. Sometimes all of the pertinent process extremes can be fabricated on a single lot; however, often at least two lots are needed to obtain all of the process extremes necessary to address each of the parametric mechanisms.

Investigations to find structural mechanisms can also be nicely satisfied with lots made with processing to the extreme edge of the allowed manufacturing range. In this case it is not quite so obvious which extremes are required and lots made with nominal processing should also be investigated. However large samples are not required since by the definition given previously, if a chip is made with this particular set of features, it will fail. The challenge then is to ensure that a chip is made with those features. But, of course, unless one has already seen a fail, one does not know what set of features would cause a fail. Here, past experience is the best guide for finding that set of process conditions that might result in structural failures.

As suggested previously, large, representative sample sizes are needed for the purely statistical mechanism. For today's technology few mechanisms are purely statistical in nature. That is the good news in terms of sample size. The other good

news is that what one lacks in terms of representative lots and wafers, one can often compensate through the use of good test site designs. Also appropriate test site designs can provide a sufficient effective sample size even though the total wafer count may be limited.

There are several other types of variables in the realm of sampling that should be mentioned for the reader's awareness. Sampling wafers from a tool that has just gone back into production after a preventative maintenance cycle can cause problems, as can wafers from new chemical suppliers even though they have been approved.

1.4.4 Criteria, Censoring, and Plotting Points

Data can be collected either in terms of specific shifts or in terms of pass/fail criteria. Electromigration, for example, will cause resistance increases. Depending on the process details, the lines undergoing electromigration may or may not actually go to a high resistance state or an open. The data taken and recorded may simply be those lines that shifted by more than a given value or even just the number of lines shifting by more than the given value. The data taken and recorded may include the resistance shift of each line. Clearly, if the resistance shift of each line is taken, the data can later be expressed in terms of a pass/fail limit.

For technology qualifications, it is usually desirable to record as much data as possible including the actual shifts. These data are useful at several levels including failure analysis decision making and isolating wafer and lot dependencies. The total shift data are vital for the parametric mechanisms but are also desirable for the structural and statistical mechanisms. Once the product is in manufacturing and minimal stressing and testing can be done, often pass/fail criteria are adequate as part of the shipping qualifications. Even during the wafer stressing and testing for the shipping qualifications, chip failure location on the wafer can be used to great advantage and flag process problems early.

One manufacturer [14] has shown excellent results using product yield as an indicator of defect density for neighboring chips. They divide the chip into an edge chip region and a center chip region. Chips in a ''bad'' neighborhood, in either region, are 10 times more likely to fail during burn-in than those chips from a ''good'' neighborhood in either region. This work covered one hundred thousand to one million chips and included technologies from 0.25 μm down to 0.09 μm as well as both aluminum and copper. They give a quantitative expression for the number of chips that constitute a ''bad'' neighborhood and show excellent correlation between their modeled behavior and actual burn-in data. This paper also provides a nice set of references for earlier work in this area.

Most stress equipment built today has *in situ* capability. That is, it has the ability to both stress the DUT as well as to test that device. Hence, obtaining exact times-to-fail is simple. For example, the exact time for a 20% shift in the line resistance due to electromigration can be obtained from the test equipment. In this case the time for each fail or predetermined-shift is known uniquely and can be plotted uniquely.

Occasionally, times are known only for groups of fails. This may happen if older stress equipment is used or if stresses are done in ovens because of large sample sizes. Here the parts are stressed for a given period of time, removed from stress, taken to a tester, tested, and then put back on stress. For reference, this equipment is typically referred to as *ex situ* equipment. Now only those parts failing after the given time period are known. The cumulative population failing at that time is typically then plotted against that time. However, depending on the design of the experiment and the specific issues being investigated, it is sometimes preferable to break any experiment into intervals and plot the cumulative fallout in the middle of that interval be it on a log or linear scale. Plotting positions are discussed in more detail below.

One of the challenges of *ex situ* testing is that care must be taken that the mechanism does not relax. For example if one were stressing for ionic contamination in *ex situ* equipment, one would need to apply a bias to the parts as they came off the *ex situ* equipment so that the parts did not recover while awaiting test. Historically large automotive batteries were used for this task when *ex situ* equipment was used to stress mechanisms that relaxed.

Data may be censored for many reasons. There are statistical procedures for handling nearly every type of censoring. The most common reason for censoring is likely lack of stress time. Parts can only be stressed for a given time period and it is expected that some part of the sample will survive beyond that time. Normally for parametric mechanisms, one would have a model that predicted how long the product should last given the shift, the stress conditions and the time under stress. The model may allow the entire sample to be used.

If most DUTs failed within a relatively narrow time range but several did not, two questions arise. Were these DUTs significantly better than the rest and potentially not subject to the mechanism that caused the majority to fail, or was the stress or test somehow compromised on these parts? Again, failure could be defined as a certain shift or an actual open, short, or cessation of operation. The nonfailing parts might be candidates for censoring if the stress was not fully applied, or they might be part of a second and better population so that a bimodal distribution should be used. There are at least three common ways of handling these cases depending on the exact circumstances. One characterization method would be to characterize the failing distribution with a bimodal distribution. Another way would be to censor the unfailing or late-failing parts, subtract them from the sample, and recalculate the failing distribution and times based on the new sample size. A third method would be to leave them in the sample. It is clearly highly desirable to know the reason for the preponderance of early or late fails before making the decision as to how to characterize them. However, the statistics in doing so are straightforward. Another pertinent example is those DUTs that pass the initial test but fail during the first application of the stress. These are not yield fails, nor are they real reliability fails since they failed as the initial stress was being applied. These would actually be in the ship product quality level (SPQL) category of fails. This is the class of fails that cause a product to look good as shipped from the supplier, but when the customer first turns on the machine, it

fails to function. This is obviously a very serious class of fails but can show up on reliability plots and unless handled correctly can compromise the reliability conclusions.

Step stressing is a technique by which multiple stress conditions are obtained on the same set of hardware. These techniques are advantageous because a smaller overall sample is then required since the same parts are used for every condition. Usually, but not always, the steps have increasing acceleration as outlined below, and if done carefully, step stressing is powerful. One of the attributes of a well designed experiment is equal representation across all cells so that if one stress condition or cell gives results unlike the other stress cells, it cannot be due to one wafer or lot dominating that cell or lacking in that cell.

Historically, the key to step stressing is that each new applied condition must result in an acceleration of about $100\times$ in time from the previous cell. With a $100\times$ acceleration, it is possible to calculate all of the acceleration parameters to the same degree as if two independent cells had been stressed. If this cannot be achieved, the separation in the results may not be adequate to interpret the data, compromising the entire experiment. If a step stress is contemplated, a focus on experimental design is necessary. This type of step stress is discussed further in Section 2.3 with a dielectric example.

1.4.4.1 Plotting Position. The simplest and most straightforward method of plotting the cumulative fails would be to simply plot the fail percent against the time of fail. For example, assume that a stress of four parts resulted in one fail at the one-hour readout, a second fail at the second-hour readout, a third fail at the third-hour readout and finally the fourth fail at the fourth-hour readout. The concept of a plotting position is to plot the fail in the most representative position in the interval of failure. Thus instead of plotting the first fail at 25% in our example, we plot it someplace between 0 and 25%. One option is to plot the fails at the midpoints of the intervals such the fails would be plotted at the 12.5%, 37.5%, 62.5% and 87.5% points on the y axis corresponding to 1, 2, 3, and 4 hours respectively on the x axis. Conceptually, this y-interval plotting position may be the simplest of the plotting positions. This plotting position is shown for 4 data points and 20 data points in Figure 1.5. There can be several choices for plotting positions and good statistical arguments for each. Some of the more common plotting positions are given below, but a thorough treatment of the statistical rationale behind each choice is beyond the scope of this book. The interested reader is referred to the statistics texts. In addition to the statistical arguments suggesting the requirement of a plotting position, there would be the practical problem that plotting simplistically as mentioned above, would require plotting a point at 100% fallout which for log plots does not exist.

The smaller the sample, the more crucial is the use of an appropriate plotting position. Several authors have suggested different plotting positions, any one of which may be best under a given set of circumstances. Two common plotting positions are $F(t_i) = i/(n+1)$ or $F(t_i) = (-1/2)/n$ [19] and $F(t_i) = (-0.3)/(n+0.4)$

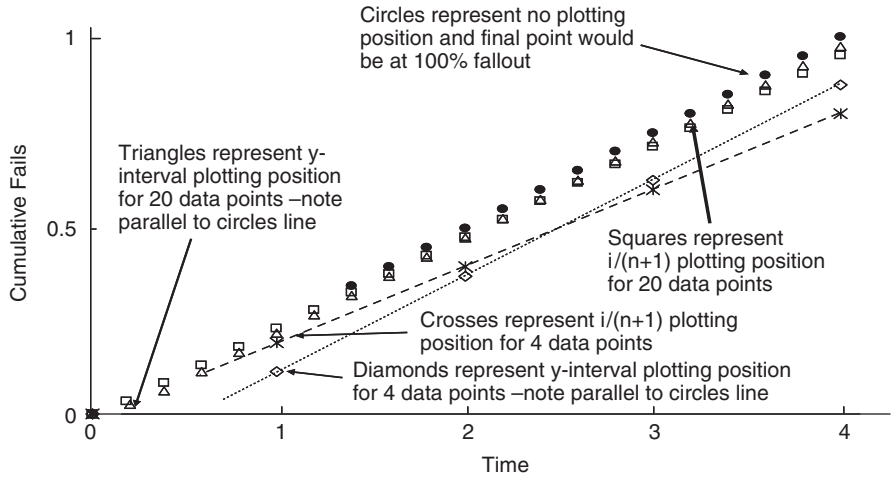

Figure 1.5. Plot comparing *y*-interval and *i*/(*n*+1) plotting positions with no plotting position for 4 and 20 data points.

[20]. Other authors have suggested various alternatives to the above plotting positions. It is a simple matter to compare the possibilities with one's actual data especially in today's world where the data is typically analyzed using a computer.

Note that the above discussion is only applicable and pertinent when graphical procedures are used. Typically the analysis and estimation procedures will be done using computer procedures.

As can be seen by Figure 1.5, there is little difference between the values of the *y*-interval plotting position and the $F(t_i) = i/(n+1)$ plotting positions for samples of 20 or greater, although the slopes are somewhat different. There are significant differences both in value and slopes for both of the plotting positions shown on Figure 1.5 for the case of only four points.

1.4.5 Definitions (Normal)

1.4.5.1 Failure Figures-of-Merit. The first set of definitions that are given are for the generic failure language that is used within the reliability field. Wearout, intrinsic failures, or end-of-life failures are defined at the end of the expected life of the product. Defect or extrinsic failures are typically expressed both in terms of end-of-life failures and various points of time throughout the life of the product especially after one year. As seen below, these expressions may either be in terms of the cumulative fails or the failure rates of the hazard function. The terminology used for cumulative fails is typically parts per million (ppm), and the terminology used for failure rates may be fails per 1000 hr or fails per billion device hours (FITs).

1.4.5.2 General. The independent variable for nearly all practical cases within reliability statistics is time. Sometimes the distribution parameters will be used as the independent variables when comparing distributions or models. This might take the form of plots comparing the mean and standard deviation for various lognormal distributions. (See Section 1.4.8) For Weibull distributions this might take the form of a comparison of the various shape parameters at the characteristic life for each distribution (See Section 1.4.7).

Recall that our first assumption above was that the variation in the results was due only to random variations. A ramification of this definition is that the actual value is not known at any precise point in the domain of interest (nearly always time for reliability work) due to the random statistical variation. Within the framework of reliability modeling, a distribution function is sought to describe the pattern of values that are most likely to occur.

1.4.5.3 Probability Density Function (PDF). The probability distribution function or probability density function is that function that describes the fails in each period of time. This is nothing more than a histogram with time intervals as the abscissa and the percentage of fails as the ordinate or y-axis value. Since this function gives the fails that have occurred in the previous time steps, if an analytical expression can be found that accurately describes those past times steps, that expression can then be used to predict the fails that would be expected in future time steps. For example, if a stress results in all samples failing of 50 samples on stress, and the data is then grouped within subregions, the discrete distribution in Figure 1.6 results. Note that these 50 data points were generated by a simulation with a mean fail time of $\mu = 10$ hrs and a standard deviation of $\sigma = 2$, and hence, the data are representative of some mechanism that causes the population to fail with the above characteristics. The intent of this plot is that it represents a small sample from a much larger population. The next charts will also present data representative only of the empirical sample, not of the true population. These charts will highlight the randomness of the resulting distributions when only small samples are possible.

The assumption is that the true mean and standard deviation of the population is known. The charts then depict that variation for the distributions associated only with choosing a small sample. There are several features to notice about this empirical distribution before continuing. The first is that it does indeed have the general appearance of a normal distribution with a mean of 10 hr. The second observation is that it has significant aberrations from the expected values both for the 3.5–4.5 time interval and the 6.5–7.5 time interval. But this is just a small sample from a true normal distribution so the aberrations are not a consequence of data collection issues; they are just a function of the random variation to be expected because the sample size chosen was only 50. Note that if this would have been a real experiment, the early fail at about four hours might have been attributed to something other than the primary fail mechanism, which acts between 6 and 14 hours and hence, considered as a nonrepresentative fail and eliminated from the population. Obviously, this could be the case in a real experiment; however, here it is an integral part of the normal fail distribution and

is equivalent in all respects, except its time to fail, to the later fails. As such it cannot be eliminated from the sample without compromising the experimental conclusions. The same observations could be made for some of the fails in the 6.5–7.5 time interval. Here two to three fails were expected and seven were observed for this simulation. One could wonder about power spikes and any number of other experimental problems during this time interval. However, the large number in this interval in our case is again purely random statistical variation and must be accepted as experimentally valid.

The point of the above discussion is twofold. First, when dealing with statistical mechanisms, it is crucial to have large sample sizes. Otherwise, one may not be able to differentiate between experimental data issues and purely random statistical variation for those points that appear to be far from the expected distributions. The second point is that data points should only be removed from the sample when there are very clear experimental reasons to justify it.

When the time interval approaches zero in the limit, the above discussion applies to continuous functions as well. If the discrete function is well behaved, that discrete function may be represented by a continuous distribution. The continuous PDF, which was used to generate the simulation shown in Figure 1.6, is given in Figure 1.7. Five such distributions were simulated and averaged together and are plotted in Figure 1.8 after normalization. Note that the distribution parameters were identical in all five of the simulations used to generate Figure 1.8, but each individual simulation of 50 points varied significantly from the curve shown in Figure 1.7. The combination of the five simulations of 50 points shown in Figure 1.8 after normalization, much more closely resembles the continuous distribution shown in Figure 1.7. The difference between Figures 1.6 and 1.8 is due purely to random statistical variation.

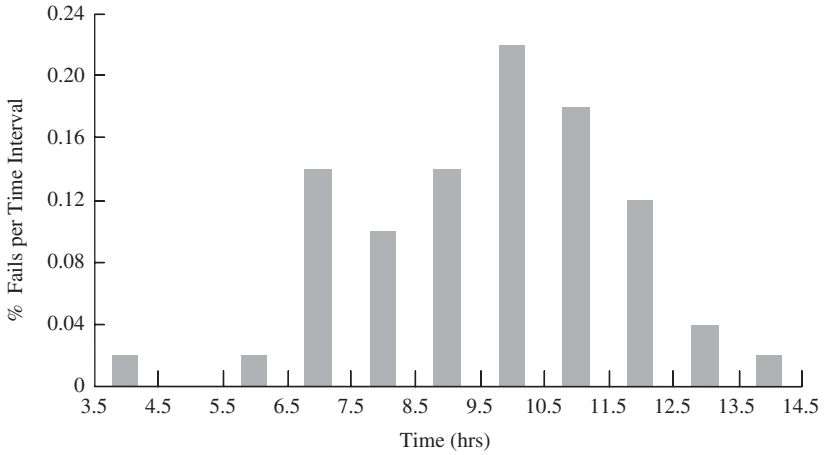

Figure 1.6. A histogram corresponding to a sample from a normal distribution with $\mu = 10$ and $\sigma = 2$; one simulation with 50 points.

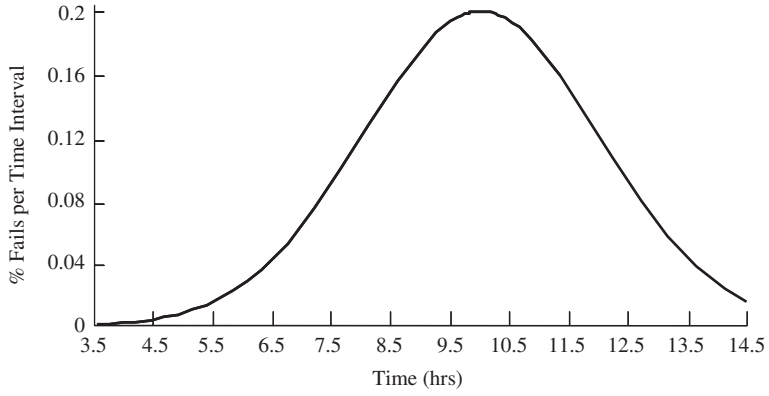

Figure 1.7. Continuous normal distribution with $\mu = 10$ and $\sigma = 2$.

Figure 1.9 shows a simulation using the same starting distribution but with 250 data points normalized to 50 data points instead of just a simulation of 50 data points. Here there is still significant deviation from the normal distribution. This is most obvious in the 8.5–9.5 and 10.5–11.5 intervals. Figures 1.8 and 1.9 show deviations because the sampling is still limited. Figure 1.10 shows that same simulated distribution but now with 1000 data points and again normalized. This curve is now very true to a normal distribution especially at the values close to the mean. The shape of all of the curves in Figures 1.6–1.10 are the classic "bell curves." But if only 50 points or less are plotted, serious deviation from the pure normal distribution should be expected.

Mathematically, one can allow the subregions to become smaller and smaller, and to in fact approach zero. This is the case shown in Figure 1.7 where the PDF becomes the continuous function, $f(t)$ and is defined as the probability that the

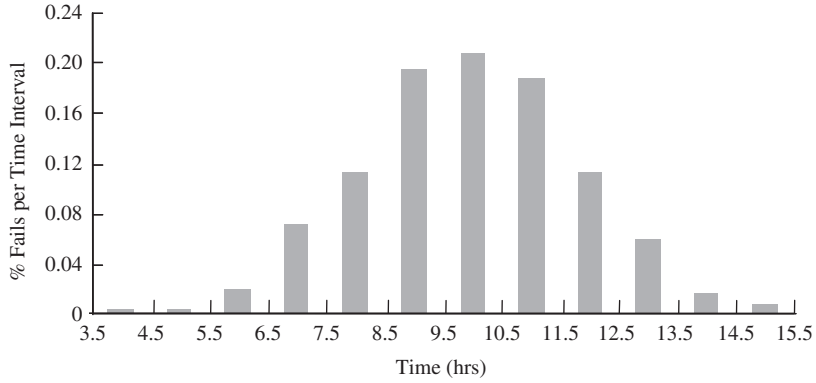

Figure 1.8. A histogram corresponding to a sample from a normal distribution with $\mu = 10$ and $\sigma = 2$ for 5 simulations each with 50 points after normalization.

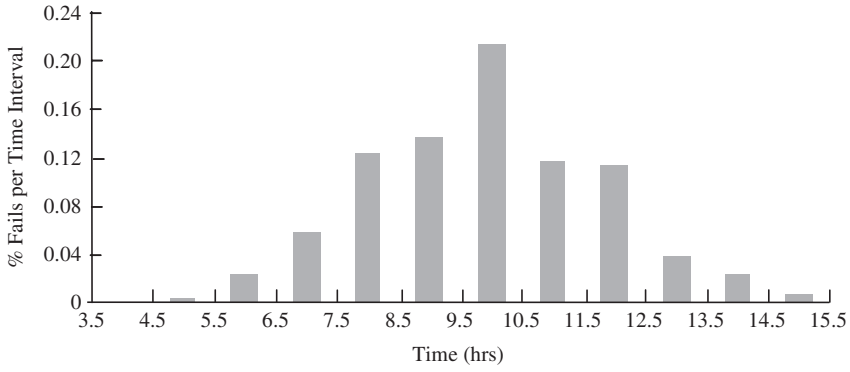

Figure 1.9. A histogram corresponding to a sample from a normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 250 points normalized to 50 points.

values of t lie between $(t-0.5dt)$ and $(t+0.5dt)$, where $0 < t < \infty$. The equation for the PDF for the normal distribution is given Equation 1.1. Note in the general case the mean may have any value whereas the standard deviation must have a positive value.

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(t-\mu)^2/2\sigma^2\right] \quad \text{where } 0 < t < \infty \tag{1.1}$$

where μ is the mean time to fail and σ is the standard deviation.

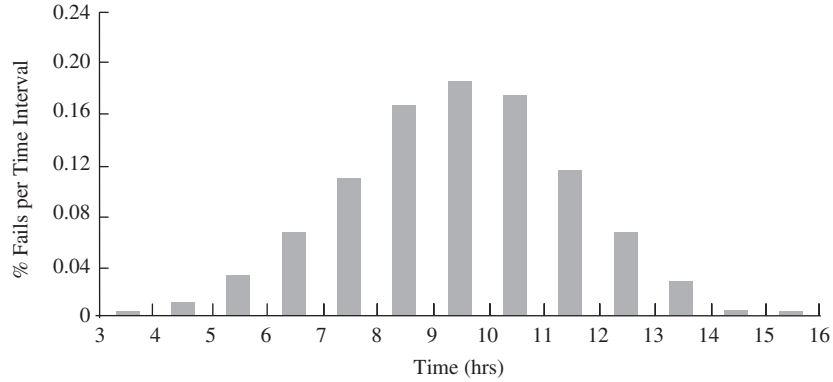

Figure 1.10. A histogram corresponding to a sample from a normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 1000 points normalized to 50 points.

1.4.5.4 Cumulative Distribution Function (CDF). The cumulative distribution function (CDF) is the cumulative sum from each subregion of the failing population for a discrete function. The CDF, or $F(t)$, is that fraction of the population that has failed by time t. Stated another way, $F(t)$ is the probability that a part will fail by time t. Hence, as one adds the failures from each time interval, the CDF increases from zero to one. The CDF for a continuous function is then simply the integral of the PDF.

The CDF will typically be the first plot that the reliability engineer makes after taking the experimental data. Notice that for the CDF, only the fails are plotted. This has the important ramification that only the failing samples will be used to determine the distribution. If the reliability engineer has designed an experiment with 1000 samples for a given time duration, and after the end of that time, only five samples have failed, only five points can be plotted and only the tail of the distribution will be displayed. The remaining 995 samples are censored before failure occurs. As we shall see later, a large variation would be expected in the results at the tail of a distribution due to large confidence bounds. Hence, very little about the sample distribution could be expected to be accurately determined in that case even though a large sample was stressed. At the other extreme, if only five samples were stressed and all failed, there would again be only five fails to plot but that would represent the entire CDF for this sample. Now because the total sample size was so small, the confidence bounds would be very large, and the distribution parameters of the population would have a such wide range of possible values for a given set of confidence bounds that little would be known about the actual distribution of the population, even though it is known for our sample of five.

The CDFs that are shown in Figures 1.11 and 1.12 are related to the discrete and continuous PDFs given in Figures 1.6 and 1.7, respectively. We need to highlight that these are empirical or sample CDFs and not those for the population. As expected, since these data were generated by a simulation of a

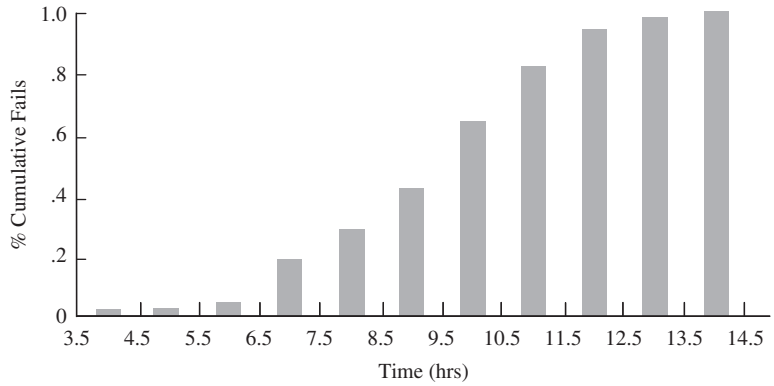

Figure 1.11. CDF for discrete normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 50 points.

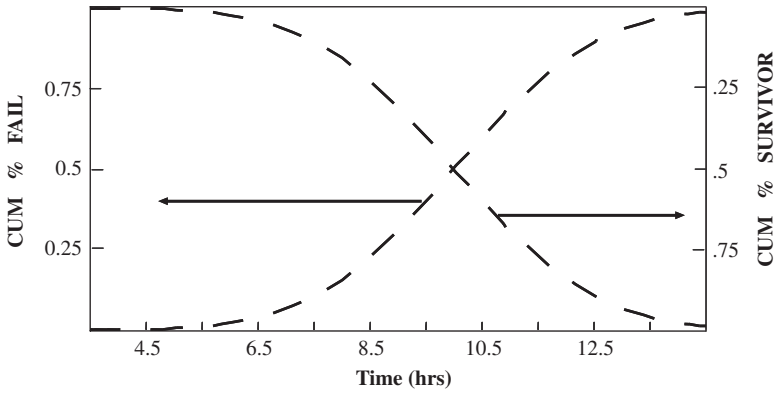

Figure 1.12. CDF and reliability function for continuous normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 50 points.

normal distribution with a mean of 10, half of the population has failed by 10 hours. Equivalently, the probability of failure by 10 hours is 50%. Expressed mathematically, the probability of fail at a given time is the area expressed by

$$F(t) = \int_0^t f(x)dx \tag{1.2}$$

From the PDF the most general equation for the CDF is given as Equation 1.3 below. The reliability subset of this equation limits t to the positive time values.

$$F(x) = \int_0^t \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x-\mu)^2/2\sigma^2\right]dx = \Phi[(t-\mu)/\sigma] \tag{1.3}$$

where Φ is the normal distribution function and $0 < t < \infty$. The probability that t lies within the range from 0 to $+\infty$, is equal to one since in our case that range includes all possible values of t. Hence, the value of the integral must be 1 representing 100% failure.

1.4.5.5 Reliability Function R(t). Whereas the CDF is the cumulative failing population, the reliability function, R, is the cumulative surviving population. It is obtained by subtracting the cumulative fails (CDF) from 1, i.e., $R(t) = 1 - F(t)$. Just as the CDF must equal one after the last fail, the reliability function must equal 0 since there are by definition no survivors. The equivalent continuous reliability distribution is also plotted in Figure 1.12 where the reliability scale is on the right. Obviously for the R, the point at which 50% of the population has survived is 10 hours or the mean of the normal distribution.

1.4.5.6 Instantaneous Failure Rate (IFR) or Hazard Function, h(t). Another important statistical concept that is used within the reliability community is that of the instantaneous failure rate (IFR) or hazard function $h(t)$. The $h(t)$ is

defined as the probability that those parts that have survived until a time, t, will fail in the next increment of time, Δt. The $h(t)$ is then a probability divided by the incremental time Δt which converts it to a rate. Hence, $h(t)$ is a failure rate expressed either in terms of fails per time or fraction failing per time. The $h(t)$ or hazard function is defined mathematically as

$$h(t) = f(t)/[1 - F(t)] = f(t)/R(t) \tag{1.4}$$

Derivations for the $h(t)$ are given in references [12] and [5]. Those derivations are beyond the scope of this book although the reader may find that a review of those derivations may be helpful in better understanding the hazard function.

This figure of merit is most useful in describing the case of fallout during the normal lifetime of the product. That fallout is also known as extrinsic fallout or defect fallout. The $h(t)$ can be integrated in a manner similar to the PDF or $f(t)$ to obtain $H(t)$ sometimes referred to as the cumulative hazard function. For more information on the cumulative hazard function than given below, the interested reader is directed to the books previously referenced and especially reference [10].

The discrete $h(t)$ related to Figures 1.6 and 1.11 is shown in Figure 1.13 and the continuous $h(t)$ related to Figures 1.7 and 1.12 is shown in Figure 1.14. For the case of this normal distribution, the $h(t)$ is an increasing function. The discrete $h(t)$ shown in Figure 1.13 has a general similarity to the continuous $h(t)$ in Figure 1.14 but because of random statistical variation is far from identical.

Because the CDF is the summation or integral of the PDF, the statistical variation can appear to be mitigated. This could be the conclusion of a casual comparison of Figures 1.6, 1.11, and 1.13 as compared to Figures 1.7, 1.12, 1.14 respectively. However, a careful inspection will still reveal the variation.

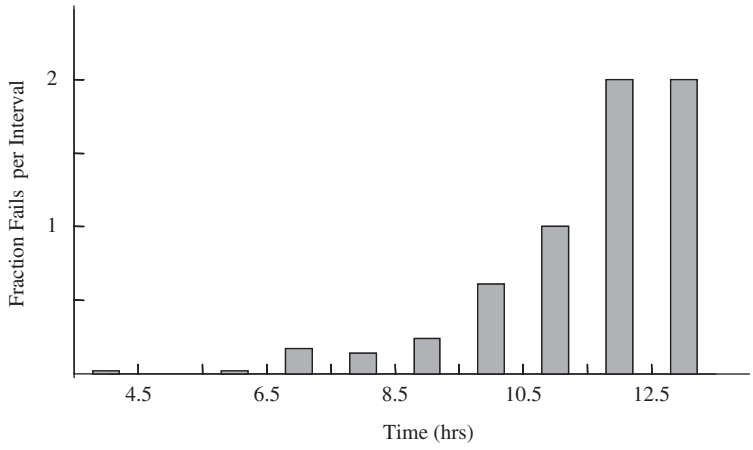

Figure 1.13. Hazard function of discrete normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 50 points.

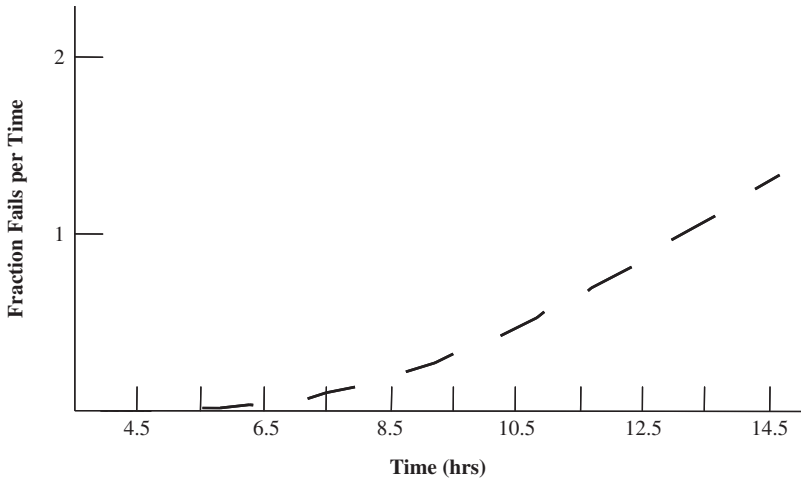

Figure 1.14. Hazard function of continuous normal distribution with $\mu = 10$ and $\sigma = 2$ for a single simulation with 50 points.

1.4.5.7 Cumulative Hazard, H(t). The cumulative hazard function is the integral of the hazard function as shown in Equation 1.5 and is another function that can provide useful insight. The cumulative hazard is directly related to the CDF as shown in Equations 1.4 and 1.5. The $H(t)$ can be more convenient for some types of censored data when plotting manually than is the CDF. However, for semiconductor technology reliability and in today's computer world, it is rarely used and is presented here for completeness.

$$H(t) = \int_0^t h(x)dx \tag{1.5}$$

Another potentially useful relationship is $H(t) = -\ln R(t)$. This equation may be verified by taking the derivative of both sides to obtain Equation 1.5. This relationship may be expressed in several different forms.

1.4.5.8 Average Failure Rate (AFR). The average failure rate (AFR) is the total number of fails within a given time interval expressed as a rate. This is a useful figure of merit because the time interval can be defined and then a single number used to characterize the reliability. Note that the hazard function is variable so the value depends on the chosen time. One could get the same value of AFR for equivalent product from two vendors using the hazard function, but the shape of the curves in arriving at that value might cause one product to be acceptable and the other one unacceptable. Because the AFR is an average, it is a simpler figure of merit and is an acceptable figure of merit in many cases. It is typically the figure of merit of choice to compare failure rates after one year and sometimes after 5 or 10 years as well. The one-year AFR is not necessarily set

exactly at one year or 8.76 khr, sometimes it is approximated as 10 khr. The AFR is just the integral of $h(t)$ between a specific time interval, divided by that time interval. Mathematically it is represented as

$$AFR(t_1, t_2) = \left(\frac{1}{t_2 - t_1}\right) \int_{t_1}^{t_2} h(t)dt. \tag{1.6}$$

1.4.5.9 Moments. Although moments of a probability density function may be defined that will characterize a PDF, the accepted practice in reliability engineering is to use the distribution attributes themselves, such as the characteristic life and shape factor for a Weibull distribution, rather than the moments of the PDFs. A given moment may be described as the relationship of every value of $f(x)$ with respect to some fixed value.

A detailed discussion of the moments of distributions is beyond the scope of this book and is generally not necessary for reliability engineering. However, the interested reader is directed to Green and Bourne [12], for a rather complete discussion on the various moments for the major distributions. The moments of most interest in semiconductor reliability are 1) the first moment about the origin which is the mean for the normal distribution and the characteristic life for the exponential distribution; and 2) the second moment about the mean which is the variance for the normal distribution and the square of the characteristic life for the exponential distribution.

First moment about some constant x_0:

$$M_1 = \int_{-\infty}^{\infty} (x - x_0) f(x)dx. \tag{1.7}$$

$$M\text{th moment:} \quad M_m = \int_{-\infty}^{\infty} (x - x_0)^m f(x)dx. \tag{1.8}$$

Then the first moment about the origin ($x_0 = 0$) is the mean:

$$M_1 = \int_{-\infty}^{\infty} xf(x)dx. \tag{1.9}$$

The second moment about the mean is the variance:

$$M_2 = \int_{-\infty}^{\infty} (x - M_1)^2 f(x)dx. \tag{1.10}$$

1.4.5.10 Fractile or Quantile Function. The p fractile, or equivalently the p quantile, is the time at which that fraction, in percent, of the sample fails. Hence, $p = 5\%$ represents the 5% failure point of the sample and $p = 50\%$ represents the 50% failure point, and $p = F(t_f)$. It is sometimes the case that the time position of the fifth, tenth, or twentieth percentile of the lifetime distribution is specified as the fail criterion. This is especially true when the tails of the population are the

primary cause of failure. The time to fail of the p fractile is then just $t_f = F^{-1}(p)$ or the inverse of the CDF. Thus for the example in Figure 1.12, the time of fail for the 25 fractile or 25% failing portion of the sample is about 8.75 hrs. Note that the 25, 50, and 75 fractiles or quantiles are also known as the quartiles.

1.4.5.11 Reliability Projections and Closure. The normal distribution was used to introduce the various reliability functions because everyone typically has had either direct or indirect experience with the normal or Gaussian distribution.

One final concept is needed before moving to the other distributions that will be commonly used in reliability statistics. That concept is the one of transforming the distribution of choice in such a manner that the CDF is linear when plotted.

The CDF plot for a continuous normal distribution was given in Figure 1.12. However, one could not make any graphical projections using that chart since the plot is not a straight line. The axis needs to be modified to achieve a straight line. Alternatively the equivalent set of equations could be numerically solved to make the projections. In practice both are typically done. A computer is used to solve equations, fit data, and do all of the calculations, and it is also used to plot the results on the appropriate axis so that one may have a visual demonstration of the solution and of the projection.

Before continuing the endeavor to make the CDF of the normal distribution a linear plot on some axis, the concept of the standard normal distribution needs to be considered. Equation 1.3 is not soluble in closed form. Originally, tabular values were used to solve these equations. A standard normal distribution was introduced so that only one set of values would be necessary and all variations of a normal distribution could be converted to that standard normal distribution. In today's world of computers and numerical integration, this procedure would not be necessary, but it is instructive and should give the reader a better intuitive understanding. Given a distribution where x is the random variable of a normal distribution that has a mean of μ and standard deviation of σ, the transformation to the standard normal distribution is given by

$$z = (x - \mu)/\sigma. \tag{1.11}$$

With this transformation and noting that for a standard normal distribution $\sigma = 1$ by definition, the PDF given in Equation 1.1 and the CDF given in Equation 1.3 become

$$\text{PDF: } \phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] \quad \text{where } -\infty < z < \infty \tag{1.12}$$

and

$$\text{CDF: } \Phi(z) = \int_{-\infty}^{z} \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-w^2}{2}\right)\right] dw \quad \text{where } -\infty < z < \infty. \tag{1.13}$$

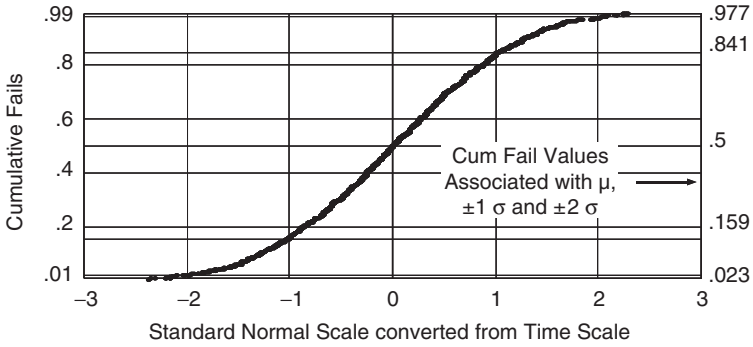

Figure 1.15. Plot of an empirical CDF of 1000 points $\mu = 0$ and $\sigma = 1$with "typical" axis.

As mentioned above, historically the values for the standard normal CDF were tabulated and all normal distributions converted to the standard normal distribution for integration based on that tabulation. Most, if not all statistics books, contain tables showing the solution of Equation 1.13, including the referenced statistics books. Plots of an empirical, normal CDF are shown in Figures 1.15 and 1.16 for a standard normal distribution for the 1000 point experiment with the points generated by simulation assuming the standard normal distribution. Note that for the standard normal distribution, the PDF is symmetric about zero so that indeed $\mu = 0$ and hence, at zero, half of the population will have failed.

We return to the second part of the normal CDF in Equation 1.3 to obtain the y-axis transformation for the normal plot that will yield a linear CDF plot if the distribution is normal. Mathematically, this may be expressed from Equation 1.3

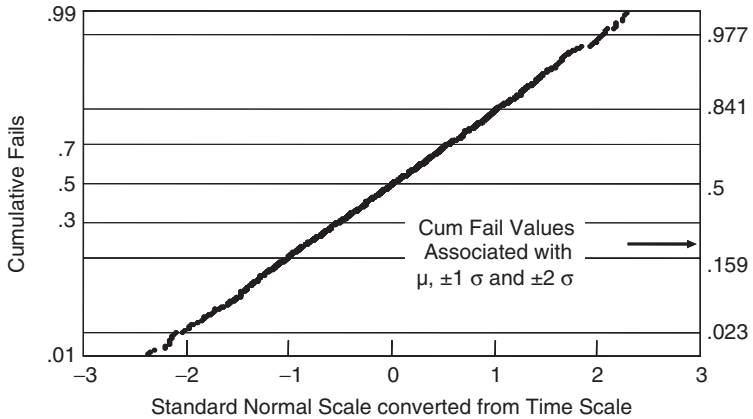

Figure 1.16. Plot of an empirical CDF of 1000 points $\mu = 0$ and $\sigma = 1$with axis transformation to yield a linear CDF plot if it fits a normal distribution.

for the normal CDF as $F(x_p) = \Phi(\{x_p - \mu\}/\sigma)$. Solving this equation for x_p by taking the inverse of the function $F(x_p)$, one obtains $x_p = \mu + \sigma\ \Phi^{-1}(x_p)$. If x_p is then plotted versus $\Phi^{-1}(x_p)$, a straight line will result if the distribution in question can be represented with a normal distribution. Note that each of the quantile pairs is plotted equidistant from the mean on this new axis. Thus $+/-\sigma$ points are equal distant from the mean as are the 0.01 and 0.99 points whereas the distance on the y axis between 0.01 and 0.023 is nearly equal to that between 0.3 and 0.5. Note that the right y-axis scale could have chosen to show the equivalent σ value instead of the actual cum fail percent. Given this choice of percentile values, the right y-axis equivalent values are 2, 1, 0, -1, and -2 σ, respectively, from top to bottom. Thus Figure 1.16 has an axis that transforms the CDF of a normal distribution into a linear function.

Note the power of this transformation. The CDF curve is now linear since the data follows a normal distribution and while the tails of the distribution vary from the straight-line projection, one could project from the data back to the time at which 1 ppm or 0.0001% would be predicted to fallout, or to any other desired fallout. Obviously this projection assumes that no new mechanisms were encountered as per the assumptions previously stated in Sections 1.2.2 and 1.4.2.

It is instructive to observe these two plots when only 50 experimental data points are available. We have already discussed the importance of sample size for at least statistical mechanisms, but since a picture is worth a thousand words, the picture is shown below. Clearly the slope and mean of the curve for Figure 1.16 are known to a much higher confidence than that for Figure 1.17, or equivalently Figure 1.18, which has the axis transformation.

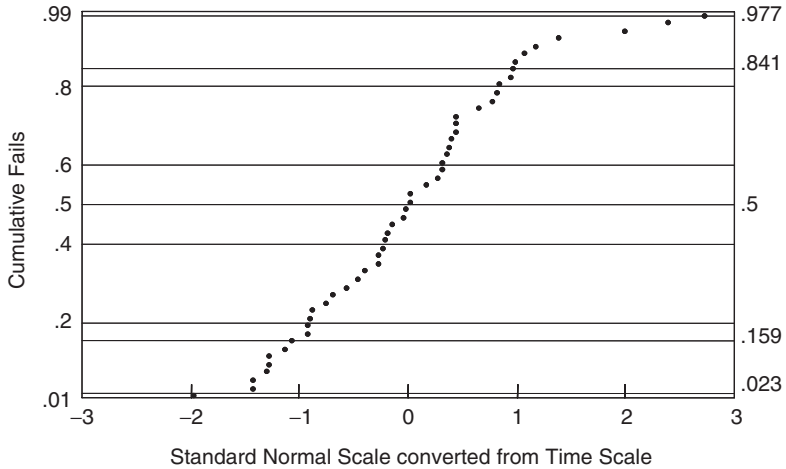

Figure 1.17. Plot of an empirical CDF for normal distribution of 50 points for $\mu = 0$ and $\sigma = 1$ with "typical" axis.

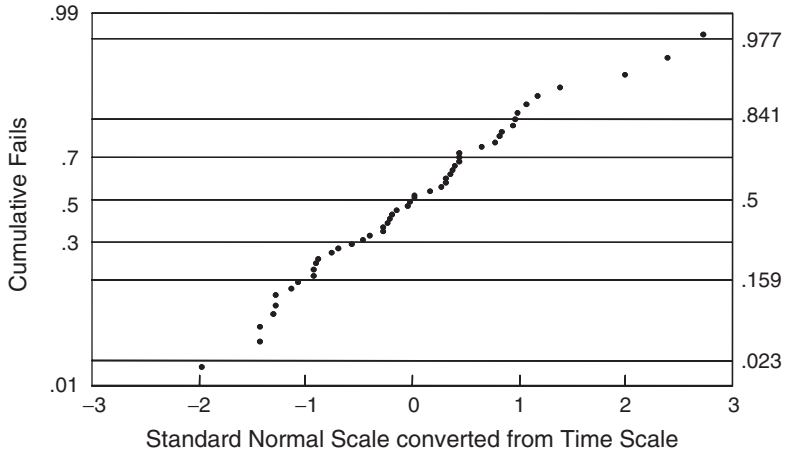

Figure 1.18. Plot of an empirical CDF for normal distribution of 50 points $\mu = 0$ and $\sigma = 1$ with axis transformation to yield a linear CDF plot if it fits a normal distribution.

The reliability concepts have been introduced using the normal distribution since most are familiar with it and its bell-shaped PDF. The normal distribution will be discussed further in terms of the lognormal distribution in Section 1.4.8 below as well as in the electromigration chapter in Sections 7.4.2 and 7.4.3. We will now move to the distributions that will be most commonly used in the remainder of this book. Those distributions are primarily the lognormal distribution and the Weibull distribution which are both two-parameter distributions as typically used in semiconductor reliability, although a third parameter could be introduced for either. It has in fact been shown that a third parameter for the lognormal distribution can more effectively describe electromigration behavior because it can be used to model an incubation period [21]. The cost is a much larger sample size to determine all three parameters than that required to determine just two parameters so that the more complex model has historically rarely been used. As features continue to shrink it may become advantageous to move to a three-parameter distribution to gain the additional accuracy that could provide additional reliability margin.

1.4.6 Exponential Distribution

The exponential distribution is a single-valued distribution and has the very important attribute that $h(t)$, the instantaneous failure rate, is a constant The PDF, CDF, and $h(t)$ are given in Equations 1.14, 1.15, and 1.16, respectively, where time, $t \geq 0$.

$$\text{PDF: } f(t) = \lambda \exp(-\lambda t) \tag{1.14}$$

$$\text{CDF: } F(t) = \int_0^t f(x)dx = 1 - \exp(-\lambda t) \tag{1.15}$$

$$\text{IFR: } h(t) = f(t)/[1 - F(t)] = \lambda \tag{1.16}$$

Because $h(t)$ for the exponential distribution is a constant, the probability of failure in the next time increment is independent of the length of time the product has already been in use or on stress. The ramification of this is that the part has no aging that impacts its failure rate. A new part has the same failure rate as does the part that is still functioning having already survived many years of operation. Another way of stating this is that the part has no memory of its past operation.

The first moment about the origin is $1/\lambda$ and the second moment about the mean is $1/\lambda^2$. These moments are given for comparison to the normal distribution and are not frequently used in semiconductor technology reliability.

We show a set of curves for the PDF, CDF, and $h(t)$ for the exponential distribution where λ has the values of 0.5, 1, and 2. Figures 1.19 and 1.22 depict the PDF and $h(t)$, respectively. The exponential CDFs are plotted in Figures 1.20 and 1.21.

Note that the vertical axis has been modified in Figure 1.21 so that the CDF for the exponential distribution will plot linearly. The procedure here is much simpler than for the normal distribution, although in principle it is similar. Equation 1.15 is solved for λt yielding $-\ln(1-F(t))$. The scale transformation is possibly most recognizable for $\lambda = 1$, since at $t = 1$, the exponential argument value is 1 and the cumulative percent failed is 63%.

Thus far we have been discussing the exponential distribution in terms of a single parameter. The exponential distribution can also be utilized with two parameters and the PDF, CDF, and $h(t)$ are shown in Equations 1.17–1.19 for

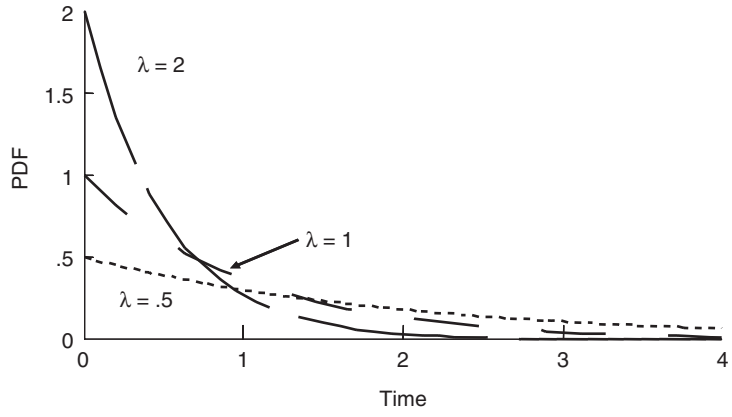

Figure 1.19. PDF of exponential distribution for three values of the distribution parameter, λ, plotted on linear axis.

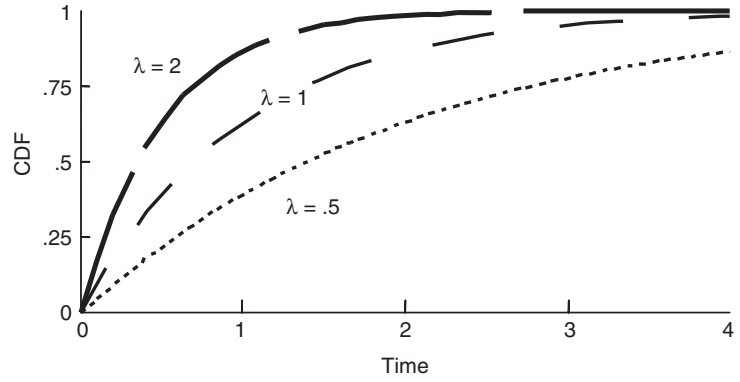

Figure 1.20. CDF of exponential distribution for three values of the distribution parameter, λ, plotted on linear axis.

that case. The second distribution parameter, t_0, would typically be a timescale shift in the case of semiconductor reliability although in some cases its interpretation might more appropriately be a threshold parameter.

$$\text{PDF: } f(t) = \lambda \exp[-\lambda(t - t_0)] \tag{1.17}$$

$$\text{CDF: } F(t) = \int_0^t f(x)dx = 1 - \exp[-\lambda(t - t_0)] \tag{1.18}$$

$$\text{IFR: } h(t) = f(t)/[1 - F(t)] = \lambda \tag{1.19}$$

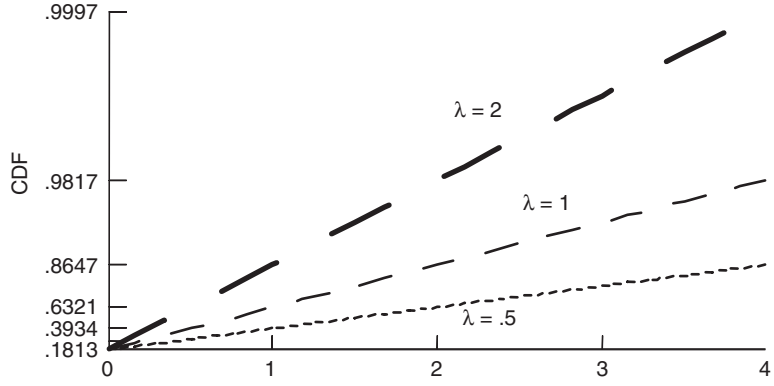

Figure 1.21. CDF of exponential distribution for three values of the distribution parameter, λ, a plotted on a transformed axis to yield linear CDFs.

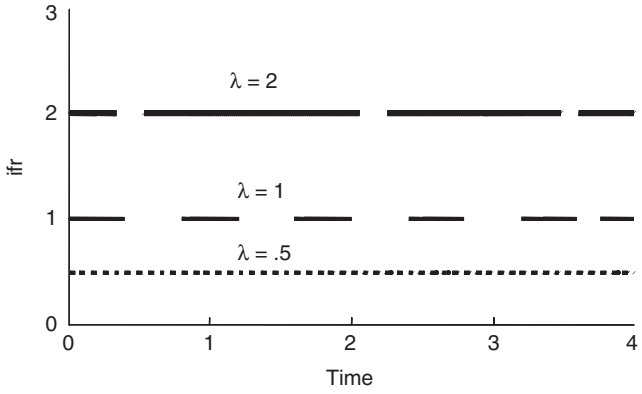

Figure 1.22. $h(t)$ of exponential distribution for three values of the distribution parameter, λ, plotted on linear axis.

1.4.7 Smallest Extreme Value and Weibull Distributions

The smallest extreme value distribution is included primarily to demonstrate its relationship to the Weibull distribution; however, it does have some application in its own right. These applications are not in general use in semiconductor reliability today and will not be discussed in this book. The smallest extreme distribution is applicable for $-\infty < x < \infty$ where κ is the location parameter and may be positive or negative and ζ is the scale parameter which must be positive. As in the Weibull distribution, 63.2% of the population has failed at the value $F(x = \kappa)$.

$$\text{PDF: } f(x) = \frac{1}{\zeta} \exp\left[\frac{x-\kappa}{\zeta}\right] \exp\left[-\exp\left(\frac{x-\kappa}{\zeta}\right)\right] \tag{1.20}$$

$$\text{CDF: } F(x) = 1 - \exp\left[-\exp\frac{x-\kappa}{\zeta}\right] \tag{1.21}$$

$$\text{IFR: } h(x) = (1/\zeta)\exp[x - \kappa/\zeta] \tag{1.22}$$

To obtain the Weibull CDF one defines $x = \ln y$ and $\kappa = \ln \varphi$, and $\zeta = 1/\beta$, then substituting in Equation 1.21 one obtains

$$F(y) = 1 - \exp[-\exp(\ln y - \ln \theta/1/\beta)] = 1 - \exp\left[-(y/\varphi)^{\beta}\right], \tag{1.23}$$

where y is now constrained to be $y > 0$.

The result in Equation 1.23 is identical to the Weibull CDF given in Equation 1.25. It is left as a student exercise to show that this special case of the smallest extreme value function reduces to the Weibull distribution for the PDF and $h(t)$. A more complete discussion is given by Nelson [10].

The Weibull distribution, as a two-parameter distribution, has a characteristic life, φ, and a shape parameter, β. The PDF, CDF, and $h(t)$ are given in Equations 1.24, 1.25, and 1.26, respectively.

$$\text{PDF: } f(t) = (\beta/t)(t/\varphi)^{\beta} \exp\left[-(t/\varphi)^{\beta}\right] \tag{1.24}$$

$$\text{CDF: } F(t) = \int_0^t f(x)dx = 1 - \exp\left[-(t/\varphi)^{\beta}\right] \tag{1.25}$$

$$\text{IFR: } h(t) = f(t)/[1 - F(t)] = (\beta/t)(t/\varphi)^{\beta} \tag{1.26}$$

Obviously for semiconductor reliability purposes, these equations are applicable only for $t \geq 0$.

Notice first the case where the shape parameter is equal to 1, that is $\beta = 1$. This is the case in which the Weibull distribution collapses to the exponential distribution. In this case, Equations 1.24–1.26 reduce to Equations 1.14–1.16, respectively, where $\lambda = 1/\varphi$.

The next observation gives the characteristic life its meaning. Substituting φ for t in the CDF, we get $F(t) = 1-\exp(-[\varphi/\varphi]^{\beta}) = 1-e^{-1} = 0.632$. Hence, 63.2% of the population fails by the characteristic life, φ, independent of the shape parameter β. This makes the characteristic life a powerful figure of merit for the Weibull distribution. Therefore when discussing distributions which follow a Weibull distribution, it is the 63.2 percentile that is the desired figure of merit, not the fallout at the 50 percentile as is the appropriate figure of merit for the normal distribution.

The shape parameter, β, gives the slope of the Weibull distribution. Usually in reliability projections, the slope has a larger impact on the final projection than the characteristic life since the projections are typically extrapolated across several or even many orders of magnitude. Any error in the shape parameter is then magnified by that extrapolation.

Many reliability systems can be and are modeled using a Weibull distribution. As the figures depicting the Weibull distribution below will show, it is a very flexible distribution. This flexibility is one of the reasons it is very useful for reliability engineers. The Weibull distribution is the distribution of choice for systems that have many, identical competing elements that can each cause a fail and for which the first element to fail causes the entire system to fail. This is discussed briefly in Section 1.4.8 and more fully in the dielectric chapters, Sections 2.1.2 and 2.4. Several authors have discussed the theoretical justification for using the Weibull distribution and how it follows from the extreme value theory when used in conjunction with the weakest link model [22–25].

The first moment about the origin is $\varphi\Gamma\{(1+\beta)/\beta\}$ for a Weibull distribution. The second moment about the mean for a Weibull distribution is $\varphi^2\Gamma\{(2+\beta)/\beta\}-\Gamma^2\{(1+\beta)/\beta\}]$ where Γ is the gamma function. These moments are not

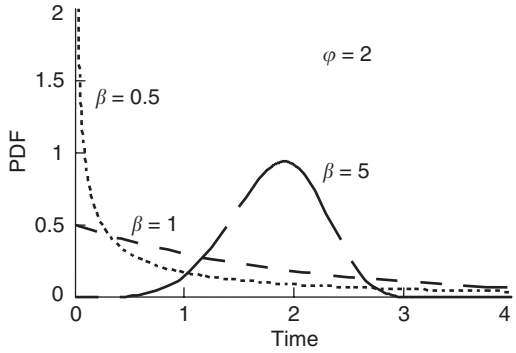

Figure 1.23. PDF for Weibull distribution with three values of the shape parameter β and a characteristic life, $\varphi = 2$.

frequently used in semiconductor reliability work but are given here for completeness and for comparison to the normal distribution.

The next topic for the Weibull discussion is the conversion of the axes for the CDF plot to those that will result in a straight line plot for the Weibull CDF. Using the same general procedure as for the exponential linearization Equation 1.25 is solved for "t" by subtraction and taking the natural logarithm twice obtaining

$$\ln[-\ln[1 - F(t)]] = \beta \ln t - \beta \ln \varphi. \tag{1.27}$$

The term on the left has been defined as a Weibit, W, where $W = \ln(-\ln(1-F(t))$. If we then plot W versus $\ln t$ on the x axis, a linear plot for the CDF is achieved if the distribution follows a Weibull distribution.

Figures 1.23, 1.24, 1.26 are plotted on a linear scale and show the PDF, CDF and $h(t)$, respectively, for the Weibull distribution. For Figure 1.25 the Weibull CDF is plotted on a scale having Weibits as the y axis and $\ln t$ as the x axis.

Notice that in Figures 1.24 and 1.25, the CDF plots for Weibull distributions having a characteristic life of two, the cumulative failure for each of the Weibull distributions is 63.2% at that characteristic life of two, regardless of the shape parameter. Also note that widely differing distributions occur for the shape factors chosen.

The $h(t)$ for a shape parameter, $\beta = 1$ is a constant as discussed above and is demonstrated in Figure 1.26. Returning to Figure 1.23, note the drastic difference in the PDF for Weibull distributions having shape parameters of 0.5, 1, and 5. For Weibull distributions having values of β less than 1, the PDF starts high at time zero and is a decreasing function. This is typical of a defect mechanism and indeed, historically, for oxides thicker than about 5 nm, $\beta \leq 1$, was typically interpreted as extrinsic or defect fallout while for values of $\beta > 1$, the interpretation was that the Weibull distribution represented the intrinsic fallout or wearout. This is a very

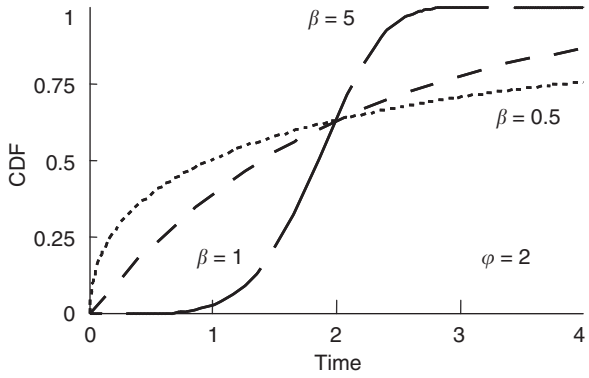

Figure 1.24. CDF for Weibull distribution for three values of the shape parameter β and for a characteristic life $\varphi = 2$.

graphic example of why the Weibull distribution is so useful in reliability modeling. Many experiments can be successfully modeled using a Weibull distribution; however, it is most powerful when a theoretical reason exists for its use to describe a particular mechanism.

The three-parameter Weibull distribution is given in the Equations 1.28–1.31. Again, the third parameter would typically be a time-shift parameter for semiconductor reliability modeling although it could be cast as a threshold parameter especially in the more general cases. Note that as semiconductors are

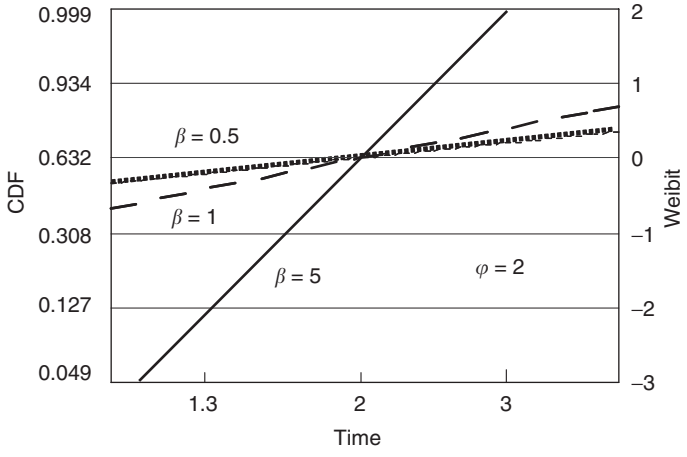

Figure 1.25. CDF for Weibull distribution for three values of the shape parameter β and for a characteristic life $\varphi = 2$ plotted on a transformed axis to yield a linear CDF plot.

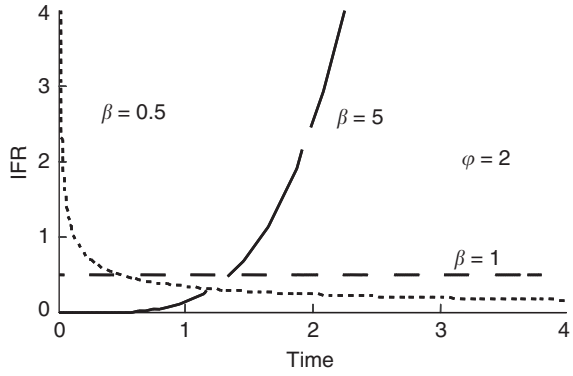

Figure 1.26. $h(t)$ for Weibull distribution with three values of the shape parameter β and a characteristic life, $\varphi = 2$.

driven to the limit, it may become necessary to include more terms in the modeling distributions to more accurately characterize the end of life predictions.

$$\text{PDF}: f(t) = (\beta/t - t_0)(t - t_0/\varphi)^{\beta} \exp\left[-(t - t_0/\varphi)^{\beta}\right] \tag{1.28}$$

This equation can be written somewhat more simply if it is recast as

$$\text{PDF}: f(t) = \left(\beta/\varphi^{\beta}\right)(t - t_0)^{\beta-1} \exp\left[-(t - t_0/\varphi)^{\beta}\right] \tag{1.29}$$

$$\text{CDF}: F(t) = \int_0^t f(x)dx = 1 - \exp\left[-(t - t_0/\varphi)^{\beta}\right] \tag{1.30}$$

$$\text{IFR}: h(t) = f(t)/[1 - F(t)] = \left(\beta/\varphi^{\beta}\right)(t - t_0)^{\beta-1} \tag{1.31}$$

1.4.8 Lognormal Distribution

The lognormal distribution, as a two-parameter distribution, has a median parameter, $\mu_{\ln}$, and a shape parameter, $\sigma_{\ln}$. $\mu_{\ln}$ is also called the log mean in some statistics texts. It is important to understand the relationship between the lognormal distribution and the normal distribution. The median parameter, $\mu_{\ln}$, is the mean of the natural log of the lifetime and the shape parameter, $\sigma_{\ln}$, is the standard deviation of the natural log of lifetime. Thus if Y is a random variable with a normal distribution having a mean of μ, and a standard deviation of σ, the lognormal distribution with a random variable of Z which would be derived from that normal distribution is given by $Z = \exp Y$. For this lognormal distribution the shape parameter $\sigma_{\ln}$, is equal to the normal standard deviation σ, and the median parameter is related to the normal mean by $\mu_{\ln} = \exp \mu$.

The PDF, and CDF equations for the lognormal distribution are shown in Equations 1.32 and 1.33 where $0 < t < \infty$.

$$\text{PDF: } f(t) = \left(\sigma_{\ln} t\sqrt{2\pi}\right)^{-1} \exp\left[-((\ln t - \ln \mu_{\ln})^2/2\sigma_{\ln}^2)\right] \tag{1.32}$$

$$\text{CDF: } F(t) = \int_0^t \left[\left(\sigma_{\ln} x\sqrt{2\pi}\right)^{-1} \exp\left[(-(\ln x - \ln \mu_{\ln})^2/2\sigma_{\ln}^2)\right]\right] dx \tag{1.33}$$

The median parameter above is designated as '$\mu_{\ln}$' to emphasize the relationship to the normal distributions but it is often designated as T_{50}. The $h(t)$ would be numerically calculated and is given by the usual $h(t) = f(t)/[1-F(t)]$ formula. Note that the hazard rate of the lognormal is not monotonic as per Figure 1.30. The mean or first moment about the origin for a lognormal distribution is $\exp[\ln(\mu_{\ln}) + 0.5\sigma_{\ln}{}^2]$. The variance or second moment about the mean for a lognormal distribution is $\exp[2\ln(\mu_{\ln}) + \sigma_{\ln}{}^2]\{\exp(\sigma_{\ln}{}^2) - 1\}$. Again these moments are given only for comparative purposes as they are rarely used in semiconductor reliability.

Figures 1.27, 1.28 and 1.30 depict the lognormal PDF, CDF, and $h(t)$ respectively. Figure 1.29 depicts the CDF after the transformation to the linear scale. Observe that in Figures 1.27–1.30 the lognormal distribution also has a great deal of flexibility as to the shape of the distributions that it can model. Hence, the lognormal distribution can also be used to model a large number of phenomena. Also note that for the CDF, the median life time-to-fail, $\mu_{\ln}$, is independent of the shape parameter, $\sigma_{\ln}$.

The lognormal scale for the CDF plot may be transformed in a manner similar to the normal scale except now instead of starting with $F(t_q) = \Phi(\{t_q - \mu\}/\sigma)$ as for the normal distribution, we start with $F(t_q) = \Phi(\{\log(t_q) - \ln(\mu_{\ln})\}/\sigma_{\ln})$. Solving for t_q by taking the inverse of the function, $\ln(t_q) = \ln \mu_{\ln} + \sigma_{\ln} \Phi^{-1}(t_q)$ is

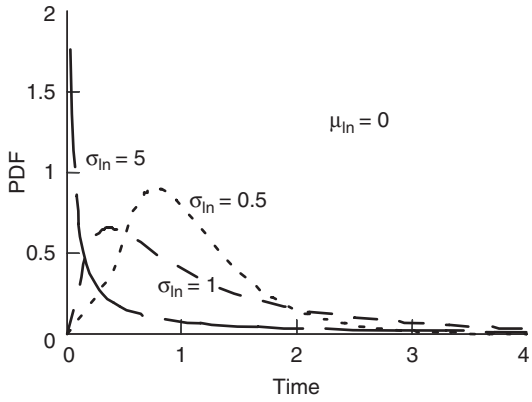

Figure 1.27. PDF for lognormal distribution with three values of the shape parameter $\sigma_{\ln}$ and a median parameter, $\mu_{\ln} = 0$.

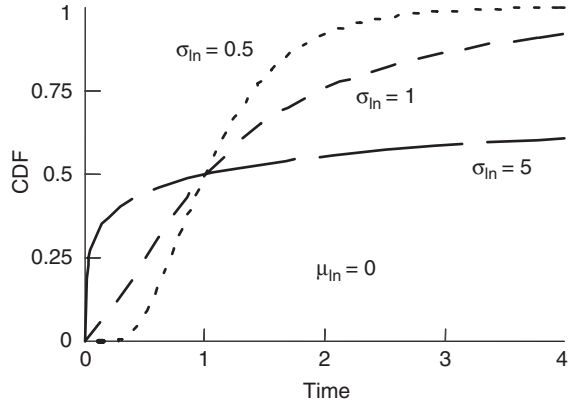

Figure 1.28. CDF for lognormal distribution with three values of the shape parameter $\sigma_{\ln}$ and a median parameter $\mu_{\ln} = 0$.

obtained. Hence, for the lognormal, a plot of $\ln(t_q)$ versus $\Phi^{-1}(t_q)$ will yield a straight line if the distribution can be modeled by a lognormal distribution. This is depicted in Figure 1.29.

Because of the flexibilities of both the lognormal distribution and the Weibull distribution it is often possible to model data using either distribution. However, typically data can only be collected across two to three orders of magnitude and projections must be made another two or three orders of magnitude in time beyond the data collection time. For example, a typical stress will last at most 100–1000 hrs and the product must last for 100 Khr. Also, the number of parts that can be stressed is usually very limited because of early hardware delivery limitations as well as stress facility limitations. The lack of stress time beyond a three month maximum

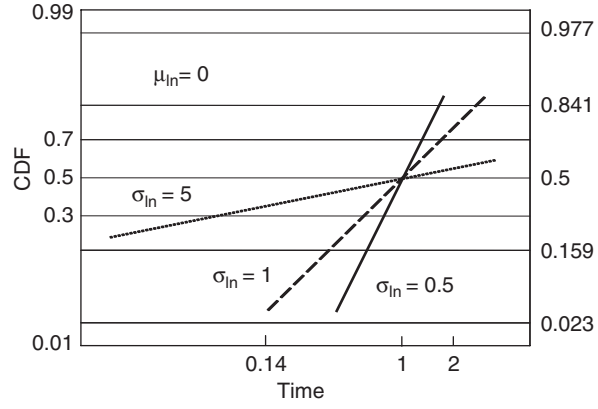

Figure 1.29. CDF for lognormal distribution with three values of the shape parameter $\sigma_{\ln}$ and a median parameter, $\mu_{\ln} = 0$ and with axis transformed to yield a linear CDF.

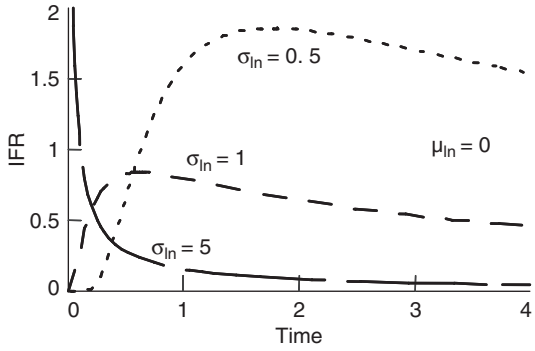

Figure 1.30. *h(t)* for lognormal distribution with three values of the shape parameter, $\sigma_{\ln}$, and a median parameter, $\mu_{\ln} = 0$.

precludes several serial stresses. Thus if only 50 parts are stressed, the first fail already represents the 2% fallout point. Based on the amount of data and the range of data, one may not be able to experimentally determine which distribution fits best and indeed, they may fit the data equally well. The results of projecting across the two to three orders of magnitude, however, could well mean the difference between a product that passes the requirement by an order of magnitude or more and one that fails the requirement by an order of magnitude or more. Recall that in addition to projecting across those orders of magnitude of time one is also usually projecting to ppm failure rates with the last data point at typically 1% or 2%.

For that reason, one prefers to appeal to a theoretical reason for choosing one distribution over the other whenever possible. It will be shown in Chapter 2, Section 2.4 that the Weibull distribution should be used when modeling phenomena that exhibit an extreme value behavior. Extreme value distributions are used to characterize systems where there are many competing mechanisms in parallel, and any one of which failing, can cause the whole system or product to fail. Dielectric breakdown is a prime example this behavior and the theory in Section 2.4 is applied to dielectric breakdown. The reader is advised to review that chapter as well as the aforementioned references to fully understand the theoretical reasons for choosing a Weibull distribution, to observe graphically the difference between choosing the lognormal and Weibull distributions, and finally to gain an appreciation for the sample size and time required to experimentally determine which distribution to use if no theoretical reason can be determined.

One basis for the use of a lognormal distribution is a proportional growth or multiplicative model [26, 27]. The model of failure in this instance is one where a shift, or a change, or a degradation starts ever so slowly, and then with time, that change grows or multiplies. Electromigration is typical, and will be modeled in this book, with a lognormal model. In the electromigration case, a void is formed, but that void growth starts with a single metallic atom exit. As the void grows from the molecular size, the volume surrounding the void increases and more atomic or molecular motion can contribute to the increasing void growth. Ultimately the void

becomes such a large percent of the line that the resistance increases to the point of failure. Crack growth and fatigue are also examples of mechanisms typically modeled by lognormal distributions as are diffusion and chemical reactive processes. Thus failures which result from small degradation processes that continue to grow until failure occurs are often best modeled by the lognormal distribution. As mentioned above, electromigration is caused by material transport and both the normal and lognormal distributions will be further developed in the support of the electromigration work in Section 7.4. The relationship of the material transport to the lognormal distribution will also be treated further there. Note that an understanding of the physical mechanism should provide significant direction as to the choice between the lognormal distribution model and the Weibull distribution model. In fact, based on the inability to take enough data to distinguish between the distribution models, the physical mechanism is typically a much better starting point to determine the correct model. The difficulty of choosing a model based on small amounts of empirical data will be clearly demonstrated in Section 2.4.3.

This introduces a further reason for understanding the physics of the mechanism in question that was not discussed in Section 1.3. Without a good understanding of the physics behind the mechanism, one might incorrectly choose a distribution with which to model the behavior and, as a consequence, be either radically optimistic or radically pessimistic without even knowing one's bias.

Sometimes it is useful to introduce a third parameter into the lognormal distribution to reflect either an incubation period, or more generally, an additional feature of the distribution. As mentioned above, this was proposed [21] for electromigration and was very successfully used to model an incubation period for electromigration and is discussed further in Section 7.4.5.

The PDF and CDF equations for the three-parameter lognormal distribution are shown in Equations 1.34 and 1.35 where $0 < t < \infty$.

$$\text{PDF: } f(t) = \left(\sigma_{\ln} t\sqrt{2\pi}\right)^{-1} \exp\left[-\left(\frac{[(\ln\, t - \mu_{\ln})/t_0]^2}{2\sigma_{\ln}^2}\right)\right] \tag{1.34}$$

$$\text{CDF: } F(t) = \int_0^t \left[\left(\sigma_{\ln} t\sqrt{2\pi}\right)^{-1} \exp\left[\frac{-[(\ln\, x - \mu_{\ln})/t_0]^2}{2\sigma_{\ln}^2}\right]\right] dx \tag{1.35}$$

The CDF for a three-parameter lognormal distribution is shown in Figure 1.31, where the third parameter is the location parameter which allows the direct modeling of an incubation period or of some threshold.

1.4.9 Poisson Distribution

The next distribution covered will be the Poisson distribution. It will be the distribution of choice for modeling some of the device negative bias temperature instability (NBTI) effects. More generally the Poisson distribution is often used to model recurrence data.

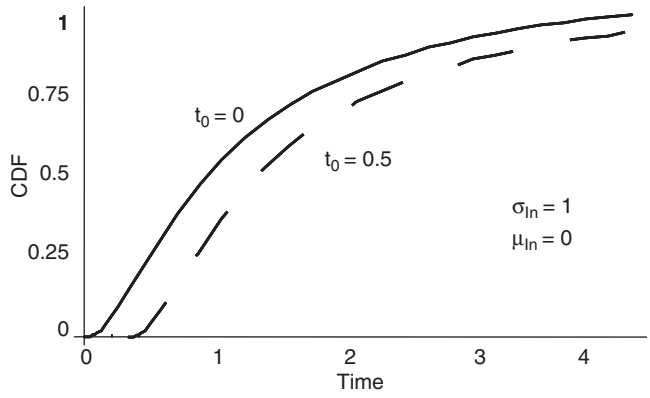

Figure 1.31. CDF for lognormal distribution having three distribution parameters with the shape parameter $\sigma_{ln} = 1$, the median parameter $\mu_{ln} = 0$, and location parameter $t_0 = 0$ and 0.5.

The Poisson distribution is very different from any of the distributions considered earlier because it belongs to the family of discrete distributions. Until now, we have modeled data with continuous distributions even though the data have been discrete. The discrete distribution gives a tool by which one can model the number of failures, or shifts, above a certain minimum shift or fail point.

The probability density function of a continuous distribution describes the fails in each period of time which can be plotted as a histogram with the time interval as the abscissa and the number of fails as the ordinate or y-axis value. A discrete distribution has an equivalent function that is called the probability function (*pf*). It consists of the probabilities of all of the occurrences that can occur in a given system. A simple discrete function is the geometric distribution and its probability function is given by $f(x) = p\ (1-p)^{(x-1)}$, where p is a probability and hence, $0 < p < 1$, x is any one of the total set of observations, and $f(x)$ is then the probability of that observation or outcome occurring.

The cumulative distribution function of a discrete function is the summation of the probabilities of all possible observations or outcomes and must equal one as x approaches $+\infty$, that is, as all of the probabilities of all possible outcomes are included in the summation. As x approaches $-\infty$, the cumulative distribution function must approach zero since the probability of none of the outcomes would be included in the summation. For the geometric and Poisson distributions, the observation or outcome factor must be an integer; however, this is not true of discrete distributions in general. A much more complete discussion of discrete distributions is given by Nelson [10].

The pf and CDF for the Poisson distribution are shown in Equations 1.36 and 1.37, where λ_P is the observation or outcome rate factor with $\lambda_P > 0$, $n \geq 0$ and is the number of observations or events, and t is the observation variable which in the case of semiconductor reliability would typically be time but could, in principle

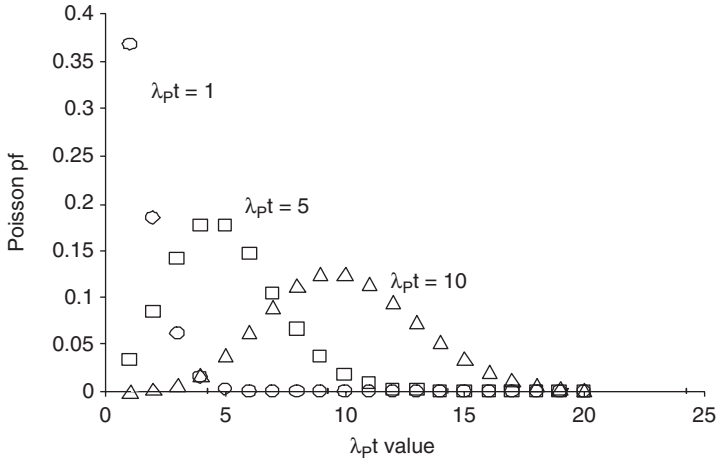

Figure 1.32. Poisson pf for λt values = 1, 5, and 10.

be any variable of interest. The pf and CDF are plotted for three values of λ_P in Figures 1.32 and 1.33, respectively.

$$\text{pf: } f(nt) = ((\lambda_P)^n/n!)\exp[-\lambda_P] \tag{1.36}$$

$$\text{CDF: } F(n) = P(N \le n) = \sum_{i=0}^{n} \left((\lambda_P)^i/i!\right)\exp[-\lambda_P] \tag{1.37}$$

The mean and variance for the Poisson distribution are equal to each other, and each is equal to λ_P. Again the reader is referred to Nelson [10], and Meeker [9], for a much more complete discussion on the Poisson distribution as well as Poisson analysis (Figure 1.33).

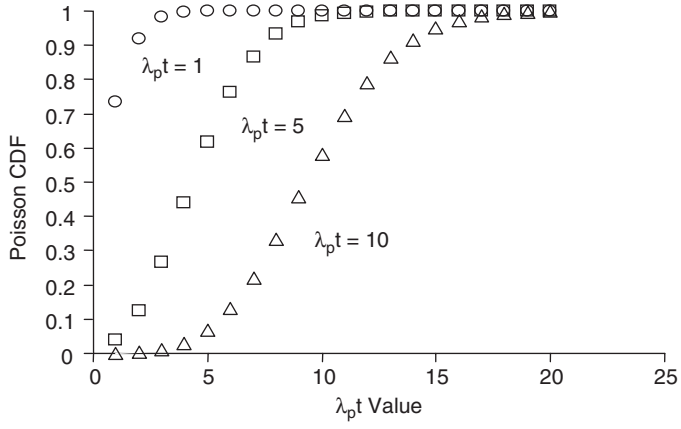

Figure 1.33. Poisson CDF for λt values = 1, 5, and 10.

1.5 CHI-SQUARE AND STUDENT *t* DISTRIBUTIONS

This section introduces those distributions that will be used not to characterize the physics of a mechanism, but rather to provide the statistical framework to test the applicability of the results.

1.5.1 Gamma and Chi-Square Distributions

The chi-square test is one of the primary metrics used to determine if the distribution chosen to characterize the mechanism is consistent with the sample data. Hence, the chi-square test is used to determine the goodness of fit. Because it is a special case of the gamma distribution, the gamma distribution is shown first. The PDF and CDF for the gamma distribution are given in Equations 1.38–1.40.

1.5.1.1 Gamma Distribution

$$\text{PDF: } f(t;\alpha_G,\beta_G) = \left(t^{\alpha_G-1}/\beta_G^{\alpha_G}\Gamma(\alpha_G)\right)\exp[-t/\beta_G] \quad \text{for } t>0. \tag{1.38}$$

$$\text{CDF: } F(t;\alpha_G,\beta_G) = (1/\Gamma(\alpha_G))\int_0^t \left(x^{\alpha_G-1}/\beta_G^{\alpha_G}\right)\exp[-x/\beta_G]dx. \tag{1.39}$$

$\Gamma(\alpha_G)$ is the gamma function given by

$$\Gamma(\alpha_G) = \int_0^\infty x^{\alpha_G-1}e^{-x}dx. \tag{1.40}$$

If $\beta_G = 2$ and $\alpha_G = \nu/2$ where ν is an integer representing the number of the degrees of freedom of the distribution, then the gamma distribution reduces to the chi-square distribution. The chi-square distribution PDF and CDF are given in Equations 1.41 and 1.42. The PDF and CDF for chi-square distributions of one and five degrees of freedom are shown in Figure 1.34.

1.5.1.2 Chi-Square Distribution

$$\text{PDF: } f(t) = \left(t^{(\nu/2)-1}/2^{(\nu/2)}\Gamma(\nu/2)\right)\exp[-t/2] \quad \text{for } t>0 \tag{1.41}$$

$$\text{CDF: } F(t) = (1/\Gamma(\nu/2))\int_0^t \left(x^{(\nu/2)-1}/2^{(\nu/2)}\right)\exp[-x/2]dx \tag{1.42}$$

It should also be noted that for the special case of two degrees of freedom ($\nu = 2$), the chi-square distribution itself reduces to an exponential distribution with a mean equal to 2. This can be seen from the PDF of Equation 1.41. $\Gamma(\nu)\} = \{1/\Gamma(\nu/2)\} \quad \{x^{(\nu/2-1)}/2^{\nu/2}\}\exp(-x/2) = \Gamma(1)\}\{1/2\}\exp(-x/2) = \{1/2\}\exp(-x/2)$. The

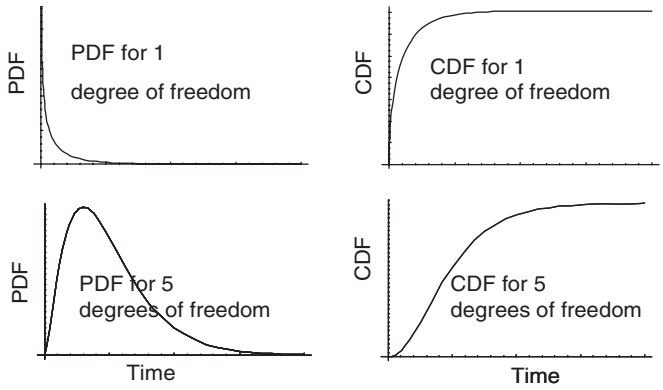

Figure 1.34. PDF and CDF for chi square distribution for 1 and 5 degrees of freedom.

gamma function could also be expressed in terms of a simple factorial expression since the gamma function argument υ has an integer value, e.g., $\Gamma(v) = (v-1)! = 1$ in the case of $v = 2$. Further discussion of the chi-square distribution is given in Green and Bourne [12].

1.5.2 Student *t* Distribution

The Student *t* test can be used to draw inferences about the sample, the population, and the statistical significance of the results. It has many uses within the field of semiconductors but its development is beyond our scope here [28]. The PDF and CDF for the Student *t* distribution is given in Equations 1.43 and 1.44, respectively.

1.5.2.1 Student t Distribution

$$\text{PDF: } f(x) = \frac{(\Gamma[(v+1)/2]/\Gamma[v/2])}{\left(\{1+(x^2/v)\}^{(v+1)/2}\sqrt{\pi v}\right)} \quad \text{where } -\infty < x < \infty \tag{1.43}$$

$$\text{CDF: } F(x) = \frac{(\Gamma[(v+1)/2]/\Gamma[v/2])}{(\sqrt{\pi v})} \int_{-\infty}^{x} \left(\{1+(y^2/v\}\right)^{-(v+1)/2} dy \tag{1.44}$$

where Γ is the gamma function and v is the degrees of freedom.

The Student *t* distribution is symmetrical about the mean and it converges to a normal distribution as v approaches infinity. Student *t* distributions are shown in Figure 1.35 for 2 and 20 degrees of freedom.

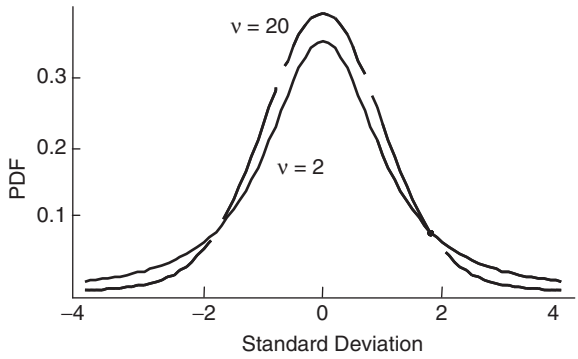

Figure 1.35. Student *t* PDF for 2 and 20 degrees of freedom.

1.6 APPLICATION

1.6.1 Readouts Versus "Exact" Time-To-Fail

Data censoring is often necessary even given the experimenter's best effort to avoid it. Various types of data censoring are discussed below. Our focus, as throughout this book, will be on the practical application to semiconductor technology reliability. Historically, one had to use separate test and stress equipment sometimes called *ex situ* stress equipment. That is still necessary sometimes today, but more often one piece of equipment can be used for both the stressing and testing of the DUTs. This is likewise sometimes referred to as *in situ* stressing. This equipment will give the "exact" time-to-fail of each DUT. *In situ* test equipment typically will be able to determine the time-to-fail within milliseconds and although this is not mathematically exact, it can be considered exact for any realistic reliability stress that is at least a few seconds or greater in duration. All of the semiconductor technology reliability stressing that is practiced today including very fast wafer-level reliability stressing is at least a few seconds in stress duration.

When *ex situ* stressing is necessary, the sample is first tested and then put into the stress chamber. Readouts should usually be done on a log-of-time basis. As an example, consider a stress sample that is to be 500 hr in duration. Assume that each test of the total sample takes one hour on the tester with a five-hour overhead of removing the sample from stress, transporting it to and from the tester, and returning the sample on stress. Further assume that either no relaxation occurs or that the samples are held at a voltage while awaiting their turn on the test equipment. The next issue to be addressed is the tradeoff between data collection, the time at which most of the failures or shifts are expected, and the total time of the stress activity. If the prime concern is the percent of the sample that survives until 500 hr, very few, intermediate readouts may be chosen especially if time is of great concern, which is often the case. As the reliability engineer, one wants to be

very cautious in this case. If the hardware all passes at the end of the 500 hr, there is no issue. However, if there are significant fails, the customer will then more than likely desire additional information. For the example requiring 6 hours off stress, the readout schedule might be 10, 30, 100, and 500 hours. Obviously if only two or three readouts were made, this becomes problematic. The best experimental design can only be achieved when the behavior is reasonably well known beforehand, so that the readouts can be taken in approximately equal steps of transformed time. Note that the first plotted fail on any CDF plot will not occur at time zero, but at the first nonzero readout. Depending on how quickly fails or shifts are expected, the reliability engineer might even choose to do a two-hour readout in addition to the above readouts or even both one- and two-hour readouts.

For the case of *in situ* stressing, a data-retention strategy is typically chosen by the equipment manufacturer. Here the issue becomes too much data when only a little information is desired. If the first significant shift or fail occurs after 18.554 hr for one of the samples on stress and the equipment has been making a readout once every 0.001 hr, someplace in the equipment storage 18,553 entries exist that present no information. Typically equipment today will give the engineer the ability to save measurements that have a shift greater than a per cent change from the last measurement or beyond some minimum threshold change. This is obviously both desirable and necessary.

The actions of the reliability engineer are the same in both cases after the data is taken. The CDF will be plotted on the appropriate axes. The difference is that for the *ex situ* case, the only points that will be plotted will be at the precise readout times or a plotting position appropriate to those times. However, what is really known about each fail in this distribution is only its time-to-fail within a certain time range. For example, for a fail that occurred at the 50-hour readout, it is only known that that fail occurred sometime between the previous readout, e.g., 20 and 50 hr. Obviously, all of the fails that occurred between 20 and 50 hr are grouped together and plotted at a single time. For this reason this type of data is called grouped data or readout data. Because these fails are only known within a range of time values, this is also called interval censoring. For the in situ case, values will be plotted at the "exact" times-to-fail, again considering the best plotting position strategy.

1.6.2 Additional Types of Censoring

Two of the most common reasons for censoring are lack of time or lack of stress equipment or both. If a stress is terminated before all of the parts have failed, it is called time censoring or Type I censoring. The times-to-fail are not known for the samples failing after the stress termination. Time censoring is very common and often necessary. The challenge for the reliability engineer is to then design the experiment in such a fashion that most of the population fails in the allotted time. Type II censoring occurs when the sample is removed from stress after a given percent of the samples have failed. This may also be called failure censoring. Failure censoring has the advantage of guaranteeing that a given, cumulative

fallout can be plotted on the CDF plot. Often the tail of the distribution may be significantly more robust than the main population so it may survive another decade or more in time. But in terms of the modeling and projection, this tail would be appropriately discounted, so time can be saved without any significant information lost in this special case of failure censoring. It is assumed that for either time censoring or failure censoring, the stress plan is decided beforehand and that the censoring is hence, unbiased. If time or fail censoring decisions are made during the stress that are based on the stress results, a bias, either favorable or unfavorable, could be introduced into the results.

Other reasons for censoring might include equipment failure or voltage surges due to power outage and return. In these cases the compromised samples would need to be removed from stress. If one were stressing on two separate stressors and only one suffered from the problem, those samples would be removed at that time, and as the stress continued on only the uncompromised equipment, the percent fails from that time on would be calculated based on the new, smaller sample.

Another reason for censoring would be the need to obtain failure analysis on an early fail or shift to determine if the signature was the expected fail signature. For example, if for an electromigration stress 20% resistance shift was considered a failure, a DUT might be pulled from the stress and sent to failure analysis if it shifted 10% much more quickly than expected. Here the data point would be sacrificed for an immediate determination as to whether a new mechanism was at work. In principle the same could hold true for a stress that uncovered several mechanisms, but for which only one was of concern. Although this would not be a common case in semiconductor reliability, it can happen.

1.6.3 Least-Squares Fit and Application

It is expected that the engineer will be using numerical techniques for plotting and analyzing the data. Hence, we do not give computer programs, nomographs, tables of calculated values for special functions, or charts showing continuous values of special functions. Once the data has been plotted, preferably by numerical means, a fitting routine can be run to give a best fit line through those values.

One of the simpler choices would be a least-squares fit to the data. In this case, the square of y-axis distance from each point, to the fitted line is summed and minimized to achieve the least-squares fit. This is shown graphically in Figure 1.36.

As discussed in Section 1.4, we choose axes which cause the function that is to be plotted to have a linear form, thus the least-squares-fit function is also a linear function.

$$y_i(a, b) = a + bt_i \tag{1.45}$$

$$Q^2(a, b) = \sum_{i=1}^{n} \left[(y_i - (a + bt_i))^2 \right] \tag{1.46}$$

Equation 1.46 is minimized when the derivative is set equal to zero. Since Equation 1.46 is a function of both 'a' and 'b,' each partial derivative must be

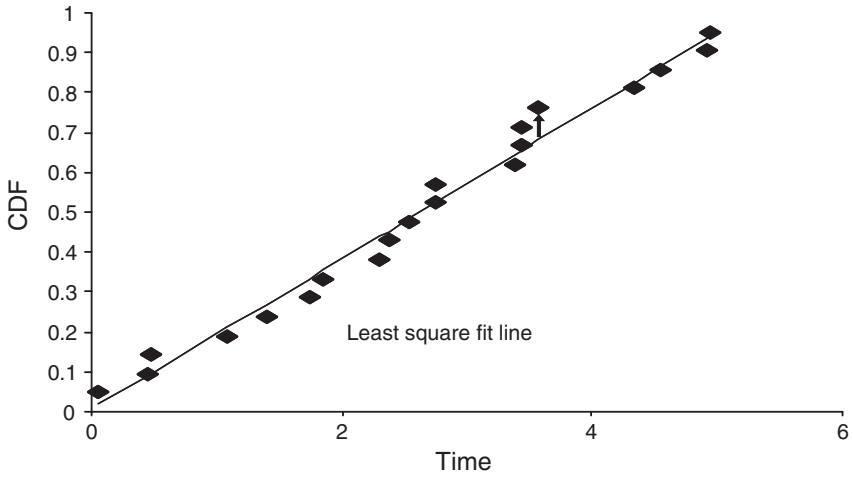

Figure 1.36. Least-squares-fit line to 20 point data set plotted with $i/(n+1)$ plotting position.

individually set to zero. Note that Q could be a function of many parameters for the nonlinear case, and in that more general case, the equivalent of Equation 1.46 would be set to zero for all partial derivatives. The derivation of the equations below is beyond the scope of this book and the reader is referred to any theoretical statistics book for a complete derivation and more complete description of these equations. The reader is directed to the end of this discussion for an example that walks through each of the following steps. Under almost every circumstance in today's world, any of several excellent software packages would be used to do these calculations; however, it is important for the reliability engineers to understand the basis of those calculations even if they have not derived each equation themselves. We start with the least-squares fit.

$$\partial Q^2(a,b)/\partial a = -2\sum_{i=1}^{n}[(y_i - (a + bt_i))] = 0. \tag{1.47}$$

$$\partial Q^2(a,b)/\partial b = -2\sum_{i=1}^{n}[(y_i - (a + bt_i))t_i] = 0. \tag{1.48}$$

The estimates of the regression coefficients, 'a' and 'b,' are given by the following equations. The estimate of the 'b' coefficient, $\hat{b}$, is the estimate of the slope of the least-squares-fit line and is given by

$$\hat{b} = \frac{\sum_{i=1}^{n} y_i t_i - \left[\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} t_i\right)/n\right]}{\left[\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2/n\right]} \tag{1.49}$$

The estimate of the '*a*' coefficient, or estimate of the intercept of the least-squares-fit line, is given by:

$$\hat{a} = \left(\sum_{i=1}^{n} y_i\right)\Big/ n - \hat{b}\left(\sum_{i=1}^{n} t_i\right)\Big/ n \tag{1.50}$$

Two figures of merit for the least-squares fit include the correlation coefficient, ρ_C shown in Equation 1.51, and a data-based estimator for the variance of the error s^2, shown in Equation 1.52. This estimator s^2 will be used throughout this chapter since it is usually true in semiconductor reliability that the variance of the population is not known. The correlation coefficient is a measure of the strength of the linear relationship between the two variables under study. ρ_C is dimensionless and ranges from -1 to 0 to $+1$, meaning an excellent negative correlation, no correlation, or a positive correlation (Figure 1.37). This is sometimes called the Pearson product moment of correlation. Although this would not normally be an issue in semiconductor reliability, note that excellent correlation does not imply a causal relationship, but only that a linear relationship exists between the variables in question. The estimator of the variance, s^2, gives a measure of the dispersion of the actual y values about the y values as fitted to the least-squares line. If the estimator of the variance was very large, it would be possible that little, if any, correlation existed even if the correlation coefficient was close to a positive or negative one. The correlation coefficient ρ_C and the variance estimator s^2 are given by:

$$\rho_C = \frac{\left[\sum_{i=1}^{n} y_i t_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} t_i\right)/n\right]}{\sqrt{\left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\Big/ n\right]}\sqrt{\left[\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2\Big/ n\right]}}. \tag{1.51}$$

$$s^2 = \left\{\frac{\left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\Big/ n\right] - b\left[\sum_{i=1}^{n} y_i t_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} t_i\right)^2\Big/ n\right]}{n-2}\right\}. \tag{1.52}$$

One very significant attribute should be noted about the least-squares fit. The error between the actual point and the fitted point contained within the line is squared. The sum of these squares is then minimized. Another and possibly preferable option would be to simply minimize the sum of the errors instead of the sum of the square of the errors. However, absolute values are not analytically soluble. The result of minimizing the sum of the squares is that points further from the fitted line have a larger contribution and are thus more heavily weighted than the closer points. Usually, but not always, the end points of the distribution, the first readouts and the last readouts, will have the largest error. These early and late points will typically be the most distant from their equivalent points on the fitted line. Hence, these early and late readouts have a greater influence on the resulting fitted line than do most of the other readouts. One would typically want to minimize the influence of these points rather than maximize it. The early points

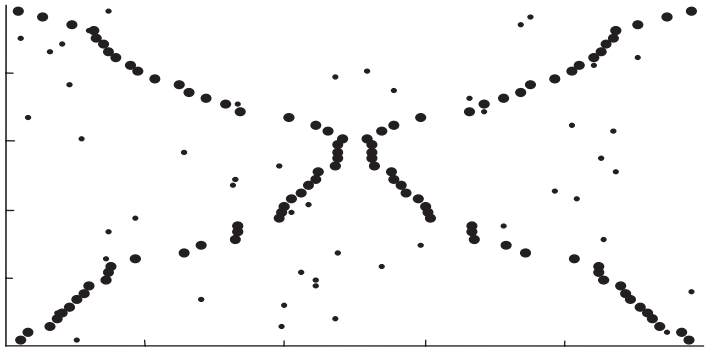

Figure 1.37. Positive correlation (increasing slope with larger dots) and negative correlation (decreasing slope with larger dots) where correlation coefficient, ρ_C, would be relatively close to + or −1, respectively, and no correlation (smaller dots) where the correlation coefficient, $\rho_C \sim 0$.

may be only slightly longer than the minimum response time of the stress/test system. The late points may not be representative of the main population due to any number of stress/test issues, or they may truly be representative of a population that has significantly longer life.

The confidence interval, at the $100(1-\alpha)$ percentile for the slope $\hat{b}$ of our straight line fit is given by Equation 1.53. Note that this will yield a range of slopes that are acceptable within our specified confidence interval. As alpha gets smaller, the range of acceptable slopes will increase. Hence, there is a tradeoff, the tighter the confidence interval desired, the larger will be the spread of values for $\hat{b}$ that must be accepted to meet that increased confidence level. This is quantified in the example at the end of this chapter. The confidence interval for the slope, C_b, is given by

$$C_b = \hat{b} \pm t_{\alpha/2} s \left\{ \sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i \right)^2 \Big/ n \right\}, \tag{1.53}$$

where $\hat{b}$ is the estimator for the slope and is given in Equation 1.49, $t_{\alpha/2}$ is taken from Student distribution tables and is also a function of the degrees of freedom, s is the square root of the variance estimator given in Equation 1.52, and n is the sample size.

The next figure of merit is for the 100 $(1-\alpha)$ percent confidence interval at a given value of time, t, for the mean value of y which in this case is the CDF_i. The reader is referred to any of the referenced statistics books for more detail. The confidence interval, C_I, is then given by

$$C_I = \hat{y} \pm t_{\alpha/2} s \sqrt{ n^{-1} + (t - \bar{t})^2 \Big/ \left(\left(\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i \right)^2 \right) \Big/ n \right) }, \tag{1.54}$$

where $\hat{y}$ corresponds to time t and is given by $\hat{y}(i) = a + \hat{b} t_i$.

Cautionary note: The confidence intervals, *t*-tests, and hypothesis tests shown above are only exact for the case where the E_i are also normal random variables. The transformations discussed in Section 1.4, which are required to cast the various failure distributions in a linear form, unfortunately cause the above regression model to be not strictly valid after the transformations. In particular, the noise terms E_i, are not independently distributed, having a tendency to be larger towards the edges of the data. Hence the maximum likelihood estimators are much preferred for the evaluation of confidence intervals, *t* tests and hypothesis testing and although they are beyond the scope of this book the reader is directed to any of the referenced statistics texts for thorough discussions of these topics. Understanding the weakness of this analyzed regression model we will continue using it for simple graphical illustrative purposes. Also note that its validity could be tested by using simulations if it were to be used instead of the maximum likelihood estimator. As already mentioned, the expectation is that one of the many computer programs for confidence interval and hypothesis testing would be used to generate the results, and that in fact a maximum likelihood estimator program would be used.

Example 1: A simple linear example of the application of Equations 1.45 through 1.54 will be given next. The simplest case to illustrate the applications of the equations and avoid getting lost in the transformation equations is used. Note that in Section 1.4 the transformations that were shown allow the normal, exponential, Weibull, and lognormal distributions to be plotted linearly after the axis transformations. The $i/(n+1)$ plotting position is chosen in this example for simplicity and there are 20 data points. The actual least-squares fit is shown in Figure 1.36. It was generated with software, but we now go back to reconstruct the method of doing it manually. Table 1.1 gives the time and CDF values as well as the additional terms required for the calculations.

First the slope and intercept estimates are calculated for the least-squares fit line drawn in Figure 1.36.

From Equation 1.45: $\hat{b} = \dfrac{34.01 - [(10)(52.34)/20]}{[178.5 - (52.34)^2/20]} = 0.189$

From Equation 1.46: $\hat{a} = 10/20 - (0.189)(52.34)/20 = 0.00571$

Hence, the formula for the line drawn in Figure 1.36 is $\hat{y} = 0.00571 + 0.189\ t$.

The correlation coefficient ρ_C is calculated from Equation 1.51 and yields $\rho_C = 0.991$. The value is very close to $+1$, indicating a very strong positive correlation. This indicates that there is a very strong linear relationship between these two variables.

$$\rho_C = \frac{[34.026 - (52.367)(10)/20]}{\left[\sqrt{6.508 - (10)(10)/20}\right]\left[\sqrt{(178.64) - (52.367)(52.367)/20}\right]} = 0.991$$

TABLE 1.1. Table Showing Time, CDF, Squares, Cross Products, and Sums for Least-Squares-Fit Procedure

	Time	Time^2	CDF	CDF^2	Time × CDF
	0.064	0.004	0.048	0.002	0.003
	0.446	0.199	0.095	0.009	0.042
	0.467	0.218	0.143	0.020	0.067
	1.096	1.200	0.190	0.036	0.209
	1.393	1.941	0.238	0.057	0.332
	1.734	3.005	0.286	0.082	0.495
	1.858	3.451	0.333	0.111	0.619
	2.306	5.317	0.381	0.145	0.878
	2.366	5.599	0.429	0.184	1.014
	2.537	6.572	0.476	0.227	1.221
	2.743	7.523	0.524	0.274	1.437
	2.753	7.582	0.571	0.327	1.573
	3.386	11.467	0.619	0.383	2.096
	3.426	11.735	0.667	0.444	2.284
	3.431	11.774	0.714	0.510	2.451
	3.572	12.759	0.762	0.581	2.722
	4.323	18.684	0.810	0.655	3.499
	4.556	20.760	0.857	0.735	3.905
	4.929	24.297	0.905	0.819	4.460
	4.955	24.553	0.095	0.907	4.719
SUM	**52.340**	**178.640**	**9.143**	**6.508**	**34.027**

The variance estimator then comes from Equation 1.52 and the square root of the variance estimator for this example is, $s = 0.0384$.

$$s = \left\{ \sqrt{\frac{[6.508 - (10)(10)/20] - [0.189(34.026 - (10)(52.367)/20]}{18}} \right\} = 0.0384$$

The final two metrics investigated in this chapter will be the confidence intervals on the value of the slope and on the mean value of $\hat{y}$ or the CDF estimator at a fixed time. The 95% confidence interval for the slope, C_b, is given by Equation 1.53 and the 95% confidence interval for the mean value of $\hat{y}$ at a fixed time is given by Equation 1.54. In terms of the reliability projection, the value of the slope is by far more critical than the mean value of $\hat{y}$ at a fixed time. That is because any slope error is magnified by the reliability projection across at least several orders of magnitude of time in the course of data being taken at highly accelerated conditions while the projection must typically reach 10 years as demonstrated in our first figure, Figure 1.1. Any error in mean value of $\hat{y}$ at a fixed time remains constant on a percentage basis as the projection

extends across the decades. These figures of merit are shown on Figure 1.38 and Figure 1.39 respectively.

$$C_b = 0.189 \pm (2.101)(.0384)/((178.5 - (52.34)^2/20))^{0.5} = 0.189 \pm .037$$

$$C_I = \hat{y} \pm (2.101)(0.0384)\sqrt{\frac{1}{20} + \frac{(t - 2.617)^2}{178.5 - (52.34)^2/20}}$$

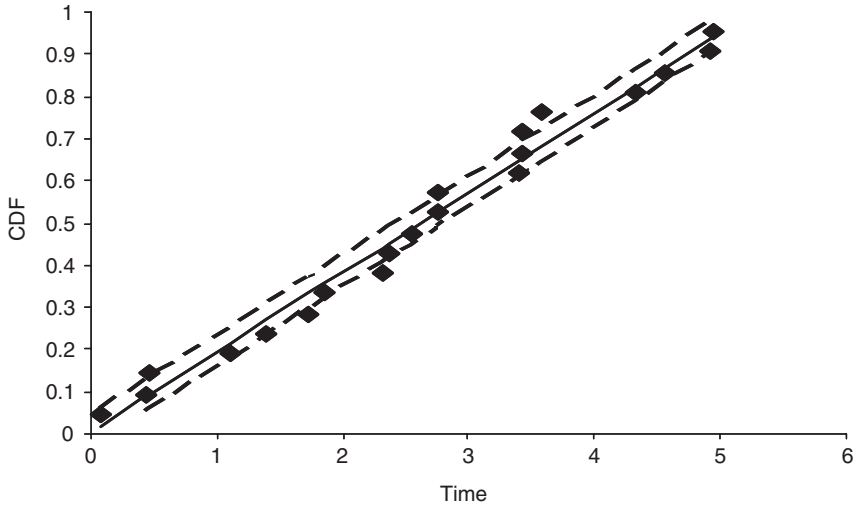

Figure 1.38. Least-squares-fit line to a 20 point data set, plotted showing point-wise confidence intervals computed for each value of *t*, using Equation 1.54.

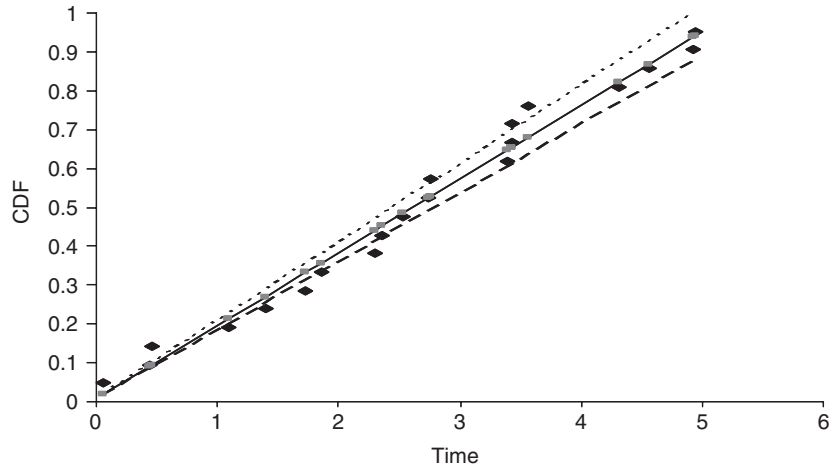

Figure 1.39. Least-squares-fit line with 20 data points plotted with confidence limits calculated based on Equation 1.53.

For the case of $t = 0$, from Equation 1.44: $\hat{y}(\hat{a}, \hat{b}) = 0.00571 + 0.189t = 0.00571$ and above

$$C_{\mathrm{I}} = (0.00571 + 0.189\,t) \pm (2.101)(0.0384)\sqrt{\frac{1}{20} + \frac{(t - 2.617)^2}{178.5 - (52.34)^2/20}}$$

$$= 0.00571 \pm 0.03736$$

Hence 95% confidence interval for the mean value of $\hat{y}$ at a fixed time of zero lies between the values of -0.0316 and $+0.043$. Obviously for CDF, the value must always be positive. The 95% confidence interval for the slope lies between 0.176 and 0.201.

A casual observation of Figures 1.38 and 1.39 may not be convincing that the conservative value for the slope has a greater impact on the final reliability projection than the conservative value of the least-squares fit due to the mean value of $\hat{y}$ at a fixed time for the 95% confidence intervals. Note that in these figures no projection has been made. Again we must emphasize that all of the results for Example 1 are only approximate because the noise terms are not independently distributed. Hence the results on the figures are also only approximate. The intent here is to give the reader a simple approximation. The more accurate computer generated results for a maximum likelihood estimator are much preferred; however, it is also good to have a simple approximate method as a sanity check.

Figure 1.40 is a qualitative depiction of three errors terms that conspire against the reliability engineer. Two terms were discussed above with respect to the linear regression model even though severe weaknesses were highlighted for its use in determining confidence bounds. Those two error terms are shown qualitatively as the slope error and the mean error in Figure 1.40. The third error is the error in the acceleration factor itself. Limited sample sizes, random statistical variation, and random process variation all limit the accuracy of the acceleration calculation. Often the models used in semiconductor reliability have an exponent that controls the acceleration so that a very small error in that exponent is magnified many times. Here we have assumed that there are no experimental or equipment errors. The slope and mean error indicators are only shown on the worst case side of the two acceleration error indicators. The data points are practically lost on a chart such as this and one can see that the total error bar is approximately an order of magnitude as it is depicted in this figure. Although this is nothing more than a sketch, it is very realistic to expect these three errors to result in an uncertainty of that magnitude. If the time is to be a given, then the uncertainty is in the value of the failure rate and again that can be on the order of a factor of ten.

Finally compare this figure to our starting point of Figure 1.1. Hopefully at this point the reader understands not only the procedure for generating plots like Figure 1.1 or 1.40, but also understands the precautions that must be considered as one goes about designing an experiment, taking the data, determining the

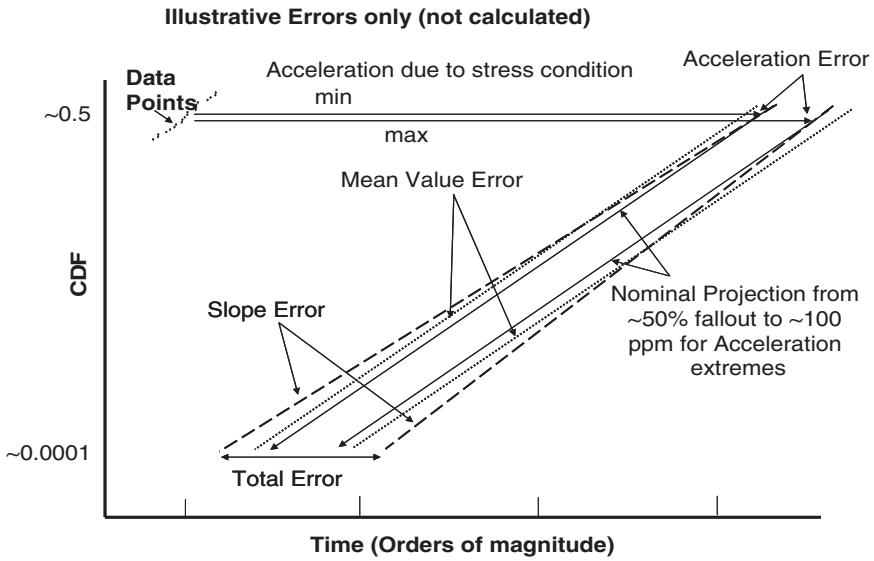

Figure 1.40. Sketch of 20 data points showing errors due to acceleration uncertainties, slope uncertainties, and mean value uncertainties, projecting from the 50% fallout to ~100 ppm.

appropriate distribution function, calculating the acceleration factors, and finally making the plots either physically or with a computer.

1.6.4 Chi-Square Goodness of Fit Application

The chi-square goodness of fit test may be used to test the hypothesis that data, sampled from a population, fit an assumed distribution. The concept behind this test is to first divide the data into intervals or cells, and to then compare the expected value in each interval based on the assumed distribution, to the observed value in that interval. The test is sensitive to both the size of the intervals and the data frequency within each interval. Note that if the expected data frequency is less than five, some intervals may need to be combined.

$$\chi^2 = \sum_{i=1}^{c} \{(Observed_i - Expected_i)^2 / Expected_i\} \tag{1.55}$$

In Equation 1.55, c is the number of the intervals or cells or groups of data and for Equation 1.55 to approach a true chi-square distribution, the sample size, n, must approach ∞. However, as long as n is large enough so that each grouping of data has a minimum of five observations, Equation 1.55 can be used for a chi-square goodness of fit test. Note that the chi-square statistic has $c-1$ degrees of freedom if no parameters need to be estimated.

Typically the parameter values will be estimated from the data in the case of semiconductor reliability. For that case, Equation 1.55 may be rewritten explicitly in terms of the sample size and the probability of the samples falling within a given grouping. The student is referred to Mann [11] for a detailed discussion of this subject.

$$\chi^2 = \sum_{i=1}^{c} \{(Observed_i - nP_i)^2 / nP_i\} \tag{1.56}$$

where c is the number of the intervals or cells or groups of data and again the sample size, n, is sufficiently large for each cell to have at least five observations.

The first step in applying the chi square goodness of fit test then, is to divide the data into groups so that a histogram of the actual data may be plotted. Typically the null hypothesis to be tested in this case will be that the data can be described by the assumed distribution and the alternative hypothesis is that the data cannot be described by the assumed distribution. The chi square figure of merit is defined by Equation 1.56. This test assumes the chi square distribution has $c-p$ degrees of freedom where c is the number of nonempty cells, and p is the number of unspecified parameters of the distribution plus one. For the two-parameter Weibull distribution in Equation 1.25 or the two parameter lognormal distribution in Equation 1.33, $p = 3$. The hypothesis that the data can be described by the assumed distribution is rejected if Equation 1.57 is true. The term $\chi^2_{(1-\alpha,\mathrm{m-p})}$ is calculated from the chi square CDF given in Equation 1.42.

$$\chi^2 > \chi^2_{(1-\alpha,\mathrm{m-p})} \tag{1.57}$$

where α is the level of significance and m–p are the degrees of freedom.

An application of the chi square test using these equations is given below. The fail times are given in Table 1.1 and the chi square test will now be used to determine whether or not these fail times belong to a uniform distribution with $0 \le t \le 5$. Three cells are chosen between 0 and 5 such that the cell boundaries are at $c_1 = 1.4$, $c_2 = 2.8$ and $c_3 = 5$ with the cells having 5, 7, and 8 observed fails respectively. Also note that since we have discrete fail times, the cell separations, c_i, must be chosen so as to avoid those fail times. The expected values for the three intervals or cells may be calculated using the equation: $P_i(t) = (t_i - t_{i-1})/(t_2 - t_1)$ where the distribution is defined between the values of t_1 and t_2, and which in this case are 0 and 5, respectively. Those results are: $P_1 = 0.28$, $P_2 = 0.28$, and $P_3 = 0.44$. After multiplying by the sample size, the estimated number of observations in each cell becomes 6, 6, and 9, respectively. Note that the requirement of a minimum of five expected occurrences per cell is met in this example. Substitution of these results into Equation 1.56 then yields:

$$\chi^2 = (5-6)^2/6 + (7-6)^2/6 + (8-9)^2/9 = 0.444$$

The hypothesis that the data is sampled from a uniform distribution is not rejected for $\alpha = 0.05$, since $\chi^2_{\ 0.95}\,(2) = 5.99$.

1.6.5 Maximum Likelihood Estimation (MLE)

Conceptually, the objective of MLE is to determine that set of distribution parameters that will maximize the likelihood of representing the sample data. MLE is both a powerful and versatile method for fitting statistical distribution models to sample data. MLE is very useful for semiconductor reliability both because it has the power to fit all of the distributions commonly used in reliability to sample data and because it provides a best fit for all of the distribution parameters across all of the cells of a stress. MLE can also be used for hypothesis testing. Although the computation required for MLE is typically complex, this is not an issue given today's software packages.

The likelihood function, L, is the joint probability shown in Equation 1.58.

$$L = \prod_{i=1}^{n} f_{xi}(x_i; \theta_1, \theta_2, ..., \theta_k) \tag{1.58}$$

where f_{xi} (x_i; θ) is the probability of all of the x_is occurring together given the distribution parameters θ_1, θ_2,…, θ_k, f is the probability density function, x_i is a random sample of size n, and θ is a vector consisting of the set of unknown distribution parameters.

Discussion of the MLE theory and technique is beyond the scope of this book and the reader is referred to, for example, Mann [11], for an excellent discussion of the MLE technique as applied to a three-parameter Weibull distribution. Li et al. recently published two excellent articles on the MLE theory and techniques for the lognormal distribution [29, 30]. Nelson [10] also addresses MLE in a comprehensive manner.

1.6.6 Closure

This brief introduction to statistics should be adequate for the reader to navigate the remainder of this book. It has been abbreviated because the focus of this book is on the reliability mechanisms themselves, not on reliability statistics. The development of the Weibull and lognormal distributions will be discussed further in Chapters 2 and 7, respectively. Also in Chapter 2, additional examples of confidence bounds are given. However, none of these statistics topics have been treated rigorously since this book is focused on CMOS reliability. Furthermore, some very important topics, for example MLE, have only been mentioned. The reliability engineer that clearly understands statistics is in the best position to both understand what information is in the data as well as advance the understanding of the physics of the mechanism. In no way should the brevity of this treatment be construed as diminishing the importance of reliability statistics. Both introductory reliability statistics books as well as advanced reliability statistics books have been referenced and the readers are encouraged to avail themselves of these references as well as any of the other of the many statistics books on the market.

Many other areas of statistics have not even been mentioned, including some the statistical considerations that go into the design of the experiments. Although several reasons for censoring have been highlighted, we stopped short of providing

the statistics to do so. The reader should also be aware that the field of statistics includes the appropriate formalisms to treat replacement parts and repairable parts. One can think of fitting multiple cells and the statistical treatment required for factorial experiments. All of these have more or less applicability to semiconductor reliability depending on the particular issues and experiments, and again the reader is referred to the full books on these subjects. Several good texts not already referenced include [31, 32].

We have discussed in some detail the mechanics of a projection. Once the mechanics are mastered, the real engineering begins. A Weibull distribution was shown to have theoretical justification for modeling a physical system that behaves as a 'weakest link' system, as discussed and referenced in Section 1.4.7. In this case the theoretical support for the Weibull distribution, a subset of an extreme value distribution, is very strong. One theoretical basis for lognormal distribution is the proportional growth or multiplicative model and is typically the case for material transport. Knowing the mechanism is an important part of choosing the correct modeling distribution.

One final word of caution for today's world where most, if not all, of reliability analyses are done by machine—this is truly progress; however, always do a sanity check on the results. Try to estimate the result so that you will not be blindsided by input errors that may have missed an order of magnitude or used inconsistent units. At the risk of nostalgia, the one and only advantage of the ancient instrument called a slide-rule was that it gave the result only to three significant digits and it did not show the decimal point. The three significant digits could be refined but the engineer had to independently calculate the decimal point. The engineer was forced to look at the problem and result in enough detail to notice an order of magnitude error.

REFERENCES

1. Ron R. Troutman. *Latchup In CMOS Technology: The Problem and Its Cure*. Kluwer Academic Publishers, Norwell, MA: 1986.
2. Steven Howard, Voldman. *ESD Physics and Devices*. John Wiley and Sons, Hoboken, NJ: 2004.
3. Ajith E. Amerasekera, Charvaka, Duvvury. *ESD in Silicon Integrated Circuits*. John Wiley and Sons, Chichester: 2002.
4. Special SER Edition. IBM J. of Res. and Dev., 40(1): Jan 1996, pp 1–136.
5. Tobias, Paul A., Trindade, David C. *Applied Reliability*. Van Nostrand Reinholdt, New York: 1986.
6. Lee, C. Tom, Deborah, Tibel, Sullivan, Timothy D., Forhan, Sheri. Comparison of isothermal, constant current and SWEAT wafer level EM testing methods. IRW: 2001, pp 194–199.
7. Tom C. Lee, Michael, Ruprecht, Deborah, Tibel, Timothy D. Sullivan, Wen. Electromigration study of Al and Cu metallization using WLR isothermal method. IRPS: 2002, pp 327–335.

8. A. E. Zitzelsberger et al. On the use of highly accelerated electromigration tests (SWEAT) on copper. IRPS: 2003, pp 161–165.
9. W. Q. Meeker, L. A. Escobar. *Statistical Methods for Reliability Data*. John Wiley and Sons, New York: 1998.
10. Wayne, Nelson. *Applied Life Data Analysis*. John Wiley and Sons, New York: 1982.
11. Nancy R. Mann, Ray E. Schafer, Singpurwalla, D. Nozer. *Methods for Statistical Analysis of Reliability and Life Data*. John Wiley and Sons, New York: 1974.
12. A. E. Green, A. J. Bourne. *Reliability Technology*. John Wiley and Sons, London: 1972.
13. Terry, Sincich. *Statistics by Example*. Dellen Publishing Co., San Francisco: 1982.
14. W. C. Riordan, R. Miller, E. R. St. Pierre. Reliability improvement and burn in optimization through the use of die level predictive modeling. IRPS: 2005, pp 435–445.
15. W. C. Riordan, R. Miller, Sherman, J. M. Hicks. Microprocessor reliability performance as a function of die location for a 0.25 μm five layer metal CMOS logic process. IRPS: 1999, pp 1–11.
16. A. W. Strong et al. Gate dielectric integrity and reliability in 0.5 μm CMOS technology," IRPS: 1993, pp 18–21.
17. C. H. Stapper. LSI yield modeling and process modeling. IBM J. of Res. and Dev., 20: 1976, pp 228–234.
18. C. H. Stapper, A. N. McLaren, M. Dreckmann. Yield model for productivity optimization of VLSI memory chips with redundancy and partially good product. IBM J. Res. and Dev., 24, 1980, pp 398–109.
19. G. H. Hahn, S. S. Shapiro. *Statistical Models in Engineering*. John Wiley and Sons, New York: 1967.
20. Johnson, Leonard G. The median ranks of sample values in their population with an application to certain fatigue studies. Industrial Mathematics, 2: 1951.
21. R. G. Filippi et al. Paradoxial predictions and minimum failure time in electromigration. App. Phys. Letters, 66(16): 1995, pp 1897–1899.
22. D. R. Wolters, J. F. Verwey. Breakdown and wear-out phenomena in SiO2 films. In: *Instabilities in Silicon Devices*, edited by Barbottin and Vapaille. North-Holland: 1986, pp 315–362.
23. R. P. Vollertsen, W. G. Kleppmann. Dependence of dielectric time to breakdown distributions on test structures. Proc. of IEEE 1991 Int. Conf. on Microelectronic Test Structures, 4: 1991, pp 75–80.
24. E. Y. Wu, J. H. Stathis, L.-K. Han. Ultra-thin oxide reliability for ULSI application. Semiconductor Sci. and Tech., 15: 2000, pp 425–435.
25. Yashchin, Emmanuel. Modeling and analyzing breakdown phenomena in insulators–A stochasitic approach. Research Thesis, Israel Institute of Technology, Haifa: May 1981.
26. D. S. Peck. Semiconductor reliability predictions from life distribution data. Proc. of the AGET Conf. on Reliability of Semiconductor Devices: 1961.
27. B. T. Howard, G. A. Dodson. High stress aging to failure of semiconductor devices. Proc. of the Seventh National Symposium on Reliablity and Quality Control: 1961.
28. T. T. Soong. *Fundamental of Probability and Statistics for Engineers*. Wiley and Sons, Hoboken, NJ: 2004.

29. B. Li, Yashchin, E. Christiansen C., J. Gill, R. Filippi, T. Sullivan. Application of three-parameter lognormal distribution in EM data analysis. Microelectronics Rel, 46: 2006, pp 2055–2049.
30. B. Li, C. Christiansen, J. Gill, T. Sullivan, E. Yashchin, R. Filippi Threshold electromigration failure time and its statistics for copper interconnects. J. of Appl. Phys., 100: 2006, pp 114516-1–10.
31. C. R. Hicks. *Fundamental Concepts in the Design of Experiments*. Holt, Rinehart and Winston, New York: 1973.
32. William J. Diamond. *Practical Experimental Designs for Engineers and Scientists*. Van Nostrand Reinholt, New York: 1989.