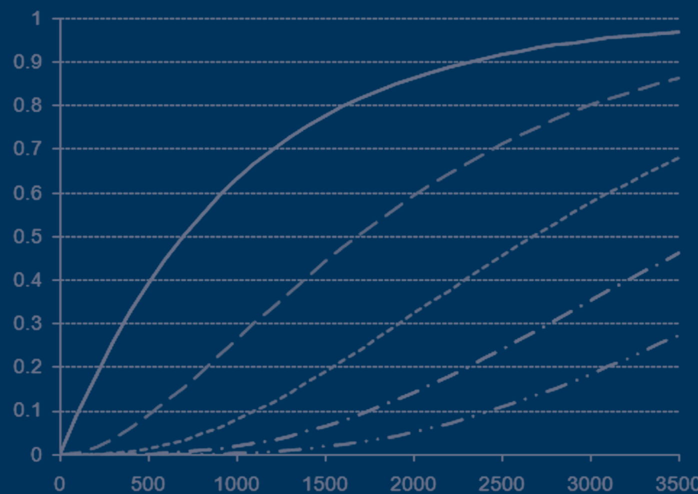
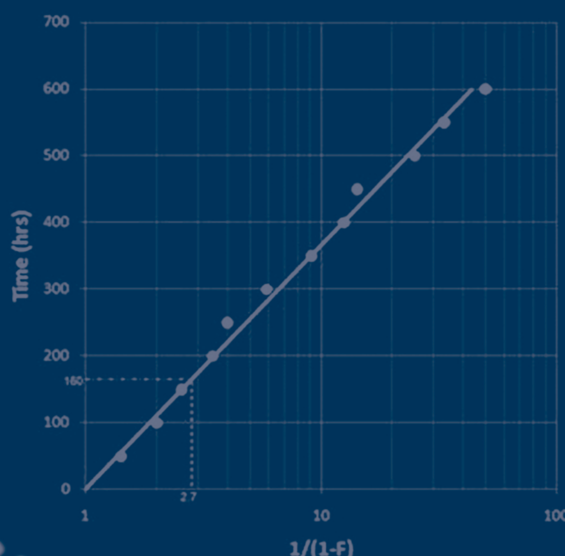
Applied Reliability

Third Edition

Paul A. Tobias
David C. Trindade





CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Applied Reliability

Third Edition

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Applied Reliability
Third Edition

Paul A. Tobias
David C. Trindade

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

International Standard Book Number: 978-1-58488-466-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

Tobias, Paul A.
Applied reliability / Paul A. Tobias and David C. Trindade. -- 3rd ed.
p. cm.
Summary: "This popular book is an easy-to-use guide that addresses basic descriptive statistics, reliability concepts, the exponential distribution, the Weibull distribution, the lognormal distribution, reliability data plotting, acceleration models, life test data analysis and systems models, and much more. The third edition includes a new chapter on Bayesian reliability analysis and expanded, updated coverage of repairable system modeling. Taking a practical and example-oriented approach to reliability analysis, it also provides detailed illustrations of software implementation throughout using several widely available software packages. Software and other files are available for download at www.crcpress.com"-- Provided by publisher.
Summary: "It's been over 15 years since the publication of the 2nd edition of Applied Reliability. We continue to receive positive feedback from old users, and each year hundreds of engineers, quality specialists, and statisticians discover the book for the first time and become new fans. So why a 3rd edition? There are always new methods and techniques that update and improve upon older methods, but that was not the primary reason we felt the need to write a new edition. In the last 15 years, the ready availability of relatively inexpensive, powerful, statistical software has changed the way statisticians and engineers look at and analyze all kinds of data. Problems in reliability that were once difficult and time consuming for even experts now can be solved with a few well-chosen clicks of a mouse. Additionally, with the quantitative solution often comes a plethora of graphics that aid in understanding and presenting the results. All this power comes with a price, however. Software documentation has had difficulty keeping up with the enhanced functionality added to new releases, especially in specialized areas such as reliability analysis. Also, in some cases different well-known software packages use different methods and output different answers. An analyst needs to know how to use these programs effectively and which methods are the most highly recommended. This information is hard to find for industrial reliability problems"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-58488-466-8 (hardback)
1. Reliability (Engineering) 2. Quality control--Statistical methods. I. Trindade, David C. II. Title.

TA169.T63 2011
620'.00452--dc22 2011011223

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

It has been more than 15 years since the publication of the second edition of *Applied Reliability*. We continue to receive positive feedback from old users, and each year, hundreds of engineers, quality specialists, and statisticians discover the book for the first time and become new fans. So, why a third edition? There are always new methods and techniques that update and improve upon older methods, but that was not the primary reason we felt the need to write a new edition. In the past 15 years, the ready availability of relatively inexpensive, powerful, statistical software has changed the way statisticians and engineers look at and analyze all kinds of data. Problems in reliability that were once difficult and time consuming for even experts can now be solved with a few well-chosen clicks of a mouse. Additionally, with the quantitative solution often comes a plethora of graphics that aid in understanding and presenting the results.

All this power comes with a price, however. Software documentation has had difficulty keeping up with the enhanced functionality added to new releases, especially in specialized areas such as reliability analysis. Also, in some cases, different well-known software packages use different methods and output different answers. An analyst needs to know how to use these programs effectively and which methods are the most highly recommended. This information is hard to find for industrial reliability problems.

The third edition of *Applied Reliability* was written to fulfill this software documentation need for reliability analysts. We chose two popular software packages that are well maintained, supported, and frequently updated: Minitab and SAS JMP. Minitab is popular in universities and JMP is widely used within leading high-technology companies. Both packages have extensive capabilities for reliability analysis and graphics that improve with every new release.

In addition, we included solutions using spreadsheet programs such as Microsoft Excel and Oracle OpenOffice Calc. With a little formula programming, spreadsheet functions can solve even very difficult reliability problems. Spreadsheet methods cannot easily produce custom, specialized reliability graphics, however, and are included primarily because they are so widely available and surprisingly powerful.

Unfortunately, producing detailed examples using software has many pitfalls. We would generate graphics of screenshots and describe how to obtain specific platforms and run analyses only to have a new release of either JMP or Minitab come out, which looked and operated somewhat differently. Even spreadsheet mechanics change with new releases. We frequently had to go back and redo problem solutions to remain current with updates.

Finally, we realized that our readers would inevitably see panels and screens coming from later releases of these software packages that might differ slightly from the screenshots shown in our text. However, it is likely that the basic methods and approaches will remain the same for a long time. Many of the suggestions we made to software developers based on methods described in the second edition are now a part of these packages or will be in future releases. Two examples are the very useful defect model (incorporated in JMP release 9) and the ability to input negative frequencies when analyzing truncated data (already in JMP 8).

We stated in the preface to the second edition: "Our goal remains that the text be application oriented, with numerous practical examples and graphical illustrations." Statements of theory and useful equations are essential building blocks, but what the industrial reliability analyst needs to know is how to apply these building blocks to numerically solve typical problems. The new edition has more than 150 worked-out examples, many done with both JMP and Minitab and even spreadsheet programs. Along with these examples, there are nearly 300 figures, and hundreds of exercises and additional problems at the end of each chapter. We also took the opportunity to add new material throughout. Sometimes, this new material increased the level of difficulty, and we chose to put this material in appendices at the end of several chapters.

Since many of the examples, exercises, and problems use lengthy spreadsheets or worksheets of failure data, we have many of these files on the publisher's website for the book. These data sets, in Excel, JMP, or Minitab format, can be accessed via the "Downloads & Updates" tab on the book's web page at http://www.crcpress.com/product/isbn/9781584884668. Data sets are organized by book chapter and given a name either mentioned in the text or based on the number of the example, exercise, or problem to which they relate. There is also a directory containing Excel templates that can be used to find maximum likelihood solutions for Weibull and lognormal multistress, life test, or field data. There are even templates incorporating the defect model or for testing equal slopes or equal parameters across several cells of data.

Another powerful software package not used in the text deserves mention: SPLUS, with the addition of Bill Meeker's SPLIDA (SPLUS Life Data Analysis) downloadable front end, which offers graphics and analysis capabilities that can also be used successfully on many of the data sets in the third edition.

Finally, we gratefully acknowledge the comments and suggestions made by our colleagues who provided feedback on the sections of the second edition and/or reviewed draft copies of many prepublication chapters of the third edition. In particular, we appreciate the comprehensive suggestions and critiques offered by Wayne Nelson, Doug Montgomery, Judy Koslov, Bill Heavlin, Ed Russell, Ken Stephens, Leon Lopez, and the many users of the text.

List of Figures

List of Tables

List of Examples

1

Basic Descriptive Statistics

One of the most useful skills that a reliability specialist can develop is the ability to convert a mass (mess?) of data into a form suitable for meaningful analysis. Raw numbers by themselves are not useful; what is needed is a distillation of the data into information.

In this chapter, we discuss several important concepts and techniques from the field of descriptive statistics. These methods are used to extract a relevant summary from collected data. The goal is to describe and understand the random variability that exists in all measurements of real world phenomena and experimental data. These concepts and techniques are basic and are applied to reliability data throughout the book.

The topics we cover include populations and samples; frequency functions, histograms, and cumulative frequency functions; the population cumulative distribution function (CDF) and probability density function (PDF); elementary probability concepts, random variables, population parameters, and sample estimates; theoretical population shape models; and data simulation.

1.1 Populations and Samples

Statistics is concerned with variability, and it is a fact of nature that variation exists. No matter how carefully a process is run, an experiment is executed, or a measurement is taken, there will be differences in repeatability due to the inability of any individual or system to completely control all possible influences. If the variability is excessive, the study or process is described as lacking control. If, on the other hand, the variability appears reasonable, we accept it and continue to operate. How do we visualize variability in order to understand if we have a controlled situation?

Consider the following example:

EXAMPLE 1.1 AUTOMOBILE FUSE DATA

A manufacturer of automobile fuses produces lots containing 100,000 fuses rated at 5A. Thus, the fuses are supposed to open in a circuit if the current through the fuse exceeds 5A. Since a fuse protects other elements from possibly damaging electrical overload, it is very important that fuses function properly. How can the manufacturer be assured that the fuses do indeed operate correctly and that there is no excessive variability?

Obviously, he cannot test all fuses to the rated limit since that act would destroy the product he wishes to sell. However, he can sample a small quantity of fuses (say, 100 or 200) and test them to destruction to measure the opening point of each fuse. From the sample data, he could then *infer* what the behavior of the entire group would be if all fuses were tested.

In statistical terms, the entire set or collection of measurements of interest (e.g., the blowing values of all fuses) define a *population*. A population is the entire set or collection of measurements of interest.

Note that a population may be finite, as in the case of a fuse lot, or it may be infinite, as occurs in a manufacturing process where the population could be all product of a specific type that has been or could ever be produced in a fabricating area.

The *sample* (e.g., the 100 or 200 fuses tested to destruction) is a subset of data taken from the population. A sample is a subset of data from the population. The objective in taking a sample is to make inferences about the population.

Note that reliability data commonly exists in one of two forms. In *variables data*, the actual measurement of interest is continuous, such as time in minutes, length in inches, or temperature in degrees Celsius. In *attributes* data, the measurements are quantified into discrete categories such as pass or fail, go or no go, in spec or out of spec, small or medium or large, and so on. Attributes data includes counts, proportions, and percentages. Although both types of data are discussed in this text, applications and analysis of attributes data is treated extensively in Chapter 11.

In the fuse data example, we record variables data, but we could also transform the same results into attributes data by stating whether a fuse opened before or after the 5A rating. Similarly, in reliability work one can measure the actual failure time of an item (variables data) or record the number of items failing before a fixed time (attributes data). Both types of data occur frequently in reliability studies. In Chapter 3, Sections 3.10–3.12, we will discuss such topics as choosing a sample size, drawing a sample randomly, and the "confidence" in the data from a sample. For now, however, let's assume that the sample has been properly drawn and consider what to do with the data in order to present an informative picture.

1.2 Histograms and Frequency Functions

In stating that a sample has been randomly drawn, we imply that each measurement or data point in the population has an equal chance or probability of being selected for the sample. If this requirement is not fulfilled, the sample may be "biased" and correct inferences about the population might not be possible.

What information does the manufacturer expect to obtain from the sample measurements of 100 fuses? First, the data should cluster about the rated value of 5A. Second, the spread in the data (variability) should not be large, because the manufacturer realizes that serious problems could result for the users if some fuses blow at too high a value. Similarly, fuses opening at too low a level could cause needless repairs or generate unnecessary concerns.

The reliability analyst randomly samples 100 fuses and records the data shown in Table 1.1. It is easy to determine the high and low values from the sample data and see that the measurements cluster roughly about the number 5. Yet, there is still difficulty in grasping the full information contained in this set of data.

Let's apply the following procedure:

1. Find the *range* of the data by subtracting the lowest from the highest value. For this set, the range is $5.46 - 4.43 = 1.03$.
2. Divide the range into ten or so equally spaced intervals such that readings can be uniquely classified into each cell. Here, the cell width is $1.03/10 \approx 0.10$, and we choose the starting point to be 4.395, a convenient value below the minimum of the data and carried out one digit more precise than the data to avoid any confusion in assigning readings to individual cells. Note that the terms "bin," "class,"

or "cell" are used interchangeably in the literature and also by statistical software programs to denote one of these equally spaced intervals.

3. Increment the starting point by multiples of the cell width until the maximum value is exceeded. Thus, since the maximum value is 5.46, we generate the numbers 4.395, 4.495, 4.595, 4.695, 4.795, 4.895, 4.995, 5.095, 5.195, 5.295, 5.395, and 5.495. These values represent the endpoints or boundaries of each cell, effectively dividing the range of the data into equally spaced class intervals covering all the data points.
4. Construct a *frequency table* as shown in Table 1.2, which gives the number of times a measurement falls inside a class interval.
5. Make a graphical representation of the data by sketching vertical bars centered at the midpoints of the class cells with bar heights proportionate to the number of values falling in that class. This graphical representation shown in Figure 1.1 is called a *histogram*.

A histogram is a graphical representation in bar chart form of a frequency table or frequency distribution. The vertical axis in a histogram may represent the actual count in a cell, or it may state the percentage of observations of the total sample that occur in a cell. Also, the range here is divided by the number 10 to generate a cell width, but any convenient number

TABLE 1.1

Sample Data on 100 Fuses

4.64	4.95	5.25	5.21	4.90	4.67	4.97	4.92	4.87	5.11
4.98	4.93	4.72	5.07	4.80	4.98	4.66	4.43	4.78	4.53
4.73	5.37	4.81	5.19	4.77	4.79	5.08	5.07	4.65	5.39
5.21	5.11	5.15	5.28	5.20	4.73	5.32	4.79	5.10	4.94
5.06	4.69	5.14	4.83	4.78	4.72	5.21	5.02	4.89	5.19
5.04	5.04	4.78	4.96	4.94	5.24	5.22	5.00	4.60	4.88
5.03	5.05	4.94	5.02	4.43	4.91	4.84	4.75	4.88	4.79
5.46	5.12	5.12	4.85	5.05	5.26	5.01	4.64	4.86	4.73
5.01	4.94	5.02	5.16	4.88	5.10	4.80	5.10	5.20	5.11
4.77	4.58	5.18	5.03	5.10	4.67	5.21	4.73	4.88	4.80

TABLE 1.2

Frequency Table of Fuse Data

Cell Boundaries	Number in Cell
4.395–4.495	2
4.495–4.595	2
4.595–4.695	8
4.695–4.795	15
4.795–4.895	14
4.895–4.995	13
4.995–5.095	16
5.095–5.195	15
5.195–5.295	11
5.295–5.395	3
5.395–5.495	1
Total count	100

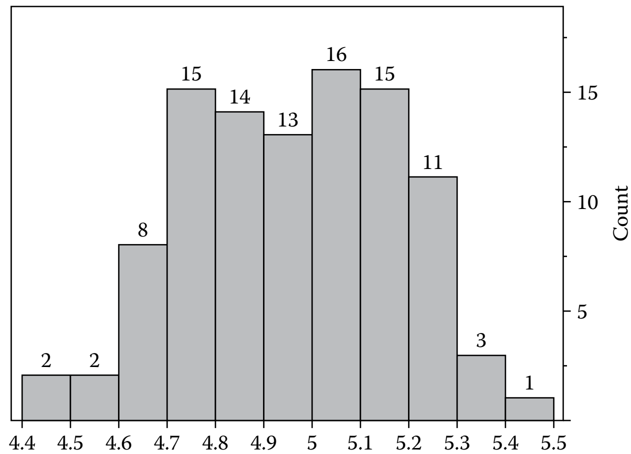


FIGURE 1.1
Histogram of fuse breakdown measurements.

(usually between 8 and 20) may be used. Too small a number may not reveal the shape of the data and too large a number can result in many empty cells and a flat-appearing distribution. Sometimes, a few tries are required to arrive at a suitable choice.

There is a useful Excel spreadsheet function called FREQUENCY that will generate a frequency table such as that shown in Table 1.2. Say the 100 fuse breakpoints are entered into column A, cells A1–A100, and the 12 cell boundaries, starting with 4.395 and ending with 5.495, are entered into column B, cells B1–B12. Next, we highlight (**click and drag**) into an empty column, say C, 13 blank rows in cells C1–C13. Then we type in the function = FREQUENCY(A1:A100,B1:B12). The expression is evaluated as a matrix operation by pressing the keys **Ctrl+Shift+Enter** together instead of pressing just **Enter** alone. This action produces the Table 1.2 frequencies in rows C2–C12 (C1 contains counts up to 4.395 and C13 contains counts after 5.495). The FREQUENCY function in OpenOffice software works the same way, except that a semicolon is used between arguments instead of a comma.

EXERCISE 1.1
Use Excel or OpenOffice to generate the frequencies given in Table 1.2, using the Table 1.1 sample data and the same interval endpoints as used in Table 1.2.

In summary, the histogram provides us with a picture of the data from which we can intuitively see the center of the distribution, the spread, and the shape. The shape is important because we usually have an underlying idea or model as to how the entire population should look. The sample shape either confirms this expectation or gives us reason to question our assumptions. In particular, a shape that is symmetric about the center, with most of the observations in the central region, might reflect data from certain symmetric distributions such as the normal or Gaussian distribution. Alternatively, a nonsymmetric appearance would imply the existence of data points spaced farther from the center in one direction than in the other, which could lead to the consideration of a distribution, such as a Weibull or lognormal.

For the data presented in Example 1.1, we note that the distribution appears reasonably symmetric. Hence, based on the histogram and the way the ends of the distribution taper off, the manufacturer believes that values much greater or much less than about 10% of the central target are not likely to occur. This variability is accepted as reasonable.

1.3 Cumulative Frequency Function

There is another way of representing the data that can be very useful. By reference to Table 1.2, let us accumulate the number of observations less than or equal to each upper cell boundary as shown in Table 1.3. This representation of the data is called a *cumulative frequency function*.

The graphical rendering of the cumulative frequency function is shown in Figure 1.2. Note that the cumulative frequency distribution is never decreasing—it starts at zero and reaches the total sample size. It is often convenient to represent the cumulative count in terms of a fraction or percentage of the total sample size used. In that case, the cumulative

TABLE 1.3

Cumulative Frequency Function for Fuse Data

Upper Cell Boundary (UCB)	Number of Observations Less than or Equal to UCB
4.495	2
4.595	4
4.695	12
4.795	27
4.895	41
4.995	54
5.095	70
5.195	85
5.295	96
5.395	99
5.495	100

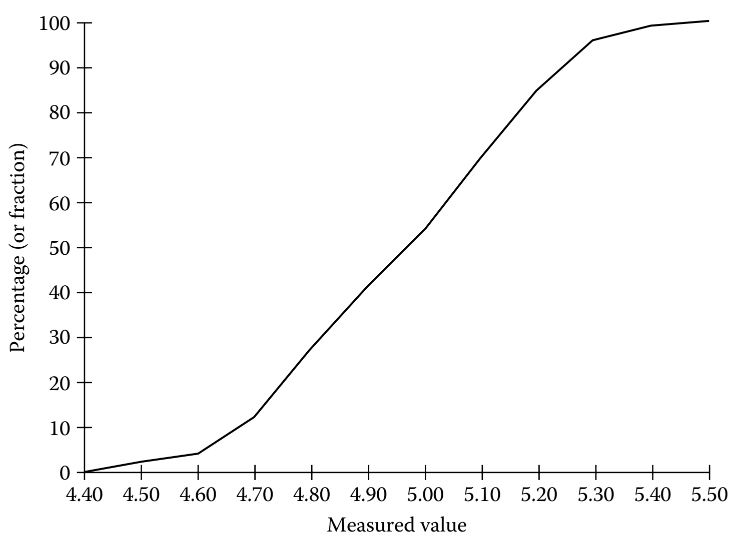


FIGURE 1.2
Plot of cumulative frequency function.

frequency function ranges from 0 to 1.00 in fractional representation or 0% to 100% in percentage notation. In this text, we often employ the percentage form.

Table 1.3 and Figure 1.2 show that the cumulative frequency curve is obtained by summing the frequency function count values. This summation process will be generalized by integration when we discuss the population concepts underlying the frequency function and the cumulative frequency function in Section 1.4.

1.4 The Cumulative Distribution Function and the Probability Density Function

The frequency distribution and the cumulative frequency distribution are calculated from sample measurements. Since the samples are drawn from a population, what can we state about this population? The typical procedure is to assume a mathematical formula that provides a theoretical model for describing the way the population values are distributed. The sample histograms and the cumulative frequency functions are the estimates of these population models.

The model corresponding to the frequency distribution is the PDF denoted by $f(x)$, where x is any value of interest. The PDF may be interpreted in the following way: $f(x)dx$ is the fraction of the population values occurring in the interval dx. In reliability work, we often have the failure time t as the variable of interest. Therefore, $f(t)dt$ is the fraction of failure times of the population occurring in the interval dt. A very simple example for $f(t)$ is the exponential distribution, given by the equation

$$f(t) = \lambda e^{-\lambda t}, \quad 0 \le t < \infty$$

where λ is a constant. The plot of $f(t)$ is shown in Figure 1.3. The exponential distribution is a widely applied model in reliability studies and forms the basis of Chapter 3.

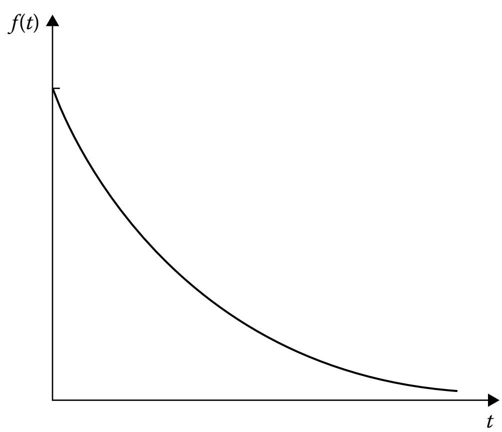


FIGURE 1.3
Plot of PDF for the exponential distribution.

The cumulative frequency distribution similarly corresponds to a population model called the CDF and is denoted by $F(x)$. The CDF is related to the PDF via the relationship

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

where y is the dummy variable of integration. $F(x)$ may be interpreted as the fraction of values less than or equal to x in the population. Alternatively, $F(x)$ gives the probability of a value less than or equal to x occurring in a single random draw from the population described by $F(x)$. Since in reliability work we usually deal with failure times, t, which are nonnegative, the CDF for population failure times is related to the PDF by

$$F(t) = \int_{0}^{t} f(y)dy$$

For the exponential distribution,

$$F(t) = \int_{0}^{t} \lambda e^{-\lambda y} dy = -e^{-\lambda y}\Big]_{0}^{t} = 1 - e^{-\lambda t}$$

The CDF for the exponential distribution is plotted in Figure 1.4.

When we calculated the cumulative frequency function in the fuse example, we worked with grouped data (i.e., data classified by cells). However, another estimate of the population CDF can be generated by ordering the individual measurements from smallest to largest, and then plotting the successive fractions

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n}{n}$$

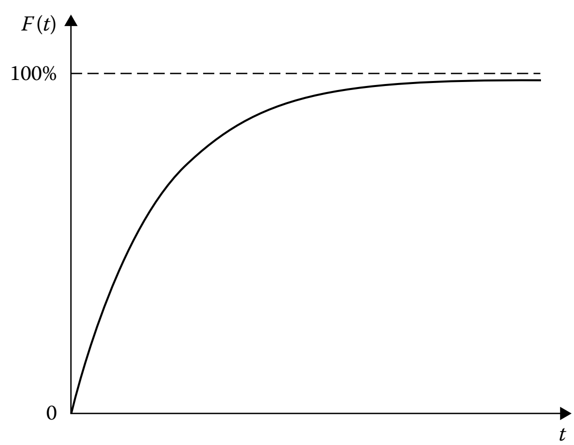


FIGURE 1.4
CDF for exponential distribution.

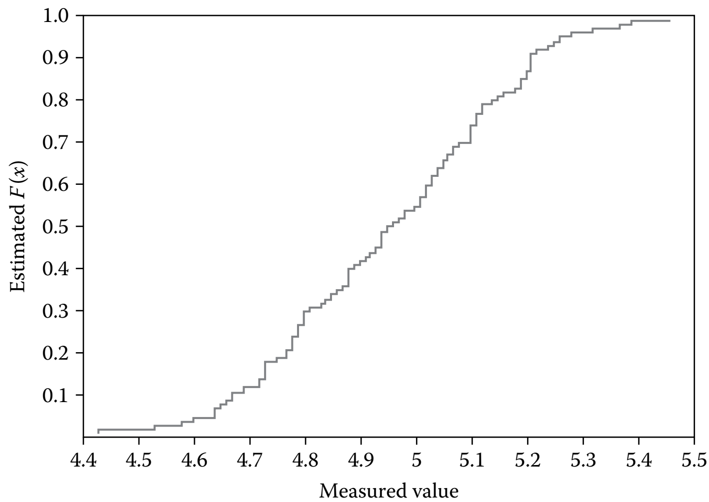


FIGURE 1.5
EDF for fuse data.

versus the ordered data points. Such a representation is called the *empirical distribution function* (EDF) and is shown in Figure 1.5 for the data from the fuse example. Note that the EDF steps up by $1/n$ at each data point and remains constant until the next point. The advantage of using the EDF instead of grouping the data is obviously that all data points are pictured; the disadvantage is that more computational effort is involved. However, spreadsheet software can easily perform the calculations and plotting. See Appendix 1A for a method to create step charts using spreadsheet plots. Since $F(x)$ is a probability, all the rules and formulas for manipulating probabilities can be used when working with CDFs. Some of these basic rules are described in Section 1.5.

EXERCISE 1.2
For the data in Table 1.1, construct a frequency table using 4.395 as the starting point and 0.2 as the interval width. Create a histogram of this frequency table. Compare it to Figure 1.1.

EXERCISE 1.3
Using the results from Exercise 1.2, construct a cumulative frequency table and create a plot of the cumulative frequency function. How does it compare to Figure 1.2?

EXERCISE 1.4
Take columns 2, 5, and 8 (left to right) from Table 1.1, for a total of 30 data points. Assume a random sample, arrange the points in order from smallest to largest, and plot the EDF. Compare to Figure 1.5.

EXERCISE 1.5
The histogram shown in Figure 1.1 was generated using JMP software. Use the JMP **Analyze, Distribution** platform and the data set shown in Table 1.1 (FuseData.jmp at the "Downloads & Updates" tab of the web page for this book at http://www.crcpress.com/product/isbn/9781584884668) to generate a histogram. The result may

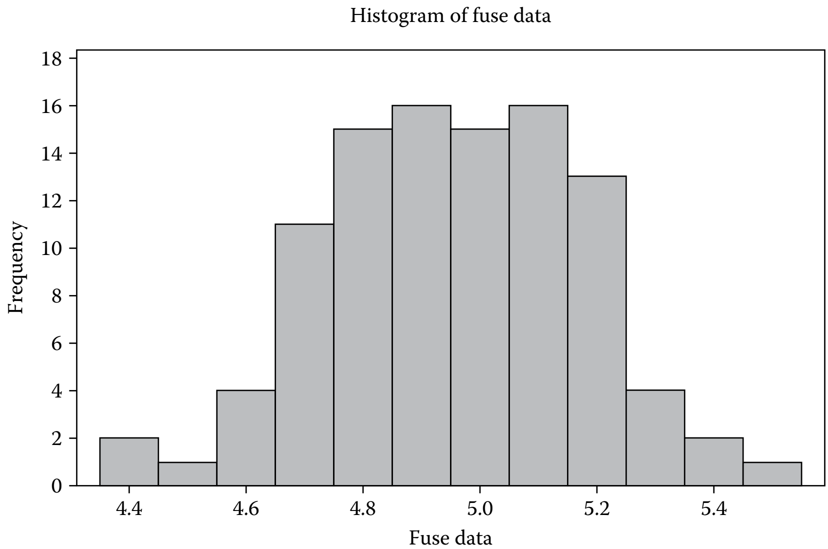


FIGURE 1.6
Minitab histogram of fuse data.

be somewhat different based on a JMP default choice for the number of bins. If so, use the "hand" cursor to change the number of bins until the same histogram as shown in Figure 1.1 is obtained. Next, use the **histogram** spreadsheet function of Excel (in **Tools, Data Analysis**) to obtain a histogram. The Table 1.1 data set for spreadsheet use is FuseData.xls. In order to get the exact same graph shown in Figure 1.1, input bin numbers starting with 4.495 and increasing in steps of 0.1 to 5.495. Finally, use Minitab to get the histogram shown in Figure 1.6 (one must change the default number of bins from 12 to 11).

NOTE: There is nothing "wrong" or "misleading" with obtaining dissimilar histograms from different software programs. It is up to the analyst to vary the bin numbers and locations to obtain one of many reasonable (although varying) views of the data. For example, the default output histogram from Minitab is shown in Figure 1.6.

1.5 Probability Concepts

In the classical sense, the term *probability* can be thought of as the expected relative frequency of occurrence of a specific event in a very large collection of possible outcomes. For example, if we toss a balanced coin many times, we expect the number of occurrences of the event "heads" to comprise approximately half of the number of outcomes. Thus, we say the probability of heads on a single toss is 0.5, 50%, or 50–50. It is typical to express probabilities either as a fraction between 0 and 1 or as a percentage between 0% and 100%.

There are two very useful relations often invoked in probability theory. These rules relate to the occurrence of two or more events. In electrical engineering terms, we are defining "and" and "or" relations. The first rule states that if $P(A)$ is the probability of event

A occurring and *P*(*B*) is the probability of event *B* occurring, then the probability of events *A* and *B* occurring simultaneously, denoted *P*(*AB*), is

$$P(AB) = P(A)P(B|A)$$

or

$$P(AB) = P(B)P(A|B)$$

where $P(A|B)$ designates the "conditional" probability of *A* given that event *B* has occurred.

Let's explain conditional probability further. We imply by the terminology that one event may be affected by the occurrence of another event. For example, suppose we ask what the probability is of getting two black cards in a row in successive draws from a well-shuffled deck of cards, without replacing the first card drawn. Obviously, the probability of the first card being a black card (call this event *A*) is

$$P(A) = \frac{\text{favorable outcomes}}{\text{total outcomes}} = \frac{26}{52} = \frac{1}{2}$$

The probability of the second card being a black card (event *B*) changes depending on whether or not the first card drawn is a black card. If yes, then the probability of the second card being a black card is

$$P\left(B|A\right) = \frac{25}{51}$$

Therefore, the probability of two successive black cards is

$$\begin{aligned} P(AB) &= P(B)P(A|B) \\ &= \frac{1}{2}\frac{25}{51} \\ &= \frac{25}{102} \end{aligned}$$

Two events, *A* and *B*, are said to be *independent* if the occurrence of one does not affect the probability of the other occurrence. The formal definition states that two events *A* and *B* are independent *if and only if*

$$P(AB) = P(A)P(B)$$

This expression is sometimes referred to as the *multiplication rule* for the probability of independent events occurring simultaneously. In general, the probability of independent events occurring is just the product of the individual probabilities of each event. For example, in the card situation, replacing the first card drawn and reshuffling the deck will make event *B* independent of event *A*. Thus, the probability of two successive black cards, with replacement and reshuffling between draws, is

$$\begin{aligned} P(AB) = P(A)P(B) &= \frac{26}{52}\frac{26}{52} \\ &= \frac{1}{4} \end{aligned}$$

Similarly, the probability of simultaneously getting a 6 on one roll of a die and an ace in one draw from a deck of cards, apparently independent events, is

$$\begin{aligned} P(AB) &= \frac{1}{6}\frac{4}{52} \\ &= \frac{1}{78} \end{aligned}$$

The extension of these conditional probability principles to three or more events is possible. For example, the rule for the joint probability of three events, A, B and C, is

$$P(ABC) = P(A)P(B|A)P(C|AB)$$

For independent events, the formula becomes

$$P(ABC) = P(A)P(B)P(C)$$

The second important probability formula relates to the situation in which either of two events, *A* or *B*, may occur. The expression for this "union" is

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

If the events are independent, then the relation becomes

$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

The last term in the above expressions corrects for double counting of the same outcomes. For example, what is the probability of getting either an ace (event *A*) or a black card (event *B*) in one draw from a deck of cards? The events are independent (see Exercise 1.6.), and therefore

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A)P(B) \\ &= \frac{4}{52} + \frac{26}{52} - \frac{4}{52}\frac{26}{52} \\ &= \frac{14}{52} = \frac{7}{13} \end{aligned}$$

Note that the term $P(A)P(B)$ subtracts out the probability for black aces. This probability has already been counted twice, once in the $P(A)$ term and once in the $P(B)$ term.

When events *A* and *B* are mutually exclusive or disjoint, that is, both events cannot occur simultaneously, then $P(AB) = 0$, and

$$P(A \cup B) = P(A) + P(B)$$

Furthermore, if both events are exhaustive in that at least one of them must occur when an experiment is run, then

$$P(A \cup B) = P(A) + P(B) = 1$$

Thus, event A is the complement of event B. Event B can be viewed as the nonoccurrence of A and designated as event $\bar{A}$. Hence, the probability of occurrence of any event is equal to one minus the probability of occurrence of its complementary event. This *complement rule* has important applications in reliability work because a component may either fail (event A) or survive (event $\bar{A}$), resulting in

$$P(\text{Failure}) = 1 - P(\text{Survival})$$

As another example, we note that the event "at least one occurrence" and the event "zero occurrences" are mutually exclusive and exhaustive events. Therefore, the probability of at least one occurrence is equal to 1 – probability of no occurrences.

An extension to three or more events is also possible. For three events A, B, and C, the formula is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(BC) - P(AC) + P(ABC)$$

For independent events, the relation becomes

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) - P(A)P(B) \\ & - P(B)P(C) - P(A)P(C) + P(A)P(B)P(C) \end{aligned}$$

For mutually exclusive, exhaustive events, we have

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) = 1$$

For four events, we begin by adding the four single-event probabilities. Then, we subtract the six possible probabilities of two events occurring simultaneously. Next, we add back in the four possible probabilities of three events occurring simultaneously. Finally, we subtract the probability of all four events occurring simultaneously. This "in and out" procedure works for any number of events, and the total number of terms in the final expression when there are n events will be $2^n - 1$.

EXAMPLE 1.2 CONDITIONAL PROBABILITIES

A tricky word problem that appears often in many forms can be stated as follows: A computer hack visits the surplus store and sees two similar hard drives displayed. The sign says, "Specially Reduced: 50–50 Chance of Working." He asks the dealer whether the hard drives operate properly. The dealer replies "at least one of them is working." What is the probability that both drives are functional? Does the probability change if the dealer says "the one on the left works"?

Solution

The first question asks for the probability that both drives work, given that at least one is working, that is, P(both work | at least one works). Let A be the event "both drives work" and let B be the event "at least one drive works." We want the $P(A|B)$. From our conditional probability formula, we can rewrite the expression as follows:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

TABLE 1.4

Possible Outcomes for Drives

No Dealer Information		At Least One Works		Left Drive Works	
Left Drive	**Right Drive**	**Left Drive**	**Right Drive**	**Left Drive**	**Right Drive**
W	W	W	W	W	W
W	N	W	N	W	N
N	W	N	W		
N	N				
Probability that both drives work					
1/4		1/3		1/2	

Now $P(AB)$ is the probability that both drives work (event A) and at least one drive works (event B). This joint event is actually the same as the probability of event A alone since event A includes event B; that is, if both drives work, then at least one works. Therefore, $P(AB) = P(A) = ([0.5][0.5]) = 0.25$, assuming the drives are independent. Since the event B "at least one drive works" and the event "both drives are not working" are mutually exclusive and exhaustive events, the denominator $P(B) = P(\text{at least one works}) = 1 - P(\text{both not working}) = 1 - (0.5)(0.5) = 0.75$. Hence, the desired probability is $P(A|B) = (0.25)/(0.75) = 1/3$.

This result surprises many individuals who incorrectly assume that the conditional probability of two working drives given at least one works should be 1/2, instead of the correct answer 1/3, since they reason that the other disk drive is equally likely to work or not work. However, the sample space of possible outcomes is listed in Table 1.4.

With no dealer information, there are four equally likely outcomes: (work, work), (work, not work), (not work, work), and (not work, not work), for the left and right drives, respectively. Thus, the probability is only 1/4 that both drives work. When we are told that at least one drive works, we eliminate the outcome (not work, not work). Therefore, we have only three equally likely outcomes remaining: (work, work), (work, not work), and (not work, work). Consequently, the probability that both drives work has increased from 1/4 to 1/3 with the added data. Alternatively, the probability that at least one of the drives does not work has decreased from 3/4 to 2/3.

On the other hand, if the dealer points out the working drive (maybe he did not have the time to test both drives), the probability that both drives work does change. Let event A be "both drives work" and C be "the left drive works." Now, $P(A|C) = (0.5)/(1) = 0.5$. In this case, there are only two possible outcomes (work, work) and (work, not work), where the first position indicates the left drive, and only one outcome of the two has both drives working.

For a set of events, $E_1, E_2, ..., E_k$, that are mutually exclusive and exhaustive, another useful relationship, sometimes called the *law of total probabilities*, applies. Any event A can be written as follows:

$$P(A) = \sum_{j=1}^{k} P(A|E_j)P(E_j)$$

In words, $P(A)$ is the weighted average of conditional probabilities, each weighted by the probability of the event on which it is conditioned. This expression is often easier to calculate than $P(A)$ directly.

EXAMPLE 1.3 TOTAL PROBABILITIES

A computer manufacturer purchases equivalent microprocessor components from three different distributors. The assembly of each computer utilizes one microprocessor, randomly chosen from an in-house inventory. Typically, the inventory consists of 30% of this component type from distributor *A*, 50% from distributor *B*, and 20% from distributor *C*. Historical records show that components from distributors *A* and *C* are twice as likely to cause a system failure as those from distributor *B*. The probability of system failure with component *B* is 0.5%. What is the probability that a computer system will experience failure?

Solution

Since there are three distributors and we randomly chose a component from one of the distributors, we have three mutually exclusive and exhaustive events. The theorem of total probability is the basis for the solution:

$$\begin{aligned} P(\text{failure}) &= P(\text{failure}|\text{distributor } A)P(\text{distributor } A) + P(\text{failure}|\text{distributor } B)\, P(\text{distributor } B) \\ &\quad + P(\text{failure}|\text{distributor } C)P(\text{distributor } C) \\ &= 2(.005)(.3) + (.005)(.5) + 2(.005)(.2) \\ &= .0075 \text{ or } .75\% \end{aligned}$$

A final key probability formula, known as Bayes' rule, allows us to "invert" conditional probabilities, that is, determine which one of the conditioning events E_j is likely to have occurred, given that event *A* has occurred. Again, for a set of mutually exclusive and exhaustive events, $E_1, E_2, \ldots, E_k$, Bayes' rule states that

$$P(E_j|A) = \frac{P(A|E_j)P(E_j)}{\sum_{j=1}^{k} P(A|E_j)P(E_j)}$$

Note that by the law of total probabilities, the denominator of this expression is just $P(A)$.

EXAMPLE 1.4 BAYES' RULE

The probability that a batch of incoming material from any supplier is rejected is 0.1. Typically, material from supplier S_1 is rejected 8% of the time, from supplier S_2, 15%, and S_3, 10%. We know that 50% of the incoming material comes from S_1, 20% from supplier S_2, and 30% from S_3. Given that the latest lot of incoming material is rejected, what is the probability the supplier is S_1?

Solution

Let *A* denote the event that the batch is rejected. Then, by Bayes' rule,

$$P(S_1|A) = \frac{P(A|S_1)P(S_1)}{P(A)} = \frac{(0.08)(0.5)}{(0.08)(0.5) + (0.15)(0.2) + (0.1)(0.3)} = 0.4$$

In this example, the starting (i.e., before we know the batch is rejected) probability of the event S_1 is 0.5. This knowledge is sometimes referred to as the "a priori" probability of S_1. After the batch rejection, Bayes' rule allows us to calculate the new (conditional) probability of S_1 as .4. The result is sometimes called the *a posteriori* probability of S_1.

EXAMPLE 1.5 BAYES' RULE APPLIED TO MISCLASSIFIED ITEMS

Assume we perform a test on a component to check for a specific defect. Historically, 1% of the components have this defect. Based on a detailed analysis of previous results, 95% of the components with the defect are detected, but 8% of the components without the defect are wrongly categorized as defective. If a component is classified as defective, what is the probability that the component actually has the defect? What is the probability that a component with a negative test has the defect?

Solution

We first solve this problem by an approach using simple average calculations. Consider a test of 2000 components. On the average, 1%, or 20, of the total components will have the defect. Of those with the defect, 95%, or 19, will be caught. However, of the 1980 without the defect, 8%, or 158, will have a false positive and be called *defective*. Therefore, 19 out of the 19 + 158 = 177 classified as defective will have the defect, and the probability of actually having the defect and a positive test result is 19/177 = 11%. This result shows that it may be a good idea to do a more extensive retest of rejected components and recover from the false positives. Also, the probability of having the defect and a negative test result is 1/(1000 − 177) = 1/1823 = .055%, which is about 1/18 of the prior probability. Next, we get the same result using the Bayes' rule formula. Let A denote a positive test and E_1 denote a defective unit. Also, let E_2 represent a unit with no defect. We want

$$P(E_1|A) = P(E_1A)/P(A) = P(A|E_1)P(E_1)/[P(A|E_1)P(E_1) + P(A|E_2)P(E_2)]$$
$$= .95(.01)/[.95(.01) + .08(.99)] = .0095/.0887 = 11\%$$

Let B = negative result. The probability of having the defect E_1 given a negative result of B is

$$P(E_1|B) = \mathrm{P}(E_1B)/P(B) = P(B|E_1)P(E_1)/[P(B|E_1)P(E_1) + P(B|E_2)P(E_2)]$$
$$= .05(.01)/[.05(.01) + .92(.99)] = .0005/.9113 = .055\%$$

EXERCISE 1.6
From a well-shuffled deck of cards, let drawing an ace event A and let drawing a black card be event B. Determine $P(AB)$, the probability of getting a black ace in a single draw, and show that events A and B are independent.

EXERCISE 1.7
Three assembly plants produce the same parts. Plant A produces 25% of the volume and has a shipment defect rate of 1%. Plant B produces 30% of the volume and ships 1.2% defectives. Plant C produces the remainder and ships 0.6% defectives. Given that a component picked at random from the warehouse stocked by these plants is defective, what are the probabilities that it was manufactured by plant A or B or C?

EXERCISE 1.8
An electronic card contains three components: A, B, and C. Component A has a probability of .02 of failing in 3 years. Component B has a probability of .01 of failing in 3 years and component C has a probability of .10 of failing in 3 years. What is the probability that the card survives 3 years without failing? What assumptions were made for this calculation?

1.6 Random Variables

In reliability studies, the outcome of an experiment may be numerical (e.g., time to failure of a component) or the result may be other than numerical (e.g., type of failure mode associated with a nonfunctional device). In either case, analysis is made possible by assigning a number to every point in the space of all possible outcomes—called the *sample space*. Examples of assigning numbers are as follows: the time to failure is assigned the elapsed hours of operation, and the failure mode may be assigned a category number 1, 2, and so on. Any rule for assigning a number creates a random variable. A random variable is a function for assigning real numbers to points in the sample space.

The practice is to denote the random variable by a capital letter (X, Y, Z, etc.) and the realization of the random variable (i.e., the real number or piece of sample data) by the lower case letter (x, y, z, etc.). Since this definition appears a bit abstract, let us consider a simple example using a single die with six faces, each face having one to six dots. The experiment consists of rolling the die and observing the upside face. The random variable is denoted X, and it assigns numbers matching the number of dots on the side facing up. Thus, $(X = x)$ is an event in the sample space, and $X = 6$ refers to the realization where the face with six dots is the side up. It is also common to refer to the probability of an event occurring using the notation $P(X = x)$. In this example, we assume all six possible outcomes are equally likely (fair die), and therefore, $P(X = x) = 1/6$ for $x = 1, 2, 3, 4, 5$, or 6.

EXAMPLE 1.6 PROBABILITY EXPRESSION FOR CDF

The CDF $F(x)$ can be defined as $F(x) = P(X \le x)$, that is, $F(x)$ is the probability that the random variable X has a value less than or equal to x. Similarly, the survival function can be defined as $S(x) = 1 - F(x) = P(X > x)$.

1.7 Sample Estimates of Population Parameters

We have discussed descriptive techniques such as histograms to represent observations. However, in order to complement the visual impression given by the frequency histogram, we often employ numerical descriptive measures called *parameters* for a population and *statistics* for a sample. These measures summarize the data in a population or sample and also permit quantitative statistical analysis. In this way, the concepts of central tendency, spread, shape, symmetry, and so on take on quantifiable meanings.

For example, we state that the frequency distribution is centered about a given value. This central tendency can be expressed in several ways. One simple method is just to cite the most frequently occurring value, called the *mode*. For grouped data, the mode is the midpoint of the interval with the highest frequency. For the fuse data in Table 1.1, the mode is 5.05.

Another procedure involves selecting the *median*, that is, the value that effectively divides the data in half. For individual readings, the n data points are first ranked in order, from smallest to largest, and the median is chosen according to the following algorithm: the middle value if n is odd, and the average of the two middle values if n is even.

Alternatively, for data that has already been grouped or binned (Table 1.2), the median occurs in the interval for which the cumulative frequency distribution registers 50%; that is, a vertical line through the median divides the histogram into two equal areas. For grouped data with n points, to get the median, one first determines the number of observations in the class containing the middle measurement $n/2$ and the number of observations in the class to get to that measurement. For example, for the fuse data in Table 1.3, $n = 100$, and the middle value is the 50th point, which occurs in the class marked 4.895 to 4.995 (width 0.1). There are 41 data points before the interval and 13 points in this class. We must count 9/13 of the interval width to get to the median. Hence, the median is

$$4.895 + \left(\frac{9}{13}\right) \times 0.1 = 4.964$$

(In reliability work, it is common terminology to refer to the median as the T_{50} value for time to 50% failures.)

The most common measure of central tendency, however, is called the arithmetic *mean* or average. The sample mean is simply the sum of the observations divided by the number of observations. Thus, the mean, denoted by $\bar{X}$, of n readings is given by the statistic

$$\begin{aligned}\bar{X} &= \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \\ &= \frac{\sum_{i=1}^{n} X_i}{n}\end{aligned}$$

This expression is called a *statistic* because its value depends on the sample measurements. Thus, the sample mean will change with each sample drawn, which is another instance of the variability of the real world. In contrast, the population mean depends on the entire set of measurements, and thus it is a fixed quantity, which we call a *parameter*. The sample mean $\bar{X}$ estimates the population mean μ. We also mention here a notation common in statistics and reliability work. A parameter estimate is commonly denoted by a caret (^) over the parameter symbol. Thus, $\hat{\mu}$ is an estimate of the population mean μ and here $\hat{\mu} = \bar{X}$.

For a discrete (i.e., countable) population, the mean is just the summation over all discrete values where each value x_i is weighted by the probability of its occurrence p_i:

$$\mu = \sum_{i} x_i p_i$$

For a continuous population, the mean parameter is expressed in terms of the PDF model as

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

For reliability work involving time, the population mean is

$$\mu = \int_{0}^{\infty} t f(t)\, dt$$

An alternate expression for the mean of a lifetime distribution is sometimes easier to evaluate. The form of the equation, when a finite mean exists, is

$$\mu = \int_0^\infty \left[1 - F(t)\right] dt$$

(See Feller 1968, page 148, for a proof.)

A common practice in statistics is to refer to the mean for both discrete and continuous random variables as the *expected value* of the random variable and use the notation $E(X) = \mu$ or $E(T) = \mu$. We occasionally use this terminology in this text.

Knowing the center of the distribution is not enough; we are also concerned about the spread of the data. The simplest concept for variability is the *range*, the difference between the highest and lowest readings. However, the range does not have very convenient statistical properties, and therefore, another measure of dispersion is more frequently used. This numerical measure of variation is called the *variance*. The variance has certain statistical properties that make it very useful for analysis and theoretical work. The variance of a random variable X is defined as the expected value of $(X - \mu)^2$, that is, $V(x) = E([X - \mu]^2)$. An alternative formula is $V(x) = E[X^2] - \mu^2$. For continuous data, the population variance for common reliability analysis involving time is

$$V(t) = \sigma^2 = \int_0^\infty (t - \mu)^2 \, dt$$

In engineering terms, we see that the variance is the expected value of the second moment about the mean.

The square root of the variance is called the *standard deviation*. The standard deviation is expressed in the same units as the observations. The sample standard deviation is denoted by s and the formula is

$$s = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n - 1}}$$

Since $\bar{X}$ is used in the formula rather than the population mean, statistical theory shows that dividing by $n - 1$ gives a better (i.e., unbiased) estimate of the population variance (denoted by $\hat{\sigma}^2 = s^2$) than just dividing by n. Alternatively, we may state that one degree of freedom has been taken to estimate the population mean μ using $\bar{X}$.

We have defined numerical measures of central tendency $\left(\bar{X}, \mu\right)$ and dispersion (s^2, σ^2). It is also valuable to have a measure of symmetry about the center and a measure of how peaked the data is over the central region. These measures are called *skewness* and *kurtosis*, and are respectively defined as expected values of the third and fourth moments about the mean, that is,

$$\text{skewness: } \mu_3 = E[(X - \mu)^3]; \quad \text{kurtosis: } \mu_4 = E[(X - \mu)^4]$$

Symmetric distributions have skewness equal to zero. A unimodal (i.e., single peak) distribution with an extended right "tail" will have positive skewness and will be referred to as skewed right; skewed left implies a negative skewness and a corresponding extended left tail. For example, the exponential distribution in Figure 1.3 is skewed right. Kurtosis, on the other hand, indicates the relative flatness of the distribution or how "heavy" the tails are.

Both measures are usually expressed in relative (i.e., independent of the scale of measurement) terms by dividing μ^3 by σ^3 and μ^4 by σ^4. The kurtosis estimate is also offset by an amount that goes to three as the sample size increases so that data from a normal population has a kurtosis of approximately zero. Sample estimates are calculated using the formulas

$$\text{Skewness estimate} = \frac{n}{(n-1)(n-2)}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{s}\right)^3$$

$$\text{Kurtosis estimate} = \frac{n(n-1)}{(n-1)(n-2)(n-3)}\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})}{s}\right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

These formulas are used by the spreadsheet SKEW and KURT functions and also by Minitab and JMP in their descriptive statistics platforms.

These various measures allow us to check the validity of the assumed model. Ott (1977) shows applications to the normal distribution. Table 1.5 contains a listing of properties of distributions frequently used in reliability studies.

The important statistical concept involved in sample estimates of population parameters (e.g., mean, variance, etc.) is that the population parameters are fixed quantities, and we infer what they are from the sample data. For example, the fixed constant θ in the exponential model $F(t) = 1 - e^{-t/\theta}$, where $\theta = 1/\lambda$, can be shown to be the mean of the distribution of failure times for an exponential population. The sample quantities, on the other hand, are random statistics that may change with each sample drawn from the population.

TABLE 1.5

Properties of Distributions Used in Reliability Studies

	Uniform	**Normal**	**Weibull**	**Exponential**	**Lognormal**	**Rayleigh (Weibull with Shape Parameter 2 and Linear Failure Rate)**	**Extreme Value**
Symmetric	Yes	Yes	No	No	No	No	No
Bell-shaped	No	Yes	No	No	No	No	No
Skewed	No Skew = 0	No Skew = 0	Yes (right)	Yes (right) Skew = 2	Yes (right)	Yes (right) Skew = 0.63	Yes (left) Skew = −1.14
Kurtosis	−1.8	0		6		0.26	2.4
Log data is symmetric and bell-shaped	No	No	No	No	Yes	No	No
Cumulative distribution shape	Straight line	"S" shaped		Exponential curve			

EXAMPLE 1.7 THE UNIFORM DISTRIBUTION

The uniform distribution is a continuous distribution with PDF for the random variable T given by

$$f(t) = \frac{1}{\theta_2 - \theta_1}, \quad \theta_1 \le t \le \theta_2$$

and zero elsewhere, where θ_1 and θ_2 are the parameters specifying the range of T. The rectangular shape of this distribution is shown in Figure 1.7.

We note that $f(t)$ is constant between θ_1 and θ_2. The CDF of T, denoted by $F(t)$, for the uniform case is given by

$$F(t) = \frac{t - \theta_1}{\theta_2 - \theta_1}$$

Thus, $F(t)$ is linear in t in the range $\theta_1 \le t \le \theta_2$, as shown in Figure 1.8.

EXERCISE 1.9

Show that the uniform distribution has expected value $E(t) = (\theta_1 + \theta_2)/2$ and variance

$$V(t) = \frac{(\theta_2 - \theta_1)^2}{12}$$

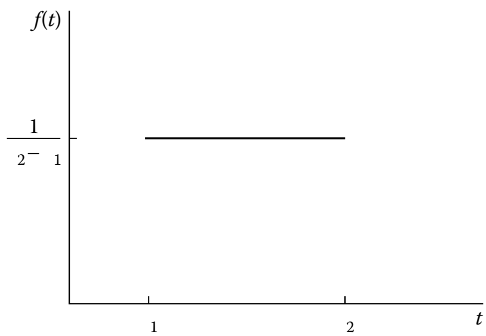


FIGURE 1.7
The uniform PDF.

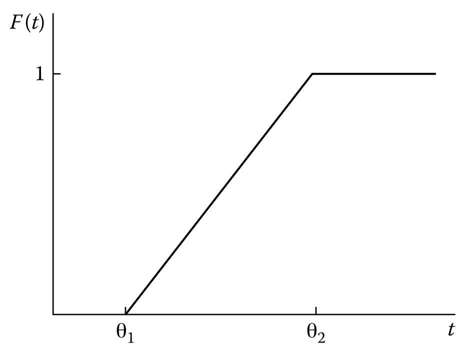


FIGURE 1.8
The CDF for the uniform distribution.

EXERCISE 1.10

The uniform distribution defined on the unit interval [0,1] is a popular and useful model—so much so that the name uniform distribution is often taken to refer to this special case. Find $f(u)$, $F(u)$, $E(u)$, and $V(u)$ for this distribution.

EXERCISE 1.11

Let $F(t)=1-(1+t)^{-1}$, $0 \le t \le \infty$. This is a legitimate CDF that goes from 0 to 1 continuously as t goes from 0 to ∞. Find the PDF and the T_{50} for this distribution. Try to calculate the mean. (Hint: Use either integration by parts or the alternate formula given in the text for calculating the mean.)

EXAMPLE 1.8 THE BETA DISTRIBUTION

The (standard) beta distribution, like the uniform distribution discussed in Exercise 1.10, is also defined on the unit interval [0,1]. However, it is a far more flexible distribution and even includes the uniform distribution as a special case. Its flexibility is one of the reasons it is an excellent choice for modeling numbers between 0 and 1, such as probabilities or proportions.

For a random variable X having a beta distribution with parameters $a > 0$ and $b > 0$, the PDF in the unit interval is given by

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

where $B(a,b)$, in terms of gamma functions (see the discussion in Chapter 4 after Exercise 4.6), is

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

For a and b integers,

$$B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$$

where we use the factorial notation $a!$ to represent $a(a-1)(a-2)\cdots 1$. Note that when $a = b = 1$, the beta distribution is the same as the uniform distribution on [0,1].

The CDF of the beta distribution is commonly called the *incomplete beta function*. For any $0 < x < 1$, the incomplete beta function $F(x)$ is given by

$$F(x) = \int_0^x \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}\,dy = I_x(a,b)$$

Like the normal distribution (covered in Chapter 5), this integral cannot be written in closed form. Tables of the incomplete beta function are available (Pearson 1968). However, both Excel and OpenOffice provide the worksheet function BETADIST for the beta CDF. The arguments are x, a, b, respectively. In Excel, for example, = BETADIST(0.5,2,3) returns the result 0.6875.

Figure 1.9 shows a variety of beta density functions for different values of the parameters a and b. The incomplete beta function $I_x(a,b)$ is closely related to the binomial distribution, a key distribution used in quality control and other sampling applications. This important relationship will be covered in Chapter 10.

EXERCISE 1.12

Find the expected value (mean) for a random variable having a beta density function defined on the unit interval, with parameters a and b. What relationship must these parameters have in order for the mean to be located in the center of the interval (or $\mu = 0.5$)?

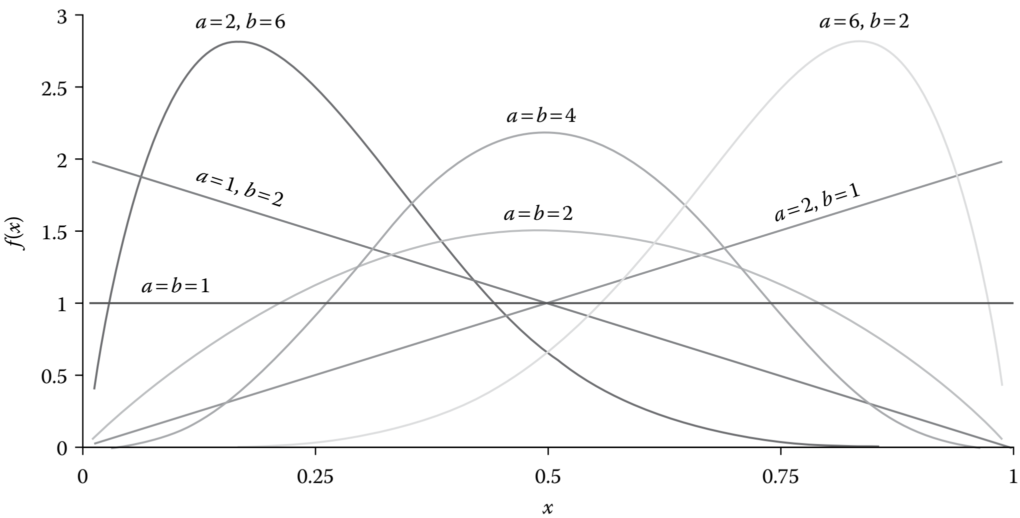


FIGURE 1.9
Beta density functions.

1.8 How to Use Descriptive Statistics

At this point, it is important to emphasize some considerations for the analyst. No matter what summary tools or computer programs are available, the researcher should always "look" at the data, preferably in several ways. For example, many data sets can have the same mean and standard deviation and still be very different—and that difference may be of critical significance (see Figure 1.10 for an illustration of this effect).

Generally, the analyst will start out with an underlying model in his mind based on the type of data, where the observations came from, previous experience, familiarity with probability models, and so on. However, after obtaining the data, it is necessary that the analyst go through a verification stage before he blindly plunges ahead with his model. This requirement is where the tools of descriptive statistics are very useful. Indeed, in many cases we utilize descriptive statistics to help us choose an appropriate model right at the start of our studies. Other useful graphical techniques include Boxplots, dot plots, stem and leaf plots, 3D plots, and so on (see Chambers et al. 1983 for further information on graphical analysis).

In this text, we focus on several key continuous distributions that are most applicable to reliability analysis: the exponential, Weibull, normal, and lognormal distributions. By learning what these distributions should look like, we can develop a yardstick to measure our data for appropriateness to some model. Graphics (frequency histograms, cumulative frequency curves) and summary values (mean, median, variance, skewness, etc.) are the means by which the characteristics of distributions are understood. In Chapter 6, we shall introduce other valuable descriptive procedures such as probability plotting.

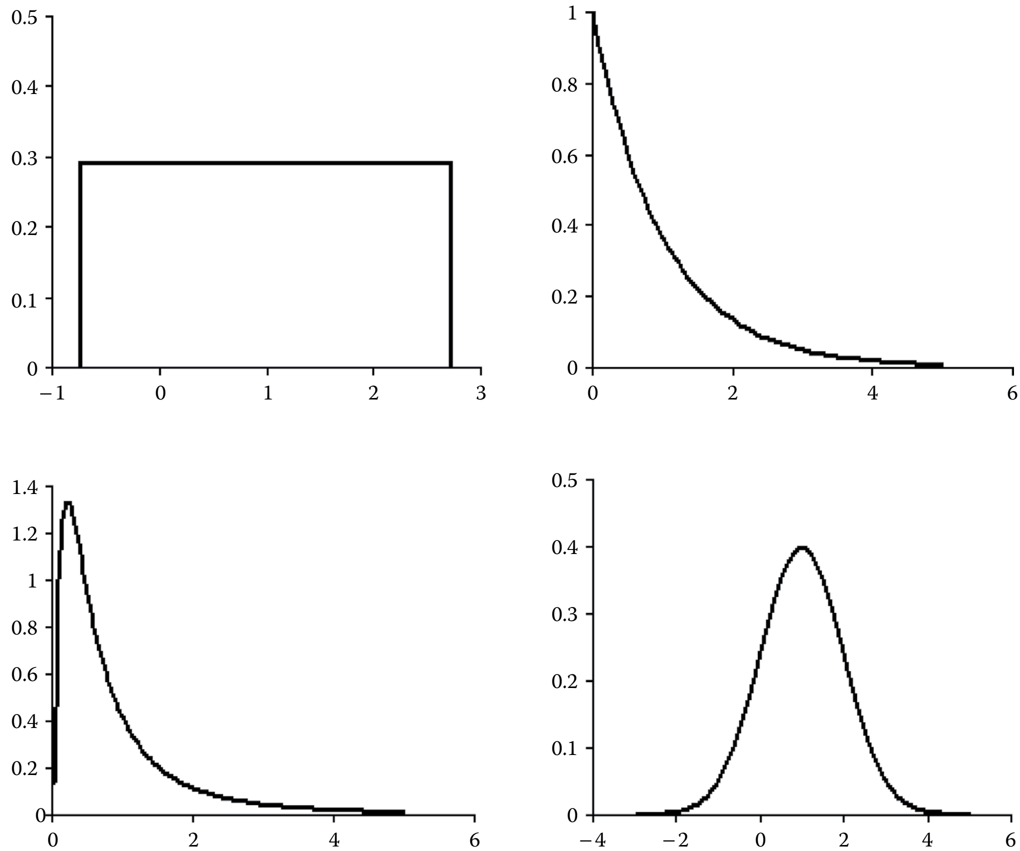


FIGURE 1.10
Mean and Sigma do not tell us enough. These four distributions have the same mean and standard deviation.

1.9 Data Simulation

Many different PDFs (and CDFs) exist, and reliability studies are often concerned with determining what model is most appropriate for the analysis. In reliability work, one may wish to simulate data from various distributions in order to do the following:

1. Determine percentiles of complicated distributions that are functions of common distributions.
2. Evaluate the effectiveness of different techniques and procedures for analyzing sample data.
3. Test the potential effectiveness of various experimental designs and sample size selections.
4. Illustrate statistical concepts, especially to understand the effects of variability in data.

Computer programs that will generate random variables from almost any desired distribution are available. However, there is a simple and general technique that allows us to produce what are called *pseudorandom numbers* from many common distributions. (The term *pseudorandom* is used because a specific computer algorithm generates the numbers to be as nearly random as possible.) To begin, we need a good table of random numbers or we can use a spreadsheet function to generate random numbers.

For simplicity, we consider only distribution functions $F(x)$ that are continuous and map one-to-one onto the unit interval (0,1), that is, $0 \le F(x) \le 1$. This class includes all the life distributions discussed in this text. Let $F(x) = u$. Then, we can define an inverse function $F^{-1}(u) = x$ that provides the specific percentile corresponding to the CDF value in the unit interval. For example, given $F(x) = .5$, then $F^{-1}(.5) =$ the median, which is the 50th percentile. F and its inverse have the following properties: $F\left(F^{-1}(u)\right) = u$ and $F^{-1}\left(F(x)\right) = x$.

To generate a random sample $x_1, x_2, \ldots, x_n$ from $F(x)$, first generate a random sample $u_1, u_2, \ldots, u_n$ from the uniform distribution defined on [0,1]. This procedure is done with random numbers. For example, if a five-digit random number is obtained from a table or a spreadsheet, divide the number by 100,000 to obtain a pseudorandom number from the uniform distribution. (The spreadsheet function RAND() provides random numbers directly in the unit interval.) Next, set $x_1 = F^{-1}(u_1), x_2 = F^{-1}(u_2), \ldots, x_n = F^{-1}(u_n)$. It is easy to show that the sample of x's is distributed according to the $F(x)$ distribution. (See the hint to Problem 1.4 at the end of this chapter.)

EXAMPLE 1.9 DATA SIMULATION

Let $F(t)$ be the distribution given in Exercise 1.11. Generate a sample of five random times from this distribution.

SOLUTION

We obtain F^{-1} by solving for t in $F(t) = u = 1-(1+t)^{-1}$ to get $t = u/(1-u) = F^{-1}(u)$. Next, we use a random number generator via a spreadsheet function to obtain the uniform distribution sample (0.880, 0.114, 0.137, 0.545, 0.749). Transforming each of these by F^{-1} gives the values $t_1 = 0.880 \div (1-0.880) = 7.333$, $t_2 = 0.129$, $t_3 = 0.159$, $t_4 = 1.198$, and $t_5 = 2.984$. The sample $(t_1, t_2, t_3, t_4, t_5)$ is the desired random sample from F.

In a typical reliability experiment, n units are placed on stress, and the exact times to failure are recorded. The successive failure times naturally occur in increasing order, that is, the first failure time is less than the second failure time, the second is less than the third, and so on. This property of ordered failure times is a key characteristic of reliability work. In contrast, consider selecting n individuals randomly and measuring, for example, their weight or height. The successive observations will not necessarily occur in increasing order. Consequently, in simulating random variables for reliability studies, one would like the values arranged in increasing order. For a single set of simulated observations, one could do a manual sort using the spreadsheet **sort** routine available under the menu item **Data**. However, for repeated simulations (involved in Monte Carlo studies), a non-manual procedure is desirable. In Excel (and OpenOffice), the spreadsheet function PERCENTILE(array, k) can be used. This function returns the kth percentile (where $0 \le k \le 1$) of values in the range defined by the array. The trick is to choose the k values to be the $(i-1)$ multiples of $1/(n-1)$, where $i = 1, 2, \ldots, n$ is the failure count and n is the sample size. We illustrate the procedure in the Example 1.10.