Chapter 14

Some issues in statistical applications: an overview

Chapter outline

Objective

In this chapter we discuss some general concepts and useful methods with applications to real-world problems.

Florence Nightingale
(Source: http://commons.wikimedia.org/wiki/File:Florence_Nightingale_1920_reproduction.jpg.)

Mathematical Statistics with Applications in R. https://doi.org/10.1016/B978-0-12-817815-7.00014-2

Florence Nightingale (1820−1910) is most remembered as a pioneer of nursing and a reformer of hospital sanitation methods. Her statistical contributions caused Karl Pearson to acknowledge Nightingale as a "prophetess" in the development of applied statistics. Nightingale used data as a tool for improving medical and surgical practices. During the Crimean War, she plotted the incidence of preventable deaths in the military and introduced polar-area charts to demonstrate the unnecessary deaths due to unsanitary conditions. With her analysis, Florence Nightingale showed the need for reform and revolutionized the idea that social phenomena could be objectively measured and subjected to mathematical analysis. In addition, she developed a model hospital statistical form for hospitals to collect and generate data and statistics. She became a Fellow of the Royal Statistical Society in 1858 and an honorary member of the American Statistical Association in 1874.

14.1 Introduction

Basically, there can be three major problems in applying the statistical methods that we have studied in the previous chapters to real-world problems. These involve sources of *bias*, *errors in methodology*, and the *interpretation of the analytical results.* Bias occurs in situations or conditions that affect the validity of statistical results. For the statistical inferences to be valid, the observed sample must be representative of the target population, and the observed variables must conform to assumptions that underlie the statistical procedures to be used. Of course, the statistical methodology chosen must also be appropriate for the problem under study. We must be careful with the interpretation of the statistical results. For example, in a regression problem, a cause-and-effect relationship may not be warranted, or in a hypothesis testing problem, we may not accept the null hypothesis, without exploring the probability of type II error. If we present the results graphically, the graphs should be accurate and reflect the data variations clearly.

In this textbook, we have assumed that a data set is available to us. Either it is a small data set that we can handle without much effort or it is in a computer-readable file. In practical situations, the proper handling of a statistical data set is not an easy task. Going from a stack of disorganized hard copy to online data that are trustworthy, that is, to input, debug, and manipulate the data, is a problem one will face even before one starts the statistical analysis. Here, we will not be dealing with these issues. Interested readers should refer to the references at the end of this book for further study on these aspects.

It is not our aim to discuss comprehensively all the problems that come up in applications. Most of the material presented in this chapter has already been discussed in various parts of the book. One of the problems we face when we study a book of this sort is that, for the problems of each chapter, say, Chapter 6 on hypothesis testing, we know that we need to use only the techniques of that section, or at most of that chapter. For the parametric analysis, in Chapter 11, we gave ways to do goodness-of-fit for choosing a particular distribution. In a real-world situation, we will not be able to look at the data analysis in a chapter-by-chapter manner. The purpose of this chapter is to present some methods in a unified way and to discuss generally the various ways in which the techniques developed in previous chapters could be applied to real-world data. Because the material in this chapter is a collection of available techniques, we will not follow the more rigorous pattern of previous chapters, and no proofs will be given.

It is very important to mention that every parametric statistical method and also some nonparametric methods are subject to certain assumptions, and when we apply them to real-world problems, we should make every effort to justify these assumptions. If you cannot, it is necessary, when you conclude your analysis and make decisions, that you state that your results are subject to certain assumptions that you could not justify.

14.2 Graphical methods

We first present some useful graphical methods that were not introduced in Chapter 1 on descriptive statistics. Graphical analysis is a very important aspect of any statistical study. Before attempting a complex statistical analysis, summarize the data with a graph. Graphical displays of data analysis help in data exploration, analysis, and presentation and in communication of results. In data analysis, one of the significant steps is to summarize and plot the data. Graphs help in the communication of final results and recommendations inferred from quantitative models. A statistical model is often suggested by an initial graphical analysis. Adequacy of statistical models depends on the model conditions. Because the violations of these model assumptions may sometimes occur as nonlinearities, graphical methods provide an easy and perhaps very effective method of detection. Some examples of graphical displays are histograms, dot plots, box plots, and scatterplots. Methods of graphing multivariate data are more complex and include scatterplot matrices and icon plots. These are beyond the level of this book.

If we have a data set with one variable (univariate), we first create a dot plot and summary of basic statistics. In a dot plot, we plot the data as dots (one dot for each observation) above the horizontal axis that covers the entire range of observations (see Fig. 14.1). The dot plot will provide us with an idea of the distribution of the data and any unusual behavior of the data that may not be apparent from summary statistics such as mean, median, or standard deviation. The dot plots allow us to visualize the entire distribution of the data set by listing each possible outcome and the frequency of the variable. Other ways of summarizing univariate data, such as histograms, have been discussed in Chapter 1. The histogram differs from the dot plot in that it groups data into categories. We illustrate these problems with several examples.

EXAMPLE 14.2.1

The following data give the lifetimes of 30 light bulbs (rounded to nearest hour) of a particular type:

1122	922	1146	1120	1079	905	1095	977	1138	966
1150	977	1137	1088	1139	1055	1082	1053	1048	1132
1088	996	1102	1028	1130	1002	990	1052	1116	1135

Construct a dot plot.

Solution

Fig. 14.1 is the dot plot for these data.

The dot plot suggests a distribution that is skewed toward the right, because most of the observations are located to the right.

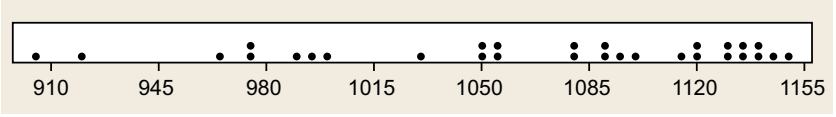

FIGURE 14.1 Dot plot for lifetime of light bulbs.

Some of the graphing methods can also be applied to compare two variables—for example, their frequency distributions. For instance, dot plots could also be used to compare bivariate (two variables) or multivariate (many variables) data. When we have independent samples, *side-by-side box plots* could be used for comparing two-sample distributions in terms of their centers, dispersions, and skewnesses.

When there are two variables, a *scatterplot* is used as one of the basic graphic tools to examine the relationship between two variables.

The scatterplot in Fig. 14.2 for two variables, x and y, indicates a possible linear relation between x and y. The strength of the relationship between two variables is often represented through a correlation statistic. It should be noted that the correlation coefficient is a single number that is easy to calculate and comprehend, though it measures only the strength of a linear relationship and hence is often used as the primary statistic of interest. However, scatterplots provide information

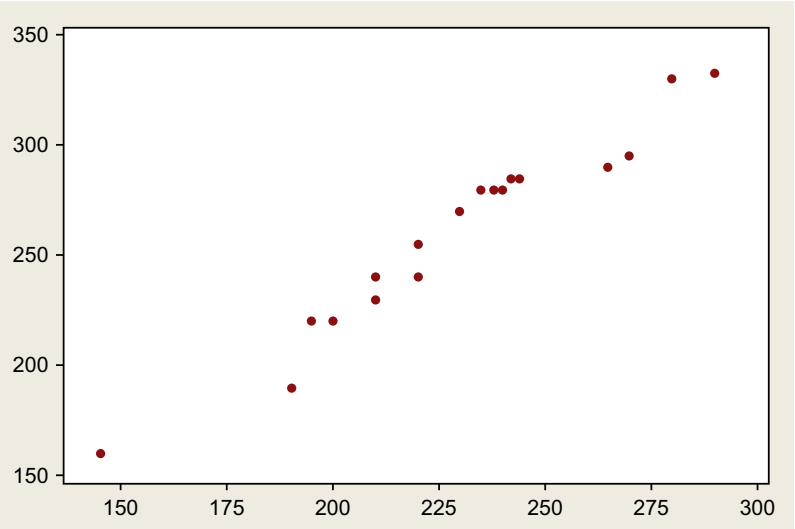

FIGURE 14.2 Scatterplot.

about the strength of association, not necessarily linear, between variables. In addition, scatterplots help us understand other aspects of the data, such as the range. Given n observations on two variables, X and Y, we plot a character or symbol at n points representing (x_i, y_i). If two or more observations in a scatterplot are identical, the plotted symbols will coincide, masking possibly important information.

EXAMPLE 14.2.2

The following data give the cholesterol levels before a certain treatment and after 4 months of the treatment:

Before	235	212	277	262	162	212	226	252	185	276
	216	315	289	283	234	223	275	282	311	285
After	233	214	200	266	146	212	238	284	191	247
	244	268	241	289	220	202	221	196	212	247

Draw a scatterplot. Also find the correlation between before-treatment and after-treatment values.

Solution

Fig. 14.3 is a scatterplot of the data.

Looking at the scatterplot in Fig. 14.3, we see a trend in the cholesterol levels before and after the treatment. Correlation of before-treatment and after-treatment data is measured by r*, the correlation coefficient, and is given by:*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}.$$

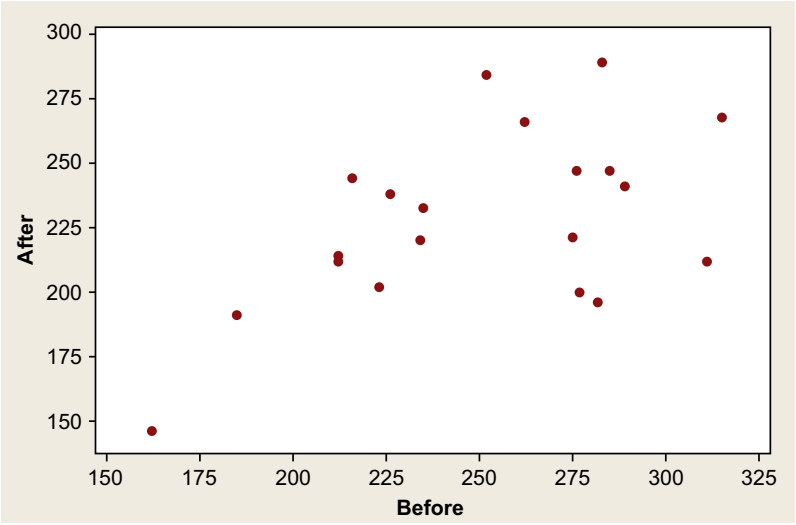

FIGURE 14.3 Scatterplot for cholesterol levels.

The *quantile−quantile* (QQ) *plot* is another useful technique for comparing bivariate data. In a QQ plot, the quantiles of the two samples are plotted against each other. For two distributions that are almost the same, their quantiles would be nearly equal. As a result, the quantiles would plot along the 45-degree line. Deviation of plots from this line can be used to draw inferences about how the two samples differ from one another. If the two sample sizes n_1 and n_2 are equal, then we can draw the QQ plot by graphing the order statistics $x_{(i)}$ and $y_{(i)}$ against each other. If the two samples are not of the same size, then we can use the following procedure to create the QQ plot. If $n_1 > n_2$, then draw the $(1/(n_i + 1))$th quantiles of the two samples against each other. For a large sample, they are the order statistics, $x_{(1)} < \ldots < x_{(n_1)}$. For the smaller sample sizes, the pth quantile value is obtained by using the following formula:

$$\widetilde{x}_p = \begin{cases} x_{p(n+1)}, & \text{if } p(n+1), \text{ is an integer} \\ x_{(m)} + [p(n+1) - m]\left(x_{(m+1)} - x_{(m)}\right), & \text{if } p(n+1), \text{ is a fraction} \end{cases} \tag{14.1}$$

where m denotes the integer part of $p(n + 1)$. It should be noted that a QQ plot is not useful for paired data because the same quantiles based on the ordered observations do not, in general, come from the same pair.

EXAMPLE 14.2.3

Draw a QQ plot for the data given in Example 14.2.2.

Solution

Here $n_1 = n_2 = 20$. *First sort the data in ascending order:*

Before	162	185	212	212	216	223	226	234	235	252
	262	275	276	277	282	283	285	289	311	315
After	146	191	196	200	202	212	212	214	220	221
	233	238	241	244	247	247	266	268	284	289

Because the QQ plot points lie mostly below the 45-degree line, we may conjecture that the cholesterol level before is generally higher than that after (Fig. 14.4).

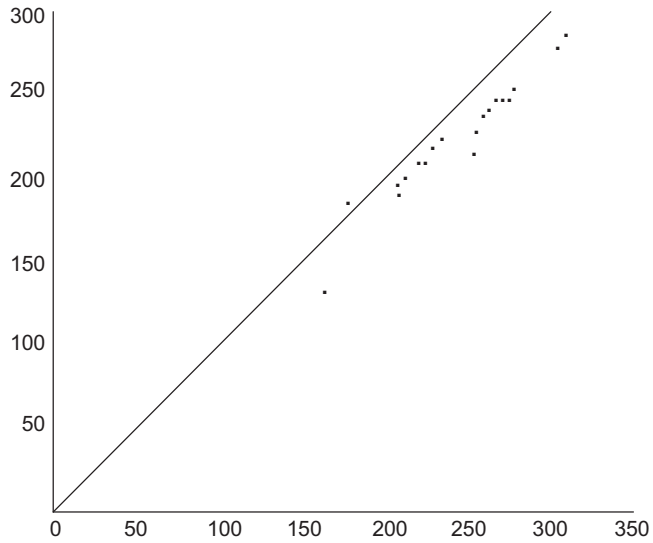

FIGURE 14.4 QQ plot for cholesterol levels.

We saw in Chapter 1 that box plots could be used for identification of outliers. To summarize, we emphasize that graphical procedures, although preliminary, are an integral part of any statistical analysis.

Exercises 14.2

14.2.1. To study any possible relationship between expense and return, the following data give percentage of expense ratio and total 1-year return for randomly selected stock mutual funds for the year 2000 (source: *Money*, February 2000):

% Expense ratio	1.03	1.80	1.90	1.53	1.03	2.06	3.20	0.49	1.10	1.07
	1.48	1.30	1.23	1.22	1.60	1.50	1.81	1.75	0.97	1.28
% Return	7.3	9.5	32.2	11.0	19.5	7.3	25.1	10.2	1.5	7.9
	18.9	26.1	3.4	3.7	23.5	2.9	14.5	14.9	22.7	21.9

Draw a scatterplot. Also find the sample correlation of percentage expense ratio and percentage return.

14.2.2. To study any possible relationship between age and change in systolic blood pressure (BP) (mm Hg) in 24 hours in response to a treatment, the following data were obtained from 11 individuals:

Age	70	51	65	70	48	70	45	48	35	48	30
Systolic BP change	−28	−10	−8	−15	−8	−10	−12	3	1	−5	5

(a) Draw a scatterplot.
(b) Find the sample correlation of age and systolic BP.
(c) Fit a least-squares regression line.
(d) Interpret (a), (b), and (c).

14.2.3. The following data represent 15 randomly selected state finances: revenue and expenditures (in millions of dollars) for the fiscal year 1997 (source: *The World Almanac and Book of Facts*, 2000).

Revenue	9,439	8,845	14,520	24,028	39,038	5,215	20,128	7,467
	26,538	5,537	6,494	2,818	49,318	4,229	7,724	
Expenditure	5,722	7,685	13,862	21,975	35,302	4,441	16,200	7,145
	25,791	4,808	5,130	2,426	39,296	4,002	6,818	

(a) Draw a scatterplot.
(b) Find the sample correlation between revenue and expenditure.
(c) Draw a QQ plot.
(d) Interpret (a), (b), and (c).

14.2.4. The following data give birth rates (per 1000 population) for 20 selected states in 1998 (source: *The World Almanac and Book of Facts*, 2000).

14.4 16.3 13.5 14.6 13.7 15.6 10.9 12.8 13.0 14.2
13.4 13.9 15.9 13.3 14.1 15.7 15.2 13.9 15.4 11.3

Construct a dot plot and interpret.

14.2.5. The following data give the median prices (rounded to nearest $1000) of single-family homes for 18 randomly selected US cities in 1998 (source: *The World Almanac and Book of Facts*, 2000).

128 146 109 90 105 152 79 89 109
93 108 128 188 158 93 78 123 137

Construct a dot plot and interpret.

14.3 Outliers

All statistical procedures make assumptions about a population and the sample values obtained from the population. Before we proceed to analyze the data, we must check to see if there are any outliers, that is, data points that do not belong in the data set or are not in line with the rest of the data.

Outliers are observations that appear to have an abnormal value compared with the rest of the values in the data set; that is, the value of an outlier is either much higher or significantly lower than any other value in the data set. An outlier could be a discordant observation or a contaminant. A discordant observation is one that appears surprising or discrepant to the investigator and is to some extent subjective. A contaminant is an observation that is from a different distribution compared with the rest of the data. Outliers may occur as a result of some limitations on measuring techniques or recording errors. They may also be due to the sample not being entirely from the same population. Extreme values in a data set could also be due to a skewed population. It should be noted that sometimes a data point that is labeled as an outlier may really be indicative of a novel phenomenon. In these cases, an extreme observation may not be classified as an outlier.

The presence of outliers can dramatically affect the estimate of the mean and variance of the sample, especially if the sample size is small. As a result, any test statistic computed from such data would be unreliable, and so would be the

statistical inferences. For example, the presence of outliers might lead to an incorrect conclusion that the variances of two samples are not equal, if the outlier is the result of a recording or measurement error.

In a controlled experiment, such as in a laboratory setting, good record keeping with a clear understanding of the phenomenon under investigation and information about all the data will minimize the occurrence of outliers due to recording errors.

What to do with outliers? As long as these points remain observations, we cannot throw them out on a whim. There are basically two methods that are employed in dealing with outliers. One method is to use statistical testing procedures to detect outliers, possibly removing them from the data set if we know that these are measurement errors, incorrectly entered values, or impossible values in real life, and letting the analysis deal only with the rest of the data. The second method is to use statistical procedures, such as nonparametric tests or data transformations, that are immune or only minimally sensitive to the presence of outliers. Of course, we could run the analysis both with and without the outliers and report both results. We now present some commonly used tests for labeling outliers.

In data analysis, it is necessary to label suspected outliers for further study. For normally distributed data, we give three simple methods to identify an outlier: z-score, modified z-score, and box plot.

In a ***z-test***, first find the z-scores of the entire data set and label any observation with a z-score greater than 3 or less than -3 as an outlier. Recall that for the observed values $x_1, \ldots, x_n$, the z-score is defined by:

$$z_i = \frac{x_i - \overline{x}}{s},$$

where s is the sample standard deviation of the sample, that is,

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Because both the sample mean and the sample standard deviation are affected by the outliers, this labeling method is not very reliable.

In a *modified z-score test,* the median of absolute deviation (MAD) is used. Let

$$MAD = median\ \ (|x_i - m|),$$

where m is the median of the observations. Then:

$$z_i = \frac{(x_i - \overline{x})}{MAD}.$$

An observation is labeled as an outlier if the corresponding modified z-score is greater than 3.5. A normal plot may be used for testing normality for the data.

If we want a reasonably robust *distribution-free test,* an observation x_0 is labeled as an outlier if:

$$\frac{|x_0 - m|}{MAD} > 5.$$

Here, the choice of 5 is somewhat arbitrary.

A box plot (also called a box-and-whisker plot) gives a method of labeling outliers through a graphical representation. We have seen the method of construction of box plots in Chapter 1. A box plot consists of a box, whiskers, and outliers. We draw a line across the box at the median. For example, in Minitab, the bottom of the box is at the first quartile ($Q1$) and the top is at the third quartile ($Q3$). The whiskers are the lines that extend from the top and bottom of the box to the adjacent values, the lowest and highest observations still inside the region defined by the lower limit $Q1 - 1.5(Q3 - Q1)$ and the upper limit $Q1 + 1.5(Q3 - Q1)$. Outliers are points outside the lower and upper limits, plotted with asterisks (*).

EXAMPLE 14.3.1

The following data give the hours worked by 25 employees of a company in a randomly selected week:

45	40	39	36	42	40	55	58	42	41
48	50	47	54	40	34	18	40	60	56
42	43	46	43	54					

Label all possible outliers using:

(a) The *z*-score test, distribution-free test, and modified *z*-score test.

(b) A box plot.

Solution

(a) *We can create Table 14.1, in which Dfree z stands for the distribution-free scores, and modified stands for the modified z-scores.*

By the z-score test, there are no outliers. Using the distribution-free test, the 18 is the only outlier. By the modified z-score test, 18 and 60 are possible outliers.

(b) *The box plot is given in Fig. 14.5.*

Hence, the observation 18 is identified as an outlier using the box plot.

TABLE 14.1 Hours Worked and Modified Scores.

Data	z-score	Dfree z	Modified
45	0.05355	0.12	0.12
40	−0.50427	1.13	−1.13
39	−0.61583	1.38	−1.38
36	−0.95053	2.13	−2.13
42	−0.28114	0.63	−0.63
40	−0.50427	1.13	−1.13
55	1.16919	2.62	2.62
58	1.50389	3.75	3.37
42	−0.28114	0.63	−0.63
41	−0.39271	0.88	−0.88
48	0.38824	0.87	0.87
50	0.61137	1.37	1.37
47	0.27668	0.62	0.62
54	1.05763	2.37	2.37
40	−0.50427	1.13	−1.13
34	−1.17366	2.63	−2.63
18	−2.95868	6.63	−6.63
40	−0.50427	1.13	−1.13
60	1.72701	3.87	3.87
56	1.28076	2.87	2.87
42	−0.28114	0.63	−0.63
43	−0.16958	0.38	−0.38
46	0.16512	0.37	0.37
43	−0.16958	0.38	−0.38
54	1.05763	2.37	2.37

Once we identify the outliers, then the question is what to do with them. If we can rule out recording errors as the source of outliers, the situation becomes more difficult. It is often impossible to say whether an outlier is really an extreme value within a skewed population or it represents a value drawn from a different population. As we indicated earlier, an outlier can be a legitimate observation representing a special feature of the sample population. In those cases, discarding the outliers may simplify the statistical analysis, although it also reduces the usefulness of such analysis. Understanding the experiment that generated the data might help in determining whether to discard or to keep the outliers.

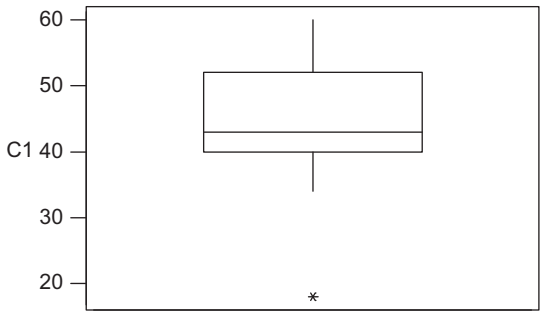

FIGURE 14.5 Box plot for hours of work per week.

Once we decide to include the outliers, there are two possible ways to deal with them. One is to transform the data, such as by taking the natural logarithm, so as to reduce the undue influence of the outliers. Another possibility is to perform the analysis twice, with and without outliers, and report both results.

If we have bivariate data, a scatterplot may reveal any possible outliers; see Fig. 14.27. There are other methods available to detect outliers in multivariate data.

Exercises 14.3

14.3.1. Motor vehicle thefts are a big problem in cities. Table 14.2 displays data on motor vehicle thefts per 100,000 population in the year 1997 for 15 randomly selected large US cities (source: *Statistical Abstract of the United States,* 1999).
Label all possible outliers using:
(a)
(i) The z-score test.
(ii) The distribution-free test.
(iii) The modified z-score test.
(b) A box plot.

14.3.2. Using the data of Example 14.2.1, label all possible outliers using:
(a)
(i) The z-score test.
(ii) The distribution-free test.
(iii) The modified z-score test.
(b) A box plot.

14.3.3. The following data represent test scores of 36 randomly selected students from a large mathematics class:

67	63	39	80	64	95	90	93	21	36	44	66
100	66	72	34	78	66	68	98	74	81	71	100
60	50	81	66	90	89	86	49	77	63	58	43

TABLE 14.2 Motor Vehicle Thefts per 100,000 Population.

Chicago, IL	1215.1	San Antonio, TX	830.0
Columbus, OH	1109.9	Charlotte, NC	780.1
Nashville, TN	1536.5	Tucson, AZ	1403.3
Albuquerque, NM	1797.8	Atlanta, GA	1869.7
Sacramento, CA	1630.5	St. Louis, MO	2152.8
Toledo, OH	939.7	Tampa, FL	1410.0
Birmingham, AL	1219.7	Anchorage, AK	532.8
Norfolk, VA	519.9		

Label all possible outliers using:

(a)

(i) The z-score test.

(ii) The distribution-free test.

(iii) The modified z-score test.

(b) A box plot.

14.3.4. The following data represent the number of days in 1997 on which selected US metropolitan areas failed to meet acceptable air-quality standards at trend sites (source: *The World Almanac and Book of Facts, 2000)*:

26	55	30	8	9	15	0	12	3	50	16
47	0	63	3	0	19	23	3	32	15	20
106	2	15	1	14	0	1	44	28		

Label all possible outliers using:

(a)

(i) The z-score test.

(ii) The distribution-free test.

(iii) The modified z-score test.

(b) A box plot.

14.4 Checking the assumptions

With some exceptions, checking data for agreement with assumptions is not a topic that is strongly emphasized in other textbooks at this level. Even in more advanced books, this step is frequently omitted. For the inferences to work correctly, the measured variables must conform to assumptions that underlie the statistical procedures, or methods, to be applied. In hypothesis testing such as the t-tests and analysis of variance (ANOVA), we made some fundamental assumptions that the random samples need to satisfy for the tests to yield correct results.

As an example, the basic assumptions underlying a t-test are:

(i) The sample comes from a normal population and is usually small, $n < 30$.

(ii) The sample is random. In cases of two-sample tests (excluding paired tests), the measurements in one sample are independent of those in the other sample.

(iii) When we are given two random samples, most of the results assume the equality of population variances, that is, $\sigma_1^2 = \sigma_2^2$. This assumption is called the homogeneity of variances. The test for equality of variance may have to be performed first if we doubt the equality of the variance.

Likewise, ANOVA is based on a model that requires the following three primary assumptions:

(i) The samples come from normal populations.

(ii) Each of the samples is randomly selected from each group, and the samples are independent of each other.

(iii) The population variances for all the samples are equal. That is, if we have k populations with variances σ_i^2, $i = 1, 2, \ldots, k$, then $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

When we say we have a random sample, we implicitly assume that the data are identically distributed. The presence of outliers in an observed sample may affect such an assumption. We now explain a few tests for checking these assumptions, such as the assumptions of normality, data transformations, and equality of variances.

14.4.1 Checking the assumption of normality

We start with the assumption of normality. Let us consider the example of randomly selected scores of 28 calculus students.

EXAMPLE 14.4.1

Given in the following table are the test scores of 28 randomly selected students from a calculus 1 class:

86	95	82	53	98	85	87	80	49	71	99	40	96	97
94	89	69	23	72	76	78	91	96	77	77	91	35	47

Construct a dot plot and a histogram, and compute the percentage of observations that fall in the intervals $\overline{x} \pm s, \overline{x} \pm 2s,$ and $\overline{x} \pm 3s$.

Solution

The dot plot is shown in Fig. 14.6.

The histogram is shown in Fig. 14.7.

We have $\overline{x} = 71.18$ *and* $s = 20.99$. *Also, 57% of the random sample (i.e., 16 observations) falls in the interval* $71.18 \pm 20.99 = (50.19, 92.17)$. *There are 27 observations, or about 96%, that fall in* $71.18 \pm 41.98 = (29.2, 113.16)$, *and all the observations fall in* $71.18 \pm 62.97 = (8.21, 134.94)$. *This suggests that the data set is approximately normally distributed. This procedure is the empirical rule.*

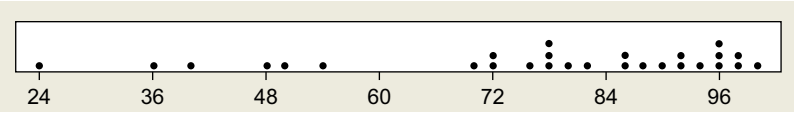

FIGURE 14.6 Dot plot of student scores.

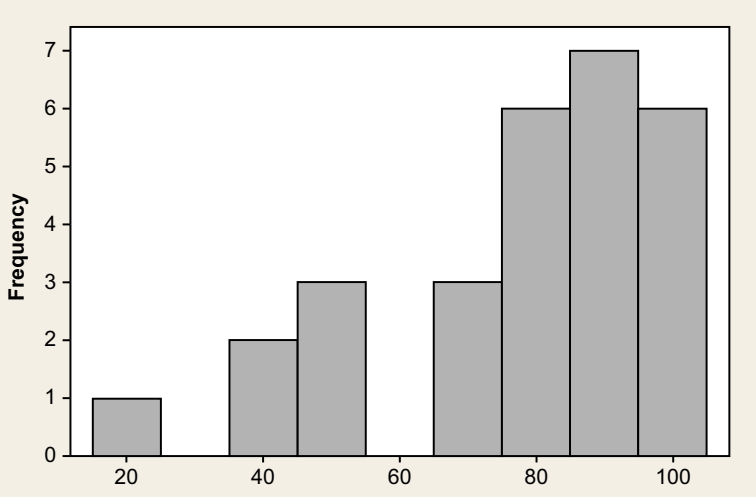

FIGURE 14.7 Histogram for student scores.

For the previous example, we have seen that the dot plot does not suggest any normality. A histogram also does not suggest any normality (see Fig. 14.7). However, if we used the empirical rule as a test for normality, the data suggest normality. Clearly this leads to a conflicting situation, with a simple theoretical check suggesting normality, while visual displays suggest nonnormality. In this case more sophisticated procedures are warranted.

Sometimes, skewness and kurtosis can be used to test for tilt in and peakedness of a distribution. After getting skewness and kurtosis from the descriptive statistics, divide these by the standard errors. If both skew and kurtosis are within the ± 2 range, the data can be considered normal.

We mention some sophisticated testing procedures for two of the most important of the parametric assumptions when running single-factor trials, namely, normality and homogeneity of variance. We have already seen in Project 4C how to construct a normal probability plot and to check for normality. In this chapter, we will use the Minitab normal plot to check for normality. Fig. 14.8 graphs a normal probability plot (using Minitab) for Example 14.4.1.

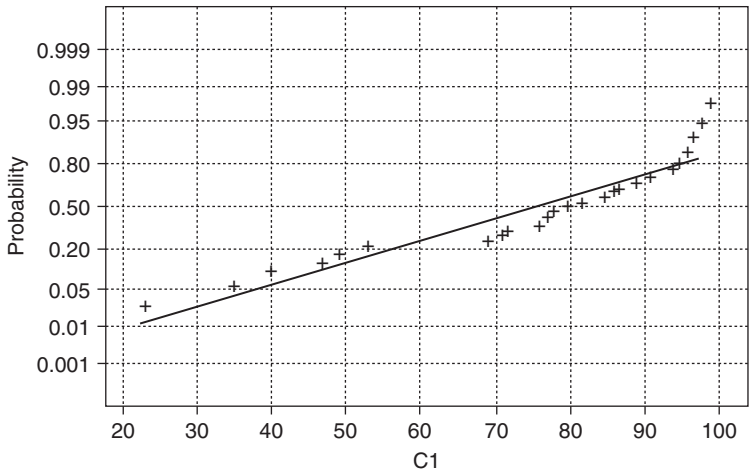

FIGURE 14.8 Normal probability plot of student scores.

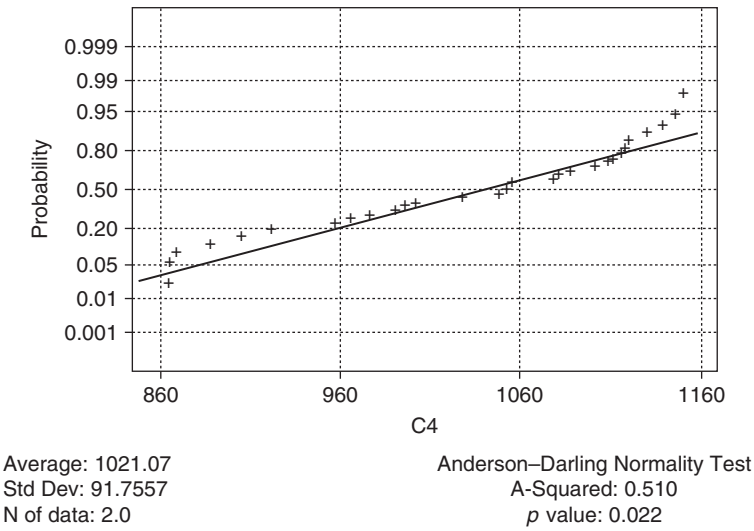

FIGURE 14.9 Normal probability plot for the lifetime of light bulbs.

We see that the test scores follow the straight line on the normal probability plot pretty well. The serious departures occur for the last four scores, because the values fall well above the line. This suggests normality with possible outliers.

It should be noted that for skewed data, in the normal probability plot, positively skewed data fall below the straight line, whereas the negatively skewed data rise above the straight line. A normal probability plot for the lifetime of 30 light bulbs in Example 14.2.1 is given in Fig. 14.9.

This graph suggests that the data may not be normal and are more toward negatively skewed. Fig. 14.10 is a normal probability plot for 30 data points generated from a standard normal distribution.

In this chapter, we have presented only simple graphical tests for testing of normality. We should mention that, in the literature, a variety of procedures for testing for normality are available, including the Kolmogorov−Smirnov test, the Shapiro−Wilk *W*-test, and the Lilliefors test. In Chapters 10 and 11, we learned how to use the Kolmogorov−Smirnov test, Anderson−Darling test, and chi-square test. Some of these tests are incorporated into statistical software packages such as R and Minitab and could be performed as easily as the graphical tests. If the sample size is very small, with any of these

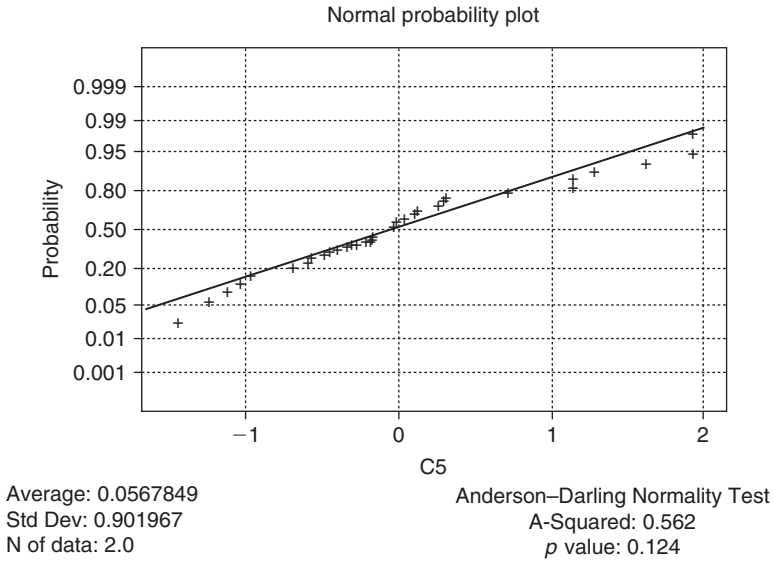

FIGURE 14.10 Normal probability plot of data from a standard normal distribution.

tests it may be difficult to detect assumption violations. It is important to keep in mind that these tests are only rough indicators of assumption violations. For small sample sizes, even when the tests show that none of the test assumptions is violated, a normality test may not have sufficient power to detect a significant departure from normality, although it is present.

14.4.2 Data transformation

Many data in real life do not meet the assumptions of parametric statistical tests: they may not be normally distributed, the variances may not be homogeneous, or both. Using most of the parametrical tests on those data may give a misleading result. Data transformation uses mathematical operations (filters) on each of the observations, transforming the original scores into a new set of scores. An appropriate transformation may (1) reduce the influence of outliers, (2) make data from a nonnormal distribution more normal, and/or (3) make the variances of different data sets more homogeneous. Some of the more commonly used transformations are (1) power transformations such as square root, (2) logarithm, (3) reciprocal, and (4) arcsine. Used correctly, data transformation can be a useful tool for the practitioner. Some of these transformations can be put into a popular class of transformations called the Box−Cox power law transformation,

$$y = \frac{x^{\lambda} - 1}{\lambda},$$

where λ can be optimally adjusted from 0 to 1. For example, as $\lambda \to 0$, we obtain the $y = \ln x$ (logarithmic filter) transformation, and when $\lambda = 1/2$, we get the square root transformation.

Even though we have done a statistical test on a transformed variable, it is not a good idea to report the summary statistics such as mean, standard errors, etc., in transformed units. We should back transform by doing the opposite of the mathematical function we used in the data transformation. For instance, if we had originally used the natural logarithm, we should use exponential transformation as the back transformation. For instance, if we got a symmetric confidence interval for transformed mean as in Chapter 5, which is symmetric for natural logarithm−transformed data, we should take exponentials of the lower and upper limits. In the process, we may lose the symmetry of the confidence interval.

As we have seen in Project 9A, it is sometimes possible to use appropriate data transformations to transform nonnormal data into approximately normal data. Then we can use this normality property to perform statistical analysis on these transformed values. For instance, if the distribution of data has a long tail (which could be seen by drawing a histogram of observations) or a few laggards on the right (which could be seen by drawing a dot plot of observations), the $\sqrt{x}$ or $\ln x$ transforms will pull larger values down further than they pull the smaller or center values. Sometimes it is necessary to try several different transformations (trial and error) to find one that is more appropriate.

EXAMPLE 14.4.2

Consider the following data from an experiment:

1.15	3.84	0.01	2.06	3.28	2.61	0.59	3.19	1.32	1.07
7.80	1.74	0.25	0.21	3.42	4.52	0.43	0.38	0.07	1.26
4.03	7.28	0.85	3.24	0.62					

(a) Draw a histogram and a normal plot.
(b) Take the transform $y = \sqrt{x}$ and draw a histogram and normal plot for the transformed data.

Solution

(a) *The histogram and normal plots for the data are shown in Figs. 14.11 and 14.12. These graphs clearly show that the data do not follow a normal distribution.*

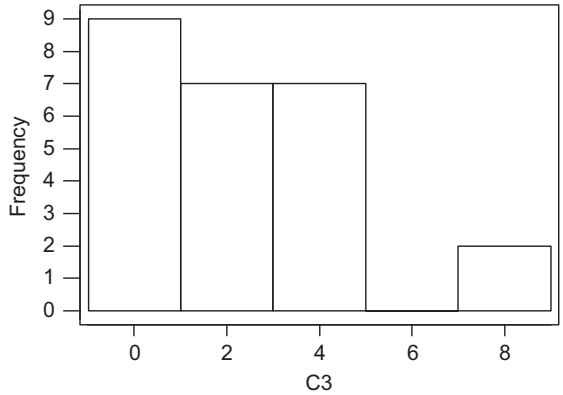

FIGURE 14.11 A histogram of the data.

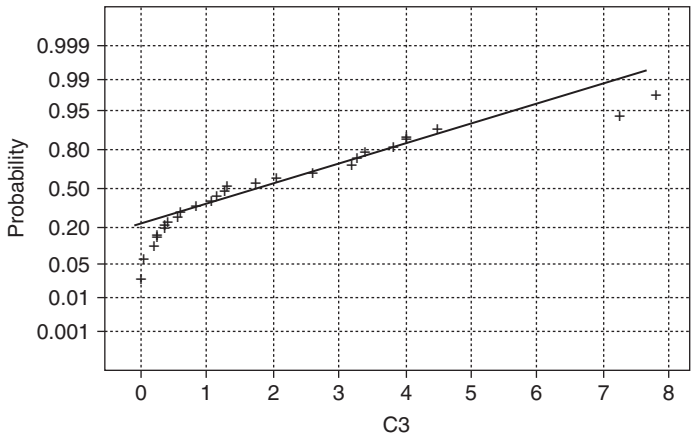

FIGURE 14.12 Normal probability plot of the data.

(b) *The histogram and normal plot for the transformed data are shown in Figs. 14.13 and 14.14. With this transformation (filter), we can see that the filtered data follow normality.*

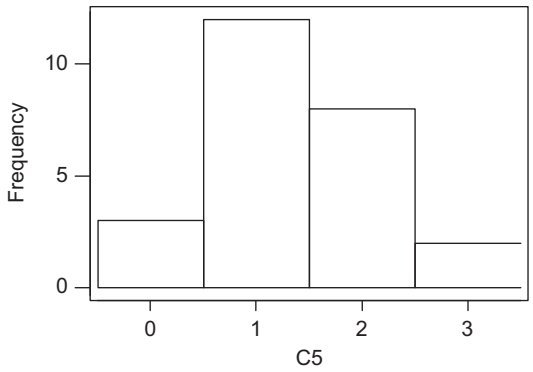

FIGURE 14.13 Histogram of the transformed data.

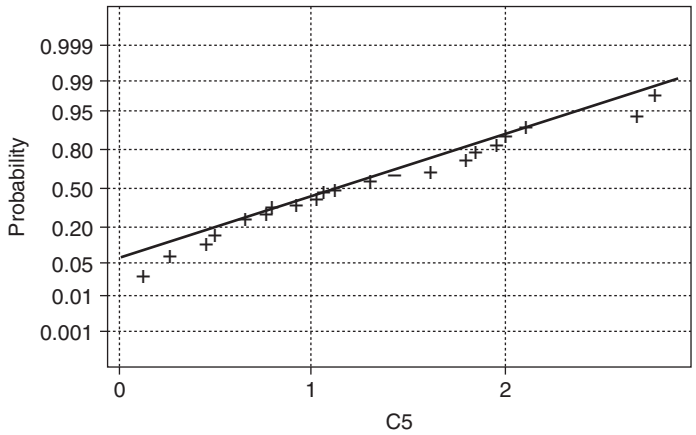

FIGURE 14.14 Normal probability plot of the transformed data.

We have only pointed out transformations in single-variable cases. The transformation methods are also useful in multivariable and multifactor studies; however, these involve more difficult analysis.

14.4.3 Test for equality of variances

Now we discuss the tests for equality of variances, that is, the tests for heteroscedasticity. Our recommendation is that, in a real-world problem, after accounting for outliers one should conduct tests for normality and heterogeneity of variance routinely before analyzing any data. Here, we give two tests. One, for the two-sample case, is based on the F-test, and for the multisampling case we give Levene's test based on ANOVA procedures. Albert Madansky's book *Prescriptions for Working Statisticians* (Springer-Verlag, 1988) gives various other tests for normality and heteroscedasticity.

14.4.3.1 Testing equality of variances for two normal populations

The following procedure has already been discussed in Chapter 6, Hypothesis testing. For the sake of completeness, here we again briefly discuss this procedure. Let $X_{11}, \ldots, X_{1n_1}$ be a random sample from an $N(\mu_1, \sigma_1^2)$ distribution and $X_{21}, \ldots, X_{2n_2}$ be a random sample from an $N(\mu_2, \sigma_2^2)$ distribution. Assume that $X_{1i}s$ and $X_{2j}s$ are independent of each other for all i, j. Let

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad i = 1, 2.$$

Assuming that μ_1 and μ_2 are unknown, we can test the hypothesis that $\sigma_1^2 = \sigma_2^2$ based on the ratio:

$$F = \frac{s_1^2}{s_2^2} = \frac{\sum_{j=1}^{n_1}\left(x_{1j} - \bar{x}_1\right)^2/(n_1 - 1)}{\sum_{j=1}^{n_2}\left(x_{2j} - \bar{x}_2\right)^2/(n_2 - 1)}.$$

We know that $(n_1 - 1)s_1^2/\sigma_1^2$ has a $\chi^2(n_1 - 1)$ distribution and $(n_2 - 1)s_2^2/\sigma_2^2$ has a $\chi^2(n_2 - 1)$ distribution. Therefore, under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, the F statistic has an $F(n_1 - 1, n_2 - 1)$ distribution.

Based on the alternate hypothesis, we will reject the equality of variance assumption if the test statistic falls into the appropriate tail of the F distribution. For example, if $H_a : \sigma_1^2 > \sigma_2^2$ with $\alpha = 0.05$, we would reject H_0 when $F > F_{0.95}(n_1 - 1, n_2 - 1)$, and if $H_a : \sigma_1^2 < \sigma_2^2$ with $\alpha = 0.05$, we would reject H_0 when $F \leq F_{0.05}(n_1 - 1, n_2 - 1)$. When $H_a : \sigma_1^2 \neq \sigma_2^2$ with $\alpha = 0.05$, we would reject H_0 when $F \geq F_{0.975}(n_1 - 1, n_2 - 1)$ or $F \leq F_{0.025}(n_1 - 1, n_2 - 1)$. It should be noted that in the case of a two-tailed alternative, this procedure is not the best one in the sense of minimizing the type II error. However, for simplicity, we will not discuss the optimal two-tailed procedure.

EXAMPLE 14.4.3

An aquaculture farm takes water from a stream and returns it after it has circulated through the fish tanks. Suppose the owner thinks that, because the water circulates rather quickly through the tank, there is little organic matter in the effluent. To find out, some samples of the water are taken at the intake and other samples are taken at the downstream outlet, and tests are performed for biochemical oxygen demand (BOD). If BOD increases, it can be said that the effluent contains more organic matter than the stream can handle. Table 14.3 gives the data for this problem.

(a) Using normal plots, check for normality of each sample.

(b) Test for the equality of variances of the BOD for the downstream and upstream samples at $\alpha = 0.05$.

Solution

(a) *The normal plots are shown in Figs. 14.15 and 14.16.*
The BOD data for the downstream and upstream samples are approximately normal.

(b) *We test* $H_0 : \sigma_1^2 = \sigma_2^2$ *versus* $H_a : \sigma_1^2 \neq \sigma_2^2$. *We have* $n_1 = n_2 = 10$, *and* $\alpha = 0.05$. *Because the normal plots of each sample conform with the normality assumption, we can use the F-statistic:*

$$F = \frac{s_1^2}{s_2^2} = \frac{(0.729)^2}{(0.654)^2} = 1.2425.$$

TABLE 14.3 Biochemical Oxygen Demand.

Upstream	Downstream
7.863	8.132
5.714	9.128
5.871	7.574
6.479	8.678
7.124	9.336
7.539	8.798
6.682	8.457
5.877	9.756
6.227	8.548
6.771	7.992

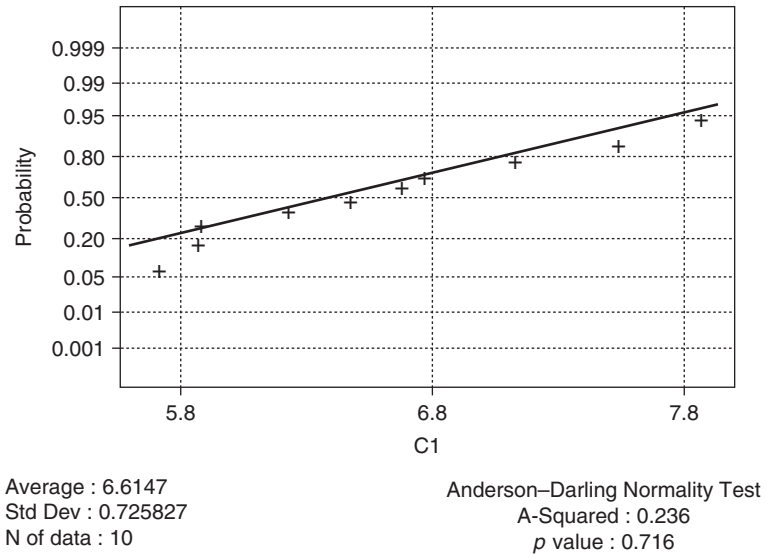

FIGURE 14.15 Normal plot of upstream data.

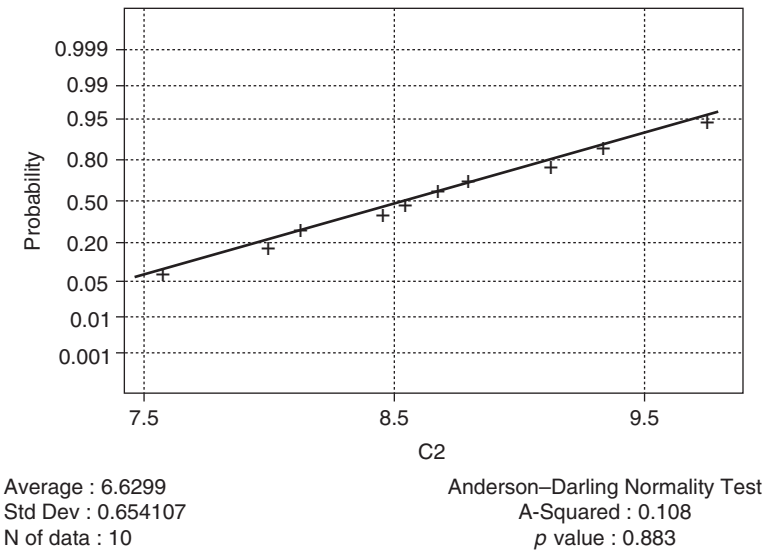

FIGURE 14.16 Normal plot of downstream data.

From the F *table, the rejection region is* $\{F \leq F_{0.025}(9, 9) = 0.248\}$ *or* $\{F > F_{0.975}(9, 9) = 4.03\}$. *Because the observed value of the test statistic does not fall in the rejection region, we conclude based on the sample evidence that the variances of the two populations are equal.*

14.4.3.2 Test for equality of variances, $k \geq 2$ populations

Generalizing to k populations, let $X_{i_1}, X_{i_2}, \ldots, X_{in_i}$, $i = 1, 2, \ldots, k$, be k random samples from $N(\mu_i, \sigma_i^2)$ distributions, with both $\mu_i s$ and $\sigma_i s$ unknown. Also assume that X_{ij}, X_{kl} are independent for all (i, j), (k, l). We wish to test the hypothesis H_0 : $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$ against H_a: at least one of the σ_i^2 is different. There are many tests available. One of the basic graphical procedures is to use side-by-side box plots (see Example 9.3.2). We describe Levene's test based on the ANOVA (source: Levene, 1960).

Let $y_{ij} = \left|x_{ij} - \overline{x_i}\right|$. Now perform an ANOVA for equality of the means of the y_{ij}. Let

$$n = \sum_{i=1}^{k} n_i, \quad \overline{y}_{i.} = \sum_{j=1}^{n_i} y_{ij} / n_i \quad \text{and} \quad \overline{y}_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} / \sum_{i=1}^{k} n_i.$$

The ANOVA statistic is given by:

$$z = \frac{\sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}_{..}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{i.}\right)^2 / (n-k)} = \frac{MST}{MSE}.$$

Recall that MST (mean square for treatments) and MSE (mean square error) were defined in Section 9.3; the MST is a measure of the variability between the sample means of the groups and the MSE is a measure of variability within the groups. For a 95% confidence level, the rejection region is $\{z > F_{0.95}\,(k-1,\, n-k)\}$.

It should be noted that y_{ij} is not independent, but the ANOVA is found to be robust against the deviation from this assumption of independence.

EXAMPLE 14.4.4

The three random samples in Table 14.4 are independently obtained from three different normal populations.

At the $\alpha = 0.05$ level of significance, test for the equality of variances.

Solution

We test $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$ *versus* H_a*: not all the* σ_i^2 *are equal. For this sample,* $\overline{x}_1 = 76$, $\overline{x}_2 = 66.33$, *and* $\overline{x}_3 = 85.67$. *Also* $n = 11$ *and* $k = 3$. *Letting* $y_{ij} = \left|x_{ij} - \overline{x}_i\right|$, *we obtain the following* y_{ij} *values:*

12	10.33	4.67000
8	7.67	6.33000
1	2.67	1.67000
1		
4		

The test statistic is:

$$z = \frac{\sum_{i=1}^{k} n_i \left(\overline{yi.} - \overline{y..}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{yi.}\right)^2 / (n-k)}$$

$$= \frac{MST}{MSE} = \frac{5.5}{16.5} = 0.33.$$

From the F *table, the 95% point is* $F_{0.05}(2, 8) = 4.46$. *Hence, the rejection region is* $\{z > 4.46\}$. *Because the observed value of* $z = 0.33$ *does not fall in the rejection region, the null hypothesis is not rejected, and we conclude that the assumption of equality of variances seems to be justified.*

TABLE 14.4 Three Independent Samples from Normal Population.

Sample 1	Sample 2	Sample 3
64	56	81
84	74	92
75	69	84
77		
80		

Through our testing, if we find that the homogeneity of variance of the data is violated significantly, then nonparametric tests are more appropriate. Another popular test for equality of variance is Bartlett's test.

14.4.4 Test of independence

Almost all the results in this book assume that we have independent random samples. In the situation where we suspect that the sample data may not be independent, perform a run test as described in Project 12B to test for independence. There are parametric procedures available to test independence; however, the run test is independent of the distributional assumptions and simpler to perform. In general, whether the two samples are independent of each other is decided by the structure of the experiment from which they arise. In the case of correlated samples, such as a set of pre- and posttest observations on the same subject that are not independent, a two-sample paired test may be more appropriate. Another popular method used to check for independence is the chi-square test of independence; see Section 7.6.2. For time series data, the Durbin–Watson test (http://www.alchemygroup.net/Permutation%20Durbin-Watson%20Final.pdf) is effective.

In practical sampling situations, the underlying populations are unlikely to be exactly normally distributed with homogeneity of variances. Both *t*-tests and ANOVA are robust for reasonable departures in some of these assumptions. However, these tests may not be robust with respect to certain other assumption violations. For example, ANOVA is quite sensitive to the violation of independence assumption. These factors need to be given special attention in data analysis.

Exercises 14.4

14.4.1. The scores of 25 randomly selected students from a large calculus class are given below:

47	73	90	22	68	86	94	32	88	86
80	97	48	70	61	82	67	73	78	55
63	59	42	46	90					

(a) Test the data for normality.
(b) If the data are not normal, try a suitable transformation (filter) to make the transformed data normal.

14.4.2. Refer to Example 14.3.1. Suppose we use the transformation $y_i = \ln x_i$ for each observation.
(a) Test whether the transformed data are normal.
(b) Determine whether the data value 18 is still an outlier in the transformed data set.

14.4.3. The data shown in the following table relate to the concealed weapons permits issued in 13 randomly selected Florida counties in 1996:

31,603	20,873	15,963	10,294	8,956	7,901	6,820
5,695	5,485	4,827	3,969	3,278	1,731	

(a) Test whether the data are normal.
(b) If not, try a suitable transformation to make the transformed data normal.

14.4.4. The following table represents a summary by state for Medicare enrollment (in thousands) for 15 randomly selected states in 1998 (source: *Statistical Abstract of the United States,* 1999):

665	3,757	623	757	541	448	478	2,728	103	771
224	86	623	1,373	713					

(a) Test to determine whether the data are normal.
(b) If not, try a suitable transformation to make the transformed data approximately normal.
(c) Test for outliers. If an observation is extreme, would you classify it as an outlier?

14.4.5. Given in the following table are 15 randomly selected state expenditures (in millions of dollars) for the fiscal year 1997 (source: *The World Almanac and Book of Facts*, 2000):

5,722	7,685	13,862	21,975	35,302	4,441	16,200	25,791
4,808	5,130	2,426	39,296	4,002	6,818	7,145	

(a) Test the data for normality.
(b) If the data are not normal, try a suitable transformation to make the transformed data approximately normal.

14.4.6. Using the data of Exercise 14.3.4:
(a) Test whether the data are normal.
(b) If not, try a suitable transformation to make the transformed data approximately normal.

14.4.7. The following data give in-city mileage per gallon for 25 small and midsize cars (source: *Money Magazine*, March 2001):

25 23 20 20 27 26 20 32 25 22
24 21 28 20 22 19 21 29 23 32
23 52 24 24 22

(a) Test to determine whether the data are normal.
(b) If not, try a suitable transformation to make the transformed data approximately normal.
(c) Test for outliers. If an observation is extreme, would you classify it as an outlier?

14.4.8. The following table gives in-state tuition costs (in dollars) for 15 randomly selected colleges taken from a list of the 100 best values in public colleges (source: *Kiplinger's Magazine*, October 2000):

3788 4065 2196 7360 5212 4137 4060 3956 3975 7395
4058 3683 3999 3156 4354

(a) Test for outliers.
(b) Test whether the data are normal.

14.4.9. Using the data given in Exercise 14.2.1, test for equality of variances.

14.4.10. Using the data given in Exercise 14.2.3, test for equality of variances.

14.4.11. The following data represent a random sample of end-of-year bonuses for lower-level managerial personnel employed by a large firm. Bonuses are expressed in percentage of yearly salary:

Females	6.2	9.2	8.0	7.7	8.4	9.1	7.4	6.7
Males	8.9	10.0	9.4	8.8	12.0	9.9	11.7	9.8

Test for equality of variances. State any assumptions you have made, and interpret your result. Use $\alpha = 0.05$.

14.4.12. In an effort to investigate the premium charged by insurance companies for auto insurance, an agency randomly selects a few drivers who are insured by three different companies. These individuals have similar cars, driving records, and levels of coverage. Table 14.5 gives the premiums paid per 6 months by these drivers with these three companies.
Test for equality of variances. State any assumptions you have made, and interpret your result. Use $\alpha = 0.01$.

14.4.13. Three classes in elementary statistics are taught by three different persons, a regular faculty member, a graduate teaching assistant, and an adjunct from outside the university. At the end of the semester, each student is given a standardized test. Five students are randomly picked from each of these classes, and their scores are as shown in Table 14.6.
Test for equality of variances. State any assumptions you have made, and interpret your result. Use $\alpha = 0.05$.

TABLE 14.5 Auto Insurance Premiums.

Company I	Company II	Company III
396	348	378
438	360	330
336	522	294
318		474
		432

TABLE 14.6 Exam Scores by Different Instructors.

Faculty	Teaching assistant	Adjunct
93	88	86
61	90	56
87	76	73
75	82	90
92	58	47

14.5 Modeling issues

A model is a theoretical description in the language of mathematical statistics of a physical phenomenon. Even though interpretations can be developed by analogy, past experience, or intuition, the scientific approach requires a model for the phenomenon of interest. Models are simplifications (or approximations) of real-world situations and are designed to make it easier to identify and understand relationships among variables. A good model is crucial for accurate estimation, forecasting, or predicting. If the observed data show a good fit to the estimates obtained through the model, we consider the model to be an adequate representation of the real-world phenomenon. If not, the model must be improved, to incorporate additional variables or modify the equations defining the relationships. In statistical modeling, it is important not to lose perspective on the essential purpose of the modeling effort. The emphasis should be on making these models work on real data sets in lieu of spending a large amount of time on the capabilities of the models. Even though the study of properties and abilities of models is important, equally important is the ability to know when and how to fit models to a particular data set. A regression line is a two-parameter model that depicts a linear dependence of one variable on another. Again, it is not our objective to discuss all the issues related to statistical modeling. We will discuss briefly only some simple issues relevant to modeling.

14.5.1 A simple model for univariate data

Suppose that we have a data set that characterizes a phenomenon of interest. Suppose our problem is to create a statistical model for the data set in the form of a probability distribution from which the data set came. First we create a dot plot and summary of the basic statistics. The dot plot will provide us with an idea of the probability distribution of the data and any unusual behavior of the data that will not be apparent from the basic statistics such as sample mean and sample standard deviation. Having identified the probability distribution of the sample statistic, we can proceed to obtain 95% confidence limits on parameters such as the mean and variance. In addition, we can obtain a 95% prediction interval of the next observation using the following expression:

$$\overline{y} \pm (t - \text{value}) s \sqrt{1 + \frac{1}{n}}.$$

Note that the prediction interval is always wider than the corresponding confidence interval. The confidence interval provides a measure of reliability for estimating a parameter. The prediction interval provides a measure of reliability for the prediction of an observation. Thus, the prediction interval needs to account for estimation error as well as the natural variability of a single observation. These steps can be considered as the first modeling effort for univariate data. Note that if we have a small sample size, using a t value in the confidence interval and/or prediction interval supposes a modeling assumption of normality for the corresponding population. The preliminary verification of this is done by the dot plot. For more detailed verification of this modeling assumption, use the normal plots.

EXAMPLE 14.5.1

Consider the following data from an experiment:

0.15	0.14	0.15	0.14	0.26	0.00	0.00	0.47	0.35	0.16
0.15	0.15	0.23	0.13	0.19	0.15	0.22	0.53	0.17	0.23
0.22	0.16	0.12	0.13	0.11	0.14	0.18	0.15	0.14	0.21
0.13	0.12	0.13	0.13	0.21	0.22	0.18	0.20	0.22	0.16
0.17	0.00	0.23	0.21	0.18	0.05	0.16	0.13	0.23	0.18
0.14	0.29	0.21	0.22	0.11	0.16	0.23	0.13	0.07	0.17
0.08	0.14	0.06	0.08	0.07	0.11	0.12	0.14	0.16	0.12
0.10	0.27	0.19	0.13	0.27	0.16	0.07	0.09	0.04	0.53
0.29	0.15	0.12	0.11	0.10	0.14	0.14	0.16	0.16	0.17
0.36	0.46	1.21	0.39	0.01	0.52	0.09	0.18	0.16	0.16
0.14	0.15	0.09	0.09	0.13	0.13	0.08	0.14	0.20	0.09
0.09	0.16	0.08	0.10	0.34	0.24	0.15	0.44	0.08	0.08
0.16	0.14	0.18	0.23	0.19	0.11	0.19	0.10	0.14	0.11
0.14	0.17	0.17	0.17	0.05	0.12	0.14	0.11	0.20	0.14
0.23	0.03	0.10	0.29	0.13	0.26	0.13	0.15	0.27	0.14
0.50	0.16	0.15	0.18	0.16	0.14	0.13	0.08	0.20	0.17
0.17	0.16	0.15	0.11	0.13	0.76	0.18	0.19	0.09	0.12
0.11	0.12	0.08	0.26	0.23	0.20	0.19	0.19	0.16	0.11
0.12	0.13	0.32	0.05	0.18	0.12	0.13	0.50	0.13	0.04
0.00	−0.11	0.18	0.15	0.14	0.15	0.02	0.20		

(a) Create a dot plot.
(b) Calculate the basic statistics, sample mean, sample median, and sample standard deviation.
(c) Obtain a 95% confidence interval for the true mean.
(d) Obtain a 95% prediction interval.

Solution
(a) *Each dot in Fig. 14.17 represents three points.*
(b) *We can use Minitab's* ***describe*** *command to obtain the following:*

	N	**Mean**	**Median**	**Tr Mean**	**StDev**	**SE mean**
C1	198	0.17038	0.15121	0.15982	0.13610	0.00967
	Min	Max	Q1	Q3		
	−0.39575	1.22076	0.12059	0.19284		

(c) *Again using Minitab commands, we can obtain (where data are stored in* ***C1****), MTB > ZInterval 95.0 0.136 c1.*

The assumed $\sigma = 0.136$

	N	Mean	StDev	SE mean	95.0% CI
C1	198	0.17038	0.13610	0.00967	(0.15143, 0.18933)

(d) *For the prediction interval use the large sample formula* $\overline{y} \pm (z_{\alpha/2})s\sqrt{1+\frac{1}{n}}$ *to obtain the 95% prediction interval for the true mean as (0.097, 0.4387).*

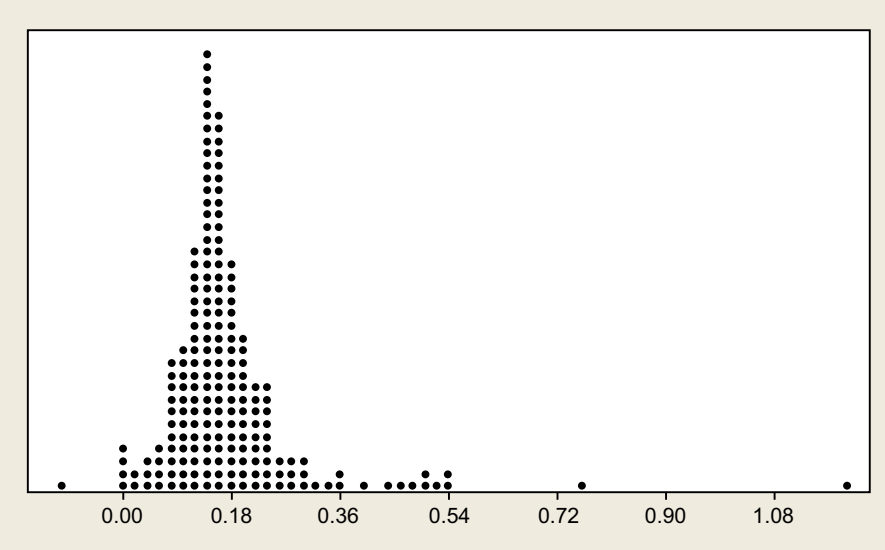

FIGURE 14.17 Dot plot of the data.

14.5.2 Modeling bivariate data

When a scatterplot of bivariate data exhibits a linear pattern, the modeling is usually done using linear regression to study their linear relationship as explained in Chapter 8. Clearly a linear relationship is desirable because it is easy to interpret, departure from linearity is easy to detect, and predicting dependent values from independent variables is straightforward. However, when a scatterplot shows a curved nonlinear pattern, finding a "good" model that fits the observed data may not be very easy. Sometimes, instead of fitting a curve, we may be able to transform the data so as to make the scatterplots of the transformed data look more linear.

A popular statistical method used to straighten a plot is the so-called power transformation. The *power transformation* is defined by specifying an exponent, k, which could be a positive or negative real number, then computing each transformed value as the original value to the power k. Note that $k = 1/2$ gives the square root transform. When $k = 0$, every transformed value is equal to 1. Instead it is customary to think of $k = 0$ as corresponding to a logarithmic transformation so as to unify the transformation concept. The power $k = 1$ corresponds to no transformation at all. Observe that these are the same transformations we have explained in Subsection 14.4.2 to transform nonnormal data into normal transformed data. The shape of the scatterplots should suggest an appropriate transformation. The four curves in Fig. 14.18 represent possible shapes of scatterplots that are usually encountered in practice.

We can use the following as a general guideline for making transformations. If we have a scatterplot that looks like plot 1 of Fig. 14.18, then to straighten the plot, we should use a power $k < 1$ for x (the independent variable) and/or use a power $k > 1$ for y (the dependent variable). Similarly, for curve 2, $k > 1$ for x and/or $k < 1$ (such as $\sqrt{y}$ or $\ln y$) for y. For curve 3, take $k > 1$ for x (such as x^2 or x^3) and/or $k > 1$ for y. Finally, for curve 4, take $k < 1$ for x and/or $k > 1$ for y. Once we straighten the data through transformations, obtain the least-squares equation of the line as explained in Chapter 8. By reversing the transformation (or solving for y in the transformed equation) we can obtain the original nonlinear relationship between x and y.

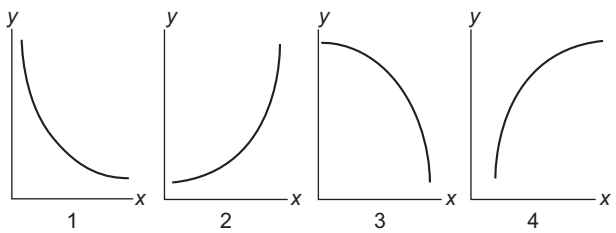

FIGURE 14.18 Possible shapes of a scatterplot.

EXAMPLE 14.5.2

For the following bivariate data:

x	0	4	8	10	15	18	20	25
y	2.4	2.6	3.1	3.6	4.1	4.2	4.6	4.7

(a) Draw a scatterplot.
(b) Use the appropriate transformation (if necessary) to linearize the scatterplot.
(c) Fit the data to an appropriate curve.

Solution

(a) *The scatterplot is shown in Fig. 14.19.*
This looks more like curve 4.

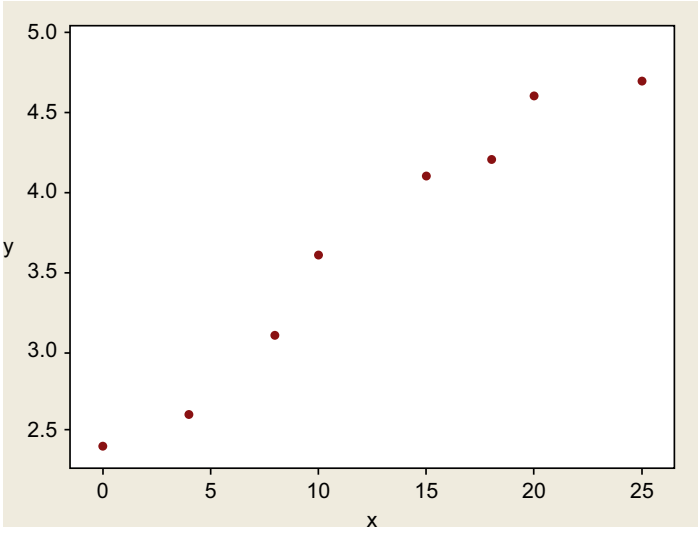

FIGURE 14.19 Scatterplot of the data.

(b) *Let us use the transformation* $x' = \ln x$ *and* $y' = y^2$. *We will get the scatterplot shown in Fig. 14.20.*
This looks more linear.

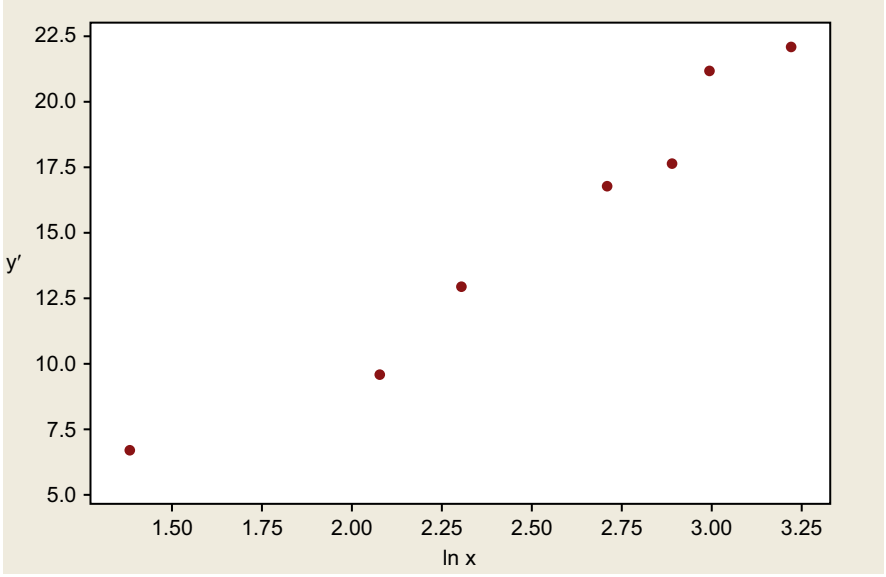

FIGURE 14.20 Scatterplot of the transformed data.

(c) *The regression line for the transformed data is* $y' = 8.86x' - 6.96$. *Therefore, for the original data,* $y^2 = 8.86 \ln x - 6.96$. *The fitted curve is shown in Fig. 14.21.*
Looking at Fig. 14.21, we can see that the data are only slightly nonlinear. In addition, using the equation, for a given value of x *we can predict the value of the response variable* y. *For instance, if* $x = 1.5$, *we estimate* y^2 *to be* -3.3676.

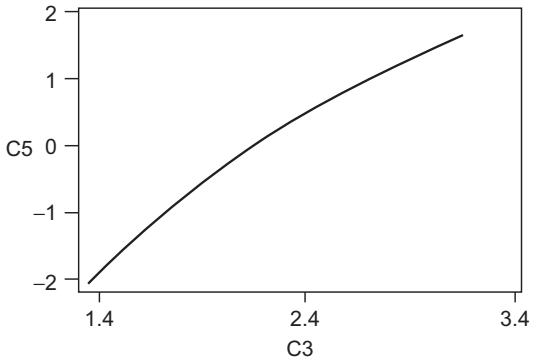

FIGURE 14.21 Fitted curve.

There are various other modeling issues that one may encounter in applications. For example, in multiple regression modeling, an investigator may have data on a number of predictor variables that might be incorporated into a model. Some of these variables may be irrelevant or may duplicate the information provided by other variables. The problem then is how to detect and eliminate the duplicating variables. However, for the sake of brevity and level of presentation, we will not go into the difficulty of these issues of model selection.

Exercises 14.5

14.5.1. Using the data of Exercise 14.4.5:
(a) Create a dot plot.
(b) Describe the data, such as mean, median, and standard deviation.
(c) Obtain a 95% confidence interval for the mean.
(d) Obtain a 95% prediction interval.
(e) Explain your solutions and state any assumptions.

14.5.2. Using the gas mileage data of Exercise 14.4.7:
(a) Create a dot plot.
(b) Describe the data, such as mean, median, and standard deviation.
(c) Obtain a 95% confidence interval for the mean.
(d) Obtain a 95% prediction interval.

14.5.3. The following represents the midterm and final exam scores for 35 randomly selected students from a large mathematics class:

Midterm	67	63	39	80	64	95	90	93	21	36
	44	66	66	72	34	78	66	68	98	43
	74	81	71	100	60	50	81	66	90	89
	86	49	77	63	58					
Final	29	33	100	33	55	20	10	5	67	64
	71	25	34	66	28	34	16	27	32	20
	14	21	16	62	50	14	61	11	14	41
	52	35	37	51	43					

(a) Draw a scatterplot.
(b) Use an appropriate transformation (if necessary) to linearize the scatterplot.
(c) Fit the data to an appropriate curve and explain the usefulness.

TABLE 14.7 Tution Amount Versus Graduatin Rate.

In-state tuition	3788	4065	2196	7360	5212	4137	4060	4354
Graduation rate	45	64	40	58	38	20	39	48
In-state tuition	3956	3975	7395	4058	3683	3999	3156	
Graduation rate	40	20	45	39	39	20	9	48

14.5.4. Using the state finance data of Exercise 14.2.3:
- **(a)** Draw a scatterplot.
- **(b)** Fit a least-squares line.
- **(c)** Explain your solutions and state any assumptions.

14.5.5. Table 14.7 gives in-state tuition costs (in dollars) and 4-year graduation rate (%) for 15 randomly selected colleges taken from a list of the 100 best values in public colleges (source: *Kiplinger's Magazine*, October 2000).
- **(a)** Draw a scatterplot.
- **(b)** Fit a least-squares line and graph it.
- **(c)** Looking at the scatterplot of (a), do you think the least-squares line is a good choice? Discuss.

14.6 Parametric versus nonparametric analysis

Up until Chapter 11, we basically assumed that random variables belong to specific probability distributions, such as a normal distribution or binomial distribution. The members of those distributions are associated by different parameters such as means or variances. Most of our efforts were concentrated on making some inferences about the unknown parameters. In this vein, we looked at point estimators, confidence intervals, and hypothesis-testing problems. In practice, the assumption that observations come from a particular family of distributions such as normal or exponential may be quite sensible. As we have already mentioned, slight violations of these assumptions in many practical cases may not significantly affect statistical inferences. However, this is not always true. Furthermore, sometimes we may want to make inferences that have nothing to do with parameters. We may not even have precise measurement data, but only the rank order of observations. For example, if we want to study the performance of students at an institution, we may not have the precise scores the students obtained; instead we may only have their letter grades such as A, B, C, D, and F. Even if we have precise measurements, we may not be able to assume a distribution, such as normality. Still, we may be able to say that the distribution is symmetric, or skewed, or has some other characteristics. Basically, if there is doubt about the parametric assumptions, or the data are not suitable for parametric inference, or we are not interested in inference about parameters, a nonparametric test that is valid under weaker assumptions is preferable. It should be noted that weaker assumptions do not mean that nonparametric methods are assumption free. The inference that can be made depends on valid assumptions that are made.

When using nonparametric tests, a common question is, "Why substitute a set of nonnormal numbers, such as ranks, for the original data?" Rank tests are often useful in circumstances when we have no idea about the population distribution. We suspect that the data are not normal, and either we cannot transform the data to make them more normal or we do not wish to do so. Few data are truly normal, despite the robustness of common parametric tests; unless we are quite sure that the nonnormality is a minor problem and would not affect the conclusions, we may often be better off using a rank test. However, there is a small penalty for using delete rank tests. If the original data are really normal, in the long run, the rank tests will be about 95.5% as efficient as a Student t-test would have been. This means that in such situations, the t-test will require about 95 samples compared with 100 for the rank test. But when data are far from normal, the rank tests will require fewer samples than the t-test; in fact, we should not use the t-test in such cases.

Basically, if we know the distribution of the underlying population, we can use parametric tests. Otherwise, for a given data set, we first perform the normality test as explained in Section 14.4. If normality fails, try transformations; if that fails, we can use nonparametric methods for the data analysis.

Another situation in which we can use nonparametric tests is when the data contain some outliers. A box plot or a normal plot, as explained in Section 14.4, will reveal the existence of outliers. However, in many applied areas, such as in most bioavailability data, there will appear to be outliers. It is not feasible to determine whether these are skewed or contaminated distributions. They are not errors. In those situations, a conservative approach will be to use nonparametric

methods. For example, because the statistic for the rank sum test is resistant to outliers, it will not be seriously affected by the presence of outliers unless the number of outliers becomes large relative to the sample size.

It should be noted that we ought to be careful even when we use nonparametric tests. For example, if the data for one or both of the samples to be analyzed by a rank sum test come from a population whose distribution violates the assumption that the distributional shapes are the same, then the rank sum test on the original data may provide misleading results or may not be the most powerful test available. Transforming the data (for example, a logarithmic transformation pulls in long tails) to obtain normality and then performing a two-sample t-test, or using another nonparametric test, may be more appropriate for the analysis. In general, nonparametric methods are appropriate when the sample sizes are small. When the data set is large, say $n > 100$, it often makes little sense to use nonparametric methods.

Finally, we must conclude that we do not perform nonparametric tests on a given set of data unless it is necessary, that is, if we cannot assume a classical probability distribution that characterizes the given data. Also, parametric statistical analysis is, in general, more powerful than nonparametric analysis. We will end this section with a quote from W.J. Conover: "Nonparametric methods use approximate solutions to exact problems, while parametric methods use exact solutions to approximate problems."

Exercises 14.6

14.6.1. Consider the following data:

0.01	0.012	0.016	0.018	0.036	0.042	0.036	0.048
0.072	0.042	0.22	0.096	0.76	0.055	0.13	0.016

(a) Test for normality and comment on whether a parametric or a nonparametric test is appropriate.
(b) Try a suitable transformation (filter) to make the transformed data normal, if possible, and then use a parametric procedure.

14.6.2. Using the Medicare data in Exercise 14.4.4, if parametric procedures are not appropriate, use a nonparametric procedure.

14.7 Tying it all together

Now we will give some real-world problems for which we will use standard methods to analyze the given data. Software reliability is a major aspect in any kind of software development. One of the ways to do this is to observe time to failure and/or time between failures (TBF). If the defects are fixed, we would expect, on average, the TBF to increase. Based on those data, one studies reliability of the software. There are a variety of methods to analyze the software reliability problems. Here we will not dwell on the reliability issues. We will only do some simple data analysis on a set of software failure data. The following data represent software failure times in the Apollo 8 software system. They were obtained from www.dacs.dtic.mil/databases/sled/swrel.shtml. It is assumed that these failure times are random.

EXAMPLE 14.7.1

The following data set consists of 26 software failure times taken from testing of the Apollo 8 software system:

T :	9	21	32	36	43	45	50	58	63
	70	71	77	78	87	91	92	95	98
	104	105	116	149	156	247	249	250	
TBF :	9	12	11	4	7	2	5	8	5
		7	1	6	1	9	4	1	3	3
		6	1	11	33	7	91	2	1	

(a) Create a dot plot and describe the TBF data.
(b) Identify any outliers and test for normality with and without outliers for TBF data. If the data are not normal, does any simple transformation make the data normal?
(c) Obtain a 95% confidence interval for TBF.

(d) For estimation problems, does a parametric or nonparametric method seem more appropriate for this data?
(e) Create a scatterplot between T and TBF and discuss its usefulness.

Solution

(a) *The dot plot for the TBF data is shown in Fig. 14.22.*
The following is the result from using the describe command in Minitab:

TBF	*N*	Mean	Median	Tr mean	StDev	SE mean
	26	9.62	5.50	6.58	17.79	3.49
TBF	Min	Max	Q1	Q3		
	1.00	91.00	2.00	9.00		

(b) *We will use the box plot shown in Fig. 14.23 to identify the outliers.*
From the box plot, observations 33 and 91 are outliers.
Figs. 14.24 and 14.25 show the normal plots with and without outliers.
It is clear that the data with outliers are not normal, whereas if we remove the outliers, the data become normal.
Fig. 14.26 gives the normal plot by taking the natural log of the TBF data with outliers. The figure shows that the data become approximately normal.

(c) *It is clear that to obtain a small-sample confidence interval, to satisfy the assumption of normality, we need to take the data without the outliers. Hence, a 95% confidence interval for TBF with the outliers removed is (3.77, 6.73). Running a nonparametric Wilcoxon test in Minitab for the 95% confidence interval with outliers gave the following:*

	Estimated		**Achieved**	
Time between failures	*N*	Median	Confidence	Confidence interval
	26	6.00	94.9	(4.00, 8.00)

(d) *If we are analyzing the data without outliers or the log-transformed data, parametric methods are better. With the original data, because the normality assumption may not be appropriate, we need to use nonparametric methods.*
(e) *Fig. 14.27 gives the scatterplot of T and TBF.*

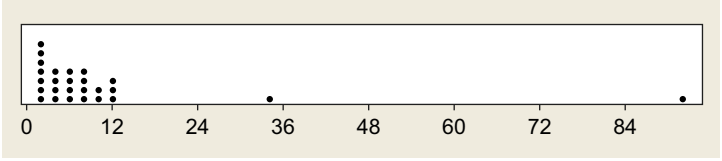

FIGURE 14.22 Dot plot of time-between-failures data.

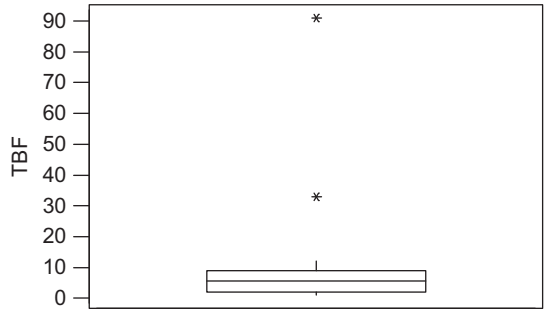

FIGURE 14.23 Box plot of time-between-failures (*TBF*) data.

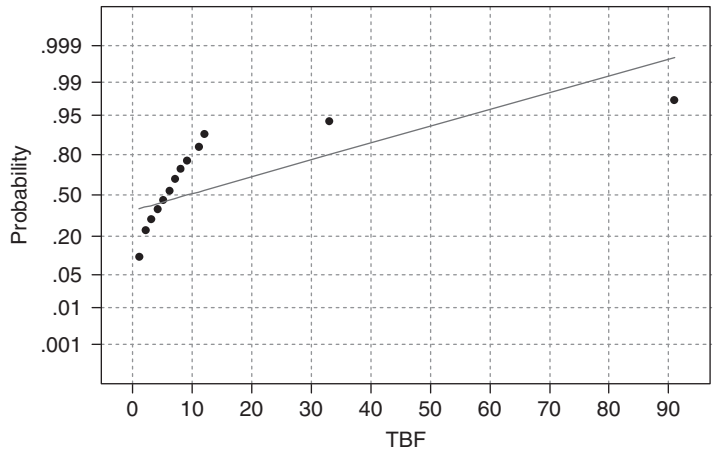

FIGURE 14.24 Normal probability plot of time-between-failures (*TBF*) data with outliers.

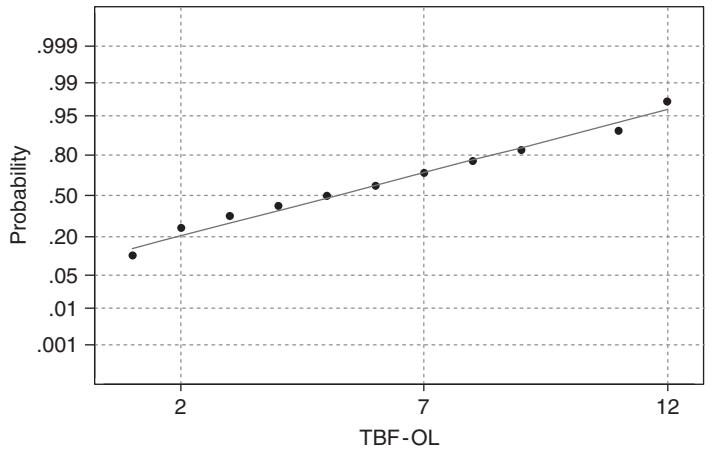

FIGURE 14.25 Normal probability plot of time-between-failures (*TBF*) data without outliers.

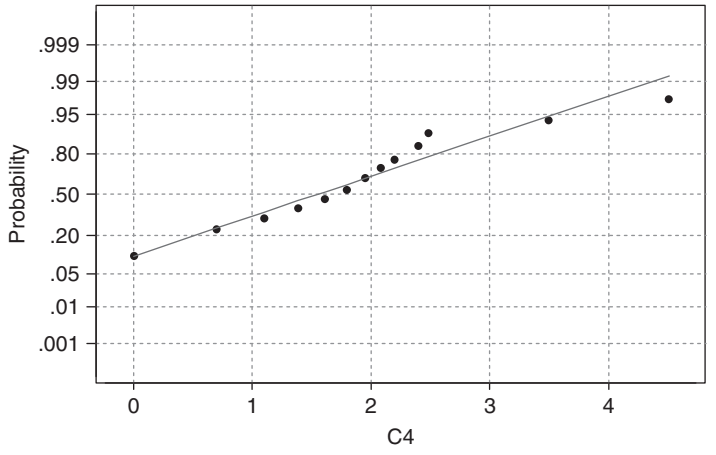

FIGURE 14.26 Normal probability plot of transformed time-between-failures (*TBF*) data with outliers.

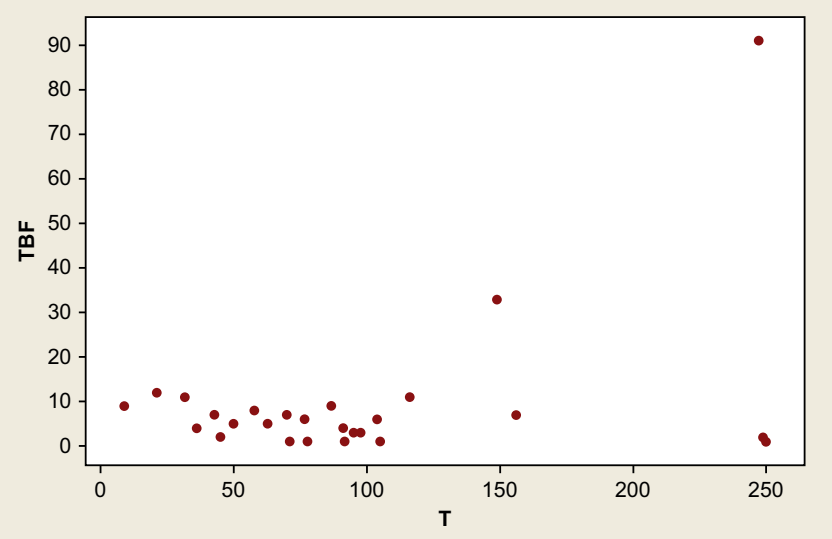

FIGURE 14.27 Scatterplot of time (T) and time between failures (TBF).

EXAMPLE 14.7.2

Table 14.8 gives dealer cost and sticker price for four-door base models of 25 small and midsize cars (source: *Money Magazine*, March 2001).

(a) Create a dot plot and describe the sticker price data.

(b) Identify any outliers and test for normality with and without outliers for sticker price data. If the data are not normal, does any simple transformation make the data normal?

(c) Obtain a 95% confidence interval for sticker price.

TABLE 14.8 Dealer Cost and Sticker Price.

Model	Dealer cost (in dollars)	Sticker price (in dollars)
Acura Integra GS	19,479	21,600
Chevy Cavalier	12,398	13,260
Chevy Impala LS	21,251	23,225
Chrysler Concord LX	20,834	22,510
Dodge Neon SE	11,856	12,715
Ford Escort	12,277	12,970
Ford Taurus SE	17,606	19,035
Honda Civic DX	11,723	12,960
Honda Accord 2.3 LX	16,727	18,790
Hyundai Sonata	13,805	14,999
Kia Sephia	9,914	10,595
Mazda 626 LX V6	18,181	19,935
Mitsubishi Mirage ES	12,534	13,627
Mercury Sable GS	17,777	19,185

Continued

TABLE 14.8 Dealer Cost and Sticker Price.—cont'd

Model	Dealer cost (in dollars)	Sticker price (in dollars)
Nissan Maxima GXE	19,430	21,249
Oldsmobile Intrigue GL	22,097	24,150
Pontiac Grand Am GT	18,790	20,535
Saturn SL	9,936	10,570
Subaru Impreza L	14,695	15,995
Toyota Corolla LE	12,042	13,383
Toyota Camry LE	18,169	20,415
Toyota Prius	18,793	19,995
VW Jetta GLS	15,347	16,500
VW Passat GLS	19,519	21,450
Volvo S40	22,090	23,500

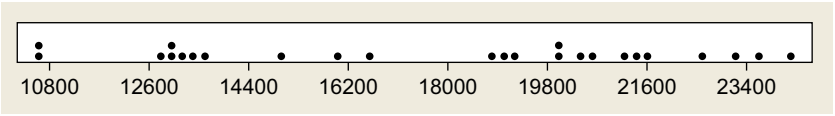

FIGURE 14.28 Dot plot for the sticker price.

(d) For estimation problems, do parametric or nonparametric methods seem more appropriate for this data?
(e) Create a scatterplot between dealer cost and sticker price.
(f) Fit a least-squares regression line and run a residual model diagnostic using Minitab.

Solution

(a) *The dot plot for the sticker price is shown in Fig. 14.28.*
The following summary statistics are obtained by using the describe command in Minitab.

	N	Mean	Median	Tr mean	StDev	SE Mean
Sticker price	25	17,726	19,035	17,758	4278	856
	Min	Max	Q1	Q3		
Sticker price	10,570	24,150	13,322	21,350		

(b) *The box plot for the sticker price is shown in Fig. 14.29.*
According to this, there are no outliers. The normal plot is shown in Fig. 14.30. This is approximately normal.

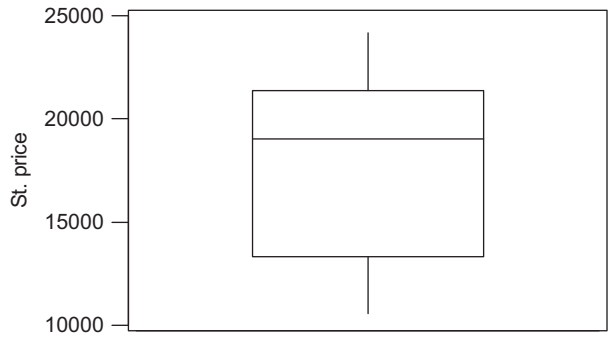

FIGURE 14.29 Box plot for the sticker (*St.*) price.

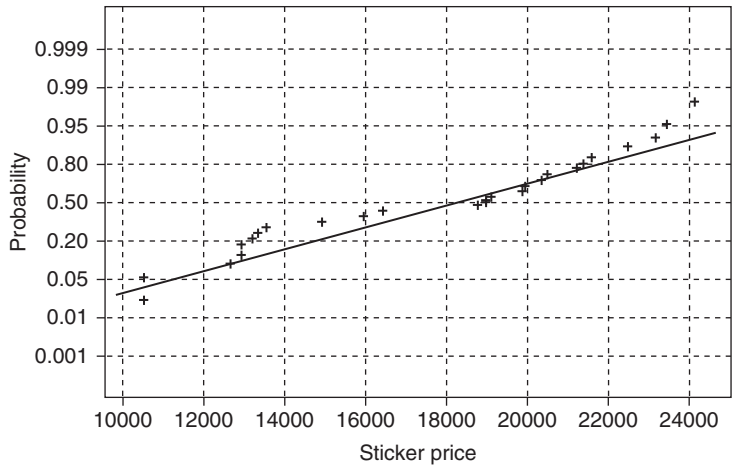

FIGURE 14.30 Normal plot for the sticker price.

(c) *The 95% confidence interval for the sticker price is:*

	N	Mean	StDev	SE mean	95.0% CI
Sticker price	25	17,726	4278	856	(15,960, 19,492)

(d) *Because there are no outliers and the data look approximately normal, parametric tests seems to be appropriate for these data.*
(e) *The scatterplot for dealer cost versus sticker price is shown in Fig. 14.31.*

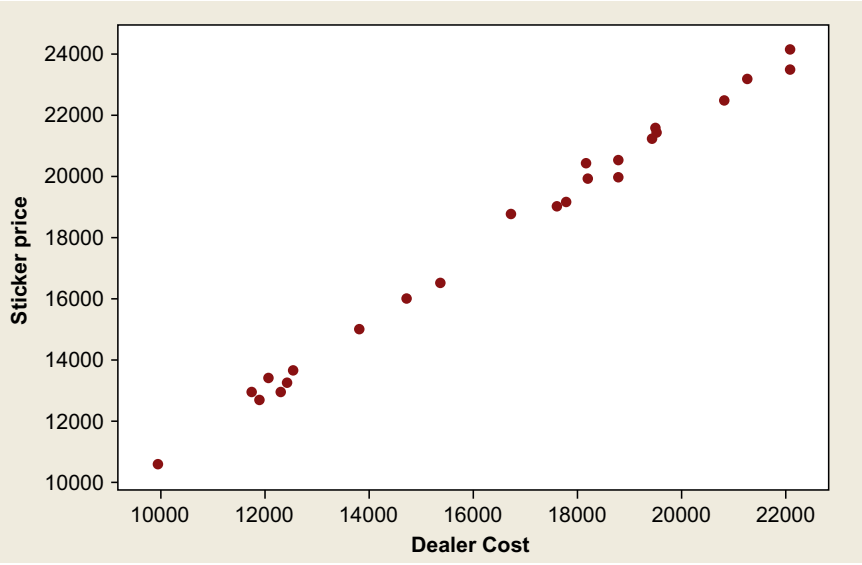

FIGURE 14.31 Scatterplot for dealer cost versus sticker price.

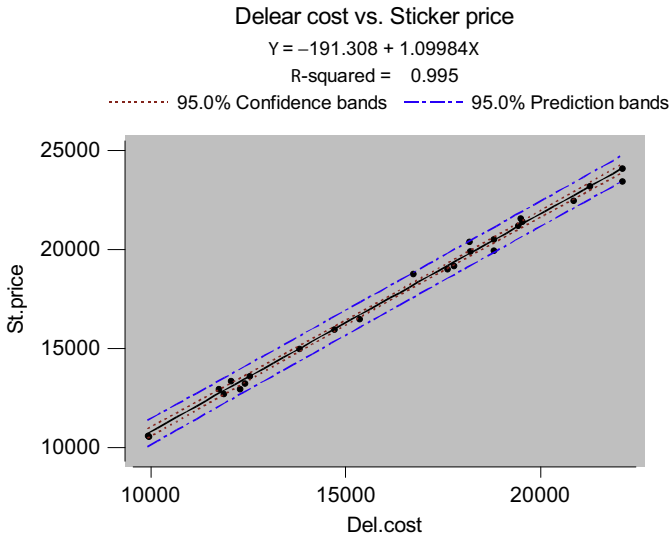

FIGURE 14.32 Regression line for dealer cost versus sticker price.

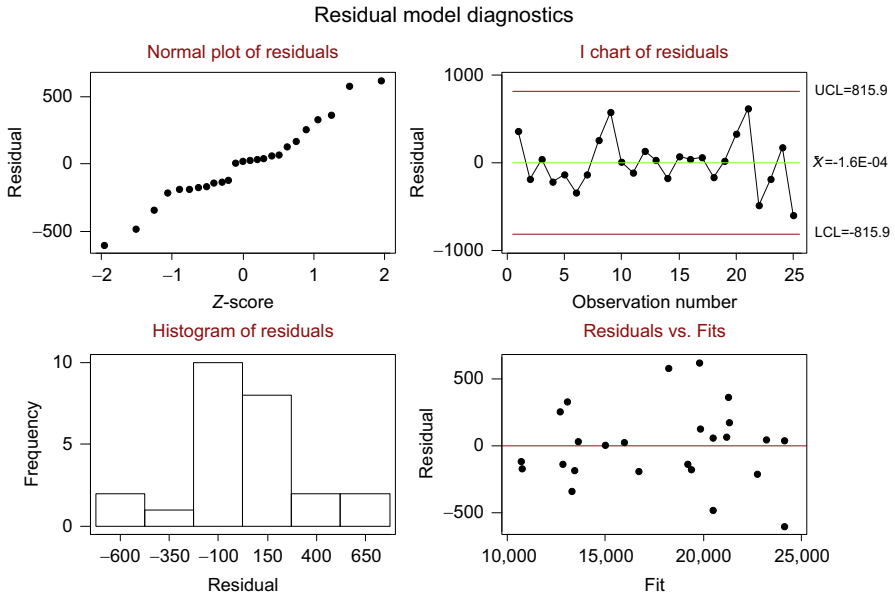

FIGURE 14.33 Residuals versus fit.

(f) *Fig. 14.32 shows the fitted regression line.*
An analysis of residuals by Minitab gives Fig. 14.33.
By looking at the residuals versus fits, we can see that we have a good fit, and hence, the model seems to be appropriate.

Exercises 14.7

14.7.1. Table 14.9 gives revenue (in thousands) for public elementary and secondary schools, by state, for 1997–98 and corresponding pupils per teacher for that state for 20 randomly selected states (source: *The World Almanac and Book of Facts*, 2000).

(a) Create a dot plot and describe the pupils per teacher data.

TABLE 14.9 School Revenue and Number of Pupils per Teacher.

State	Total revenue	Pupils per teacher
Arizona	4,388,915	19.8
Connecticut	5,112,950	14.2
Alabama	4,030,356	16.3
Indiana	7,006,752	17.2
Kansas	3,090,829	14.9
Oregon	3,119,028	20.1
Nebraska	1,688,662	14.5
New York	27,690,556	15.0
Virginia	6,661,612	14.7
Washington	6,722,916	20.2
Illinois	13,649,628	16.8
North Carolina	7,127,549	15.9
Georgia	8,579,628	16.2
Nevada	1,754,717	18.5
Ohio	12,694,407	16.7
New Hampshire	1,365,391	15.6

(b) Identify any outliers and test for normality with and without outliers for the pupils per teacher data. If the data are not normal, does any simple transformation make the data normal?
(c) Obtain a 95% confidence interval for pupils per teacher.
(d) Create a scatterplot for total revenue and pupils per teacher.
(e) Fit a regression line between total revenue and pupils per teacher.

14.7.2. Table 14.10 gives the dealer cost and sticker price for luxury cars and sports utility vehicles with popular options (source: *Money Magazine*, March 2001).
(a) Create a dot plot and describe the sticker price data.
(b) Identify any outliers and test for normality with and without outliers for sticker price data. If the data are not normal, does any simple transformation make the data normal?
(c) Obtain a 95% confidence interval for sticker price.
(d) Do parametric or nonparametric methods seem more appropriate for the data?
(e) Create a scatterplot between dealer cost and sticker price.
(f) Fit a least-squares regression line and run a residual model diagnostics using Minitab.

14.7.3. For the college tuition data of Exercise 14.5.5, fit a least-squares regression line and run a residual model diagnostics using Minitab.

14.7.4. The following data give the area (in square feet) and the sale prices (approximated to the nearest $1000) of homes that were sold in a particular city in a 6-week period of 2003.

Area	1123	1028	1490	2172	2300	1992	3200	3063	3720
	7228	720	943	904	912	1031	1152	1482	1426
	1491	1184	1650	1392	1755	2062	2495	3253	5152
	1270	1723	1161	1220	837	1446	2442	2300	2518
Price	75	75	102	149	152	154	327	425	625
	775	775	57	66	68	75	86	90	93
	95	95	104	105	135	159	169	253	725
	67	67	110	65	74	95	156	183	207

TABLE 14.10 Dealer Cost and Sticker Price for Luxury and Sport Cars.

Model	Dealer cost (in dollars)	Sticker price (in dollars)
Acura TL 3.2	26,218	29,030
Audi A6 4.2	45,385	50,754
BMW 525i	33,800	37,245
Cadillac DeVille DHS4	43,825	47,603
Infiniti I30 Touring	28,604	32,065
Jaguar XJ8	52,535	58,171
Lexus GS430	41,881	48,581
Mercedes-Benz C320	35,067	36,950
SAAB 9-3 Viggen	35,270	38,690
Volvo S80T-6	39,315	41,768
BMW X5 4.4i	45,994	50,774
Chevrolet Blazer LT	26,958	29,725
Dodge Durango	26,845	29,370
GMC Jimmy SLE	26,637	29,370
Honda CR-V LX	17,578	19,190
Isuzu Trooper LS	27,901	31,285
Jeep Cherokee SE	21,392	23,130
Lexus LX470	54,785	63,474
Mercedes-Benz ML430	42,243	45,337
Nissan Pathfinder SE	27,203	29,869
Pontiac Aztek GT	22,912	24,995
Subaru Forester S	21,990	24,190
Suzuki Vitara JS	16,063	17,079
Toyota RAV4	18,786	20,630

(a) Create a dot plot and describe the home price data.
(b) Identify any outliers and test for normality with and without outliers for home price data. If the data are not normal, does any simple transformation make the data normal?
(c) Obtain a 95% confidence interval for home price.
(d) Do parametric or nonparametric methods seem more appropriate for the data?
(e) Create a scatterplot between the square-foot area of a home and its price.
(f) Fit a least-squares regression line and run a residual model diagnostics using Minitab.

14.8 Some real-world problems: applications

In this section, we will use the goodness-of-fit methods discussed in the previous sections to identify the probability distribution that characterizes the behavior of some real-world problems that our society is facing. All the data sets used in this section are available at http://booksite.elsevier.com/9780124171138.

14.8.1 Global warming

The concept of "global warming" consists of two interacting entities, the atmospheric temperature and carbon dioxide, CO_2, in the atmosphere. The United States collects annual data for both of these variables in our observatories in Alaska and Hawaii. The actual data can be found on the website http://scrippsco2.ucsd.edu/data/atmosperic-co2.html.

Our objective is to identify the probability distribution function (pdf) that follows the CO_2 data given in thousands of metric tons annually for 31 years. Once we know the pdf that fits the CO_2 data, we can obtain useful information, such as probabilistic characterization of its behavior, the expected value of CO_2 (theoretical average), and confidence limits on the true amount of CO_2, among other interesting information.

We begin our process of identifying the pdf by structuring a histogram of the 31 randomly selected measurements of CO_2. The histogram of the subject data will give us some idea about the possible pdf that we should be testing. After some preliminary testing of some pdfs we proceed to test the following hypothesis:

H_0: the CO_2 data follow the gamma pdf

versus

H_a: the CO_2 data do not follow the gamma pdf.

To test this hypothesis, we applied the Kolmogorov−Smirnov, Anderson−Darling, and chi-square tests with a level of significance $\alpha = 0.05$. The test statistic results of the three goodness-of-fit tests are given below:

Kolmogorov−Smirnov test	$D = 0.08771$, p value 0.954
Anderson−Darling test	$A = 0.3627$, p value 0.883
Chi-square test	$\chi^2 = 0.95844$, p value 0.811

All three goodness-of-fit tests strongly support the null hypothesis that the CO_2 measurements follow the gamma pdf. We obtained the maximum likelihood estimates of the two parameters α and β of the gamma pdf, which are $\widehat{\alpha} = 635.29$ and $\widehat{\beta} = 0.557$. Thus, we can write the estimated gamma pdf for the subject data. That is,

$$f(x) = \frac{x^{635.29-1}}{(0.557)^{635.29}\Gamma(635.29)}\exp\left(\frac{-x}{0.557}\right), \quad x > 0.$$

We can use $f(x)$ to determine various probabilities of interest concerning the behavior of X, the amount of CO_2 in the atmosphere. Also, we can calculate cumulative distribution function $F(x)$, the expected amount that we would find in the atmosphere, and confidence limits, among other interesting questions about the behavior of CO_2, using procedures previously explained in this book.

14.8.2 Hurricane Katrina

One of the most devastating hurricanes in the past 100 years to hit the United States was Hurricane Katrina. The Atlantic-based hurricane, category 5 (most devastating), lasted 9 days, August 23−31, 2005. The wind pressure and velocity of Katrina are two of the most important variables and we wish to identify the pdf that characterizes its behavior. That is, we wish to perform goodness-of-fit testing to determine the pdf that follows the wind pressure data that were obtained from http://weather.unisys.com/hurricane/atlantic/2005H/KATRINA/track.dat.

We have 63 observations of the wind velocity (in mph) that reached a maximum wind velocity of 150 mph. After looking at the histogram of the data, we believe that the wind velocity of Katrina followed the two-parameter ($\delta = 0$) Weibull pdf. Thus, we proceeded to test the following hypothesis:

H_0 : the wind velocity data of Hurricane Katrina follow the two-parameter Weibull pdf

versus

H_a : the wind velocity data of Hurricane Katrina do not follow the Weibull pdf.

To test this hypothesis, we applied the Kolmogorov−Smirnov, Anderson−Darling, and chi-square tests.

All these tests strongly support the acceptance of the null hypothesis. The test results are given below:

Kolmogorov−Smirnov test	$D = 0.0792$, p value 0.795
Anderson−Darling test	$A = 0.5949$, p value 0.863
Chi-square test	$\chi^2 = 3.4031$, p value $= 0.638$

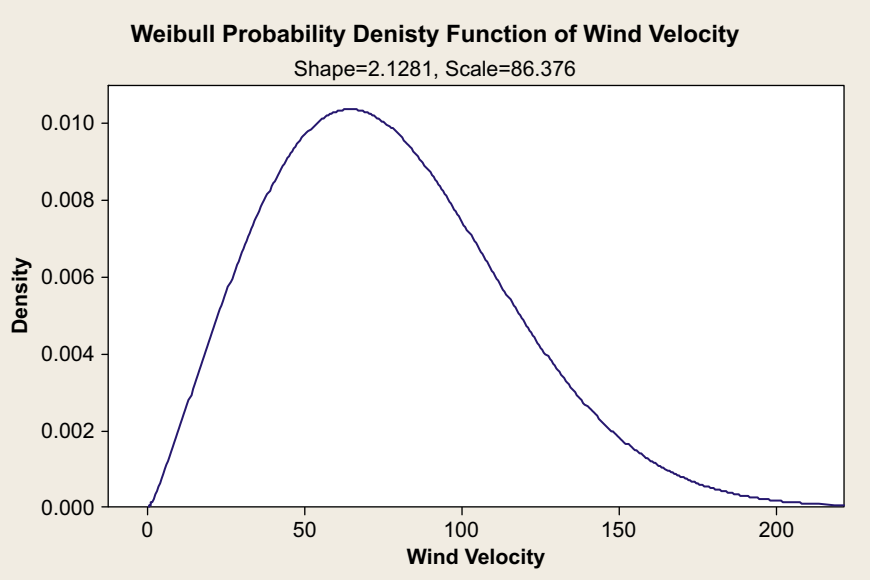

FIGURE 14.34 Weibull probability density function of wind velocity of Hurricane Katrina.

Thus, the wind velocity measurements of Hurricane Katrina follow the two-parameter Weibull pdf, with the maximum likelihood estimates of the parameter given by $\widehat{\alpha} = 2.1281$ and $\widehat{\beta} = 86.376$. The pdf of the subject data is given by:

$$f(x) = \begin{cases} \dfrac{86.376}{2.1281}\left(\dfrac{x}{2.1281}\right)^{85.376} \exp\left(\dfrac{x}{2.1281}\right)^{86.376}, & x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

A graphical display of $f(x)$ is given in Fig. 14.34.

Knowing the pdf that characterize the true-probabilistic behavior of the wind velocity of Katrina, we can calculate the expected wind velocity and confidence limits. That is,

$$E(X) = 76 \text{ miles/hour},$$

and the 95% confidence limits of the true mean of the wind velocity are 23.8 and 150.6 mph.

That is, we are at least 95% certain that the true wind velocity of Hurricane Katrina or similar hurricanes will be between 23.8 and 150.6 mph.

Also, the cumulative probability distribution, $F(x)$, of the wind velocity in its analytical and graphical form (Fig. 14.35) is given below:

$$F(x) = P(X \le x) = \int_0^x \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp - \left(\frac{t}{\alpha}\right)^{\beta} dt, \quad x > 0, \ \alpha, \ \beta > 0,$$

and for the given data we will get:

$$F(x) = \int_0^x \frac{86.37}{2.128}\left(\frac{t}{2.128}\right)^{85.37} \exp\left(- \left(\frac{t}{2.128}\right)^{86.37}\right) dt.$$

Thus, we can use the graph in Fig. 14.35 to obtain various probabilities; for example, if we are interested in the probability that the wind velocity of a category 5 hurricane is less than 150 mph we can obtain an approximate estimate from this graph, that is,

$$F(150) = P(X \le 150) \approx 0.93.$$

This means that based on the given data we are approximately 93% certain that the wind velocity will be less than 150 mph.

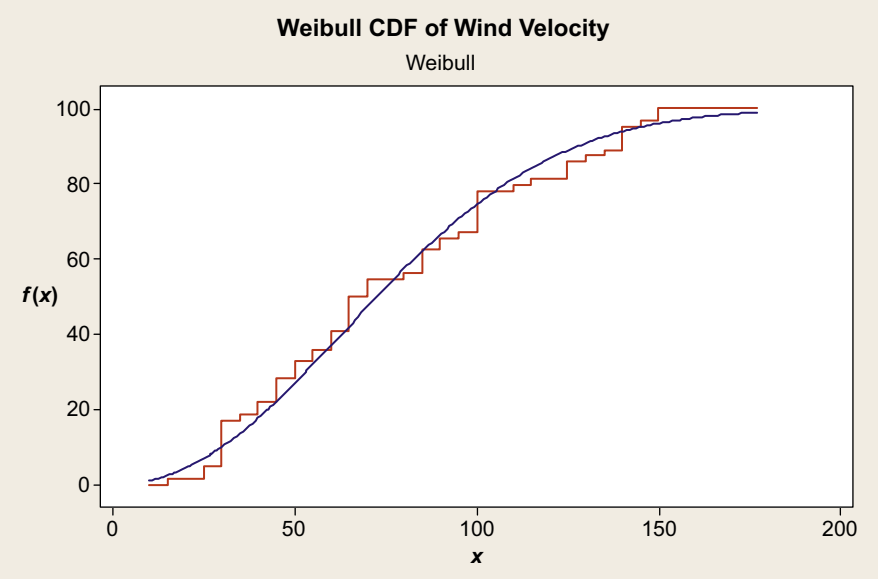

FIGURE 14.35 Weibull cumulative distribution function of the wind velocity of Hurricane Katrina.

The following is the R-code for the goodness-to-fit tests for the Katarina data:

```
HK<-read.delin("~/Documents/Hurricane Katrina.txt")
View(HK)
summary(HK) #### descriptive Stat###
xk=HK$WIND
hist(xk) ### Histogram ####
library(lessR)
dens(xk,type=c("both","normal"),xlab="Wind",ylab="f(x)")
color.density(xk)
m=mean(xk);m
std=sqrt(var(xk));std
hist(xk),density=12,breaks=8,prob=T,col="plum4",xlab="Wind",xlim=c(0,200),main=
"Histogram of Wind velocity of Hurricane Katrina")
library(vcd) ## Goodness of fit test
fitdistr(xk,'weibull') ### estimate the parameters using MLE
ks.test(xk,"pweibull",shape=1.805,scale=52.323) #Kolmogorov-Smirnov test
ad.test(xk)#Anderson-Darling
```

14.8.3 National unemployment

The aim in the present problem is to identify the probability distribution that characterizes the rates of unemployment in the United States. The subject data were obtained from the US Bureau of Labor Statistics, www.bls.gov/, under Database & Tools. The data are the annual averages of the unemployment rate in the United States from 1957 to 2008. Initially, we looked at the histogram of the data and it gave us a visual interpretation that it may follow the gamma pdf. Initially we tested for the two-parameter gamma pdf and obtained a fairly good fit, but when we tried the three-parameter gamma pdf, we obtained a better fit. That is:

H_0 : the annual average rates of unemployment in the United States follow the three-parameter gamma pdf

versus

H_a : the subject data do not fit the three-parameter gamma pdf.

Given below is the value of the goodness-of-fit test statistics for a sample of 51 data points:

Kolmogorov–Smirnov test	$D = 0.0847$, p value 0.8276
Anderson–Darling test	$A = 0.3424$, p value 0.7916
Chi-square test	$\chi^2 = 2.2353$, p value 0.8172

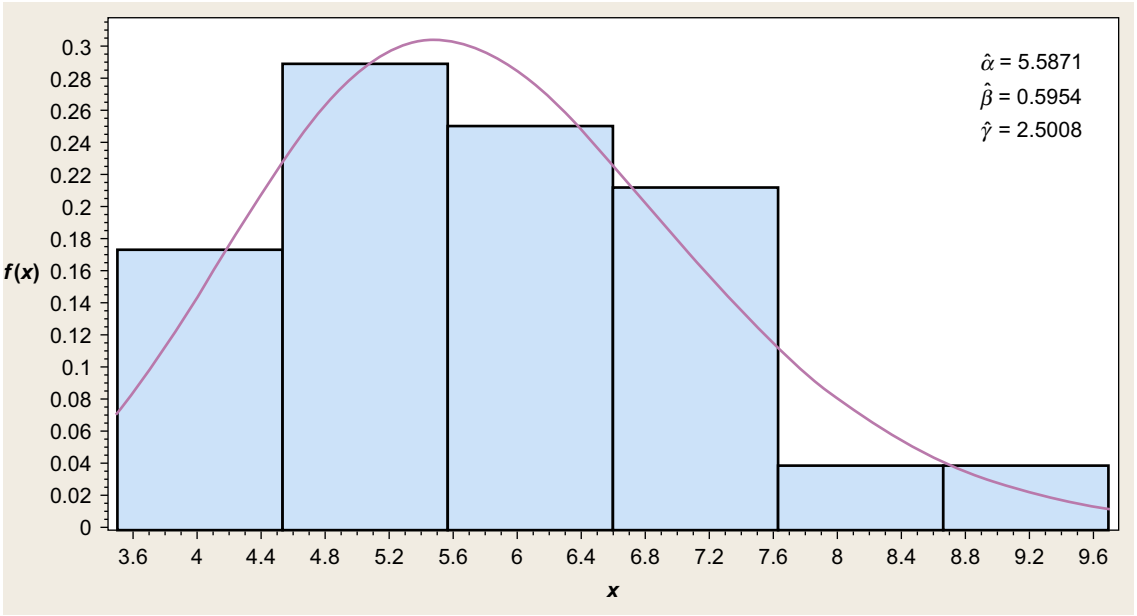

FIGURE 14.36 Three-parameter gamma probability density function for unemployment in the United States.

All three goodness-of-fit tests strongly support that the three-parameter gamma pdf is probabilistically the best to characterize the behavior of the US annual average of unemployment, with the maximum likelihood estimate of the parameters $\widehat{\alpha} = 5.5871$, $\widehat{\beta} = 0.5954$, and $\widehat{\gamma} = 2.5008$. Thus, the subject pdf is given by:

$$f(x) = \frac{1}{(0.5954)^{5.5971}\Gamma(0.5954)}(x - 2.5008)^{4.5871} \exp - \left(\frac{x - 2.5008}{0.5954}\right), \quad x > 0.$$

The expected value of the subject pdf is:

$$E(X) = \widehat{\delta} + \widehat{\alpha}\left(\widehat{\beta}\right) = 2.5008 + (5.5871)(0.5954) = 5.83.$$

Thus, one will expect the unemployment rate to be approximately 5.83% based on the actual data we analyzed. A graphical form of $f(x)$ over the initial histogram that guided us to the three-parameter gamma pdf is given in Fig. 14.36.

We can use the pdf to obtain confidence limits on the true rate of unemployment, the cumulative pdf, $F(x)$, and various probabilities of interest on the subject problem, among other useful information.

14.8.4 Brain cancer

A brain tumor is an abnormal growth of cells within the brain, which can be cancerous (malignant) or benign. It is estimated that there have been more than 43,800 new cases of cancerous brain tumors in the United States during the past few years. In this application we are interested in studying the behavior of the malignant tumor sizes in the brain. The subject data were obtained from the Surveillance Epidemiology and End Results (SEER) database. We have taken a random sample of 200 brain cancer patients from the large database with their cancerous tumor size measured in millimeters. Our aim is to find the probability distribution that characterizes the behavior of the tumor size. Thus, after testing several pdfs and looking at the histogram we believe that the three-parameter Weibull pdf is a prime candidate. Table 14.11 contains the actual data for 50 patients.

Now, we proceed to test our belief.

H_0 : the sizes of the malignant tumors in the brain fit the three-parameter Weibull pdf

versus

H_a : the subject data do not follow the three-parameter Weibull pdf.

We are applying the most commonly used goodness-of-fit tests to make a decision concerning accepting or rejecting the stated hypothesis for, say, $\alpha = 0.01, 0.05, 0.10$. The results of the three tests are given below:

Kolmogorov–Smirnov test	$D = 0.0502$, p value 0.6746
Anderson–Darling test	$A = 0.6948$, p value 0.7321
Chi-square test	$\chi^2 = 9.6143$, p value 0.2115

TABLE 14.11 Brain Cancer Data.

Frequency	1	3	3	1	1	2	3	2	4	2	1	7	1	1	2
Tumor size (mm)	7	8	10	11	12	14	15	19	20	23	24	25	26	27	28
Frequency	1	7	1	34	1	7	1	1	27	3	1	1	2	11	2
Tumor size (mm)	34	35	37	40	41	45	46	48	50	55	56	57	59	60	63
Frequency	3	2	9	1	1	1	2	6	1	1	1	2	1	1	2
Tumor size (mm)	65	67	70	72	73	74	75	80	83	85	86	90	94	100	120
Frequency	1	1	1	2	21										
Tumor size (mm)	150	160	250	21	30										

Thus, all three goodness-of-fit tests, for all levels of significance, support the null hypothesis that the sizes of cancerous tumors of the brain follow the three-parameter Weibull pdf. The approximate maximum likelihood estimates of the three parameters used are $\widehat{\alpha} = 9.4826E+7$, $\widehat{\beta} = 1.4060E+9$, and $\widehat{\gamma} = 1.3940E+9$. Thus, we can write the pdf that characterizes probabilistically the malignant tumor sizes in the brain as:

$$f(x) = \frac{9.4826E+7}{1.4060E+9}\left(\frac{x+1.3940E+9}{1.4060E+9}\right)^{9.4826E+7-1} \exp\left[-\left(\frac{x+1.3940E+9}{1.4060E+9}\right)^{9.4826E+7}\right], \quad x > 0,$$

and the cumulative pdf is given by:

$$F(x) = 1 - \exp\left[-\left(\frac{x+1.3940E+9}{1.4060E+9}\right)^{9.482E+7}\right], \quad x \geq 0.$$

A graphical illustration of the three-parameter Weibull pdf along with a frequency histogram of the data is given in Fig. 14.37.

We can use this diagram to obtain approximate probabilities of the behavior of the cancerous tumor sizes. For example, the probability that the tumor size is less 60 mm is approximately 0.25, that is,

$$P(X \leq 60 \text{ mm}) \approx 0.25, \text{P[X } \leq 60 \text{ mm]} = 0.25,$$

and the probability that the tumor size is larger than 48 mm is approximately 74%, that is,

$$P(X > 48mm) = 1 - P(X \leq 48mm) \approx 0.74.$$

We also can proceed to obtain the expected value of the tumor size and approximate confidence limits on the true size of the tumor, among other interesting information.

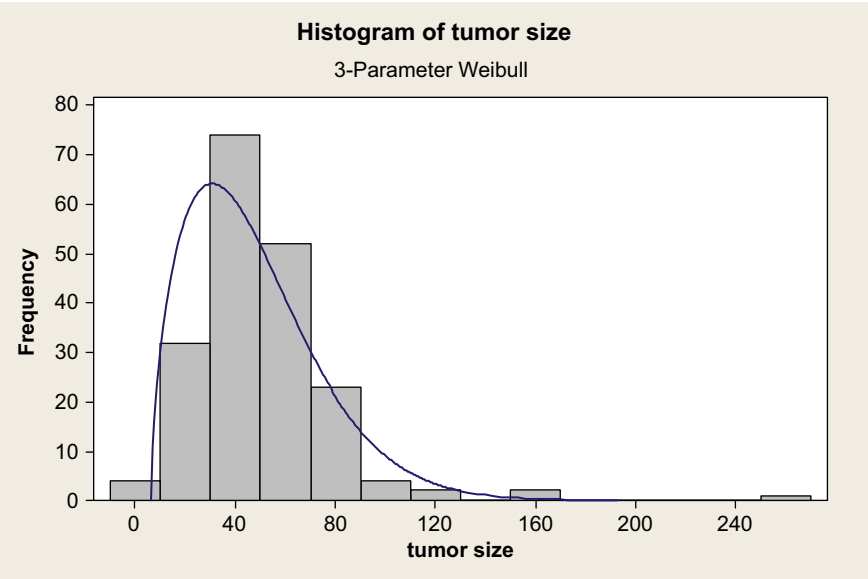

FIGURE 14.37 Three-parameter Weibull pdf for tumor size data.

TABLE 14.12 Rainfall Data.

Year	Rain	Year	Rain	Year	Rain	Year	Rain	Year	Rain	Year	Rain
1975	3.957	1981	3.68	1987	4.465	1993	4.175	1999	4.103	2005	5.22
1976	4.031	1982	5.224	1988	4.487	1994	4.889	2000	2.737	2006	3.526
1977	3.918	1983	5.639	1989	3.612	1995	5.468	2001	4.104	2007	3.211
1978	4.299	1984	3.563	1990	3.322	1996	3.668	2002	5.037		
1979	4.942	1985	3.592	1991	4.463	1997	5.029	2003	4.633		
1980	3.921	1986	4.307	1992	4.514	1998	4.73	2004	5.219		

14.8.5 Rainfall data analysis

For the southern region of the United States, we have the average annual rainfall data in inches from 1975 to 2007. The actual data are given in Table 14.12.

Using 33 measurements ($n = 33$), we wish, if possible, to identify the pdf that probabilistically characterizes the behavior of the average annual rainfall of the southern district of United States. Having such pdf we can calculate the amount of rain we will expect in the region and obtain confidence limits of the true amount of annual rainfall, among other interesting questions.

From a preliminary view of the histogram, Figure 14.38, of the data we believe that the rainfall data follows the beta pdf.

Thus, let us proceed to test our belief:

H_0 : we believe that the rainfall data follow the beta pdf

versus

H_a : the subject data do not follow the beta pdf.

Given below are the goodness-of-fit results applying the three commonly used statistical tests:

Kolmogorov–Smirnov test	$D = 0.0773$, p value 0.9806
Anderson–Darling test	$A = 0.2098$, p value 0.8836
Chi-square test	$\chi^2 = 0.2888$, p value 0.9905

For all commonly used levels of significance, $\alpha = 0.01, 0.05$, and 0.10, we strongly accept the null hypothesis that our belief is true, that is, the given rainfall data follow the beta pdf. The maximum likelihood estimates of the parameters α and β of the beta pdf are $\widehat{\alpha} = 2.2823$ and $\widehat{\beta} = 1.8754$. Thus, the beta pdf of the rainfall data is given by:

$$f(x) = \frac{\Gamma\left(2+\widehat{\beta}\right)}{\Gamma(2)\Gamma\left(\widehat{\beta}\right)} x^{\hat{\alpha}-1}(1-x)^{\hat{\beta}-1}, \quad x \geq 0,$$

or

$$f(x) = \frac{2.5237}{5.7593} x^{0.2823}(1-x)^{0.8754}, \quad x \geq 0,$$

where

$$\Gamma(2.2823+1.8754) = 2.5237, \quad \text{and} \quad \Gamma(2.2823)\Gamma(1.87554) = 5.7593.$$

The graph of the subject pdf over the histogram of the data is given in Fig. 14.38.

The expected amount of average rainfall in the southern region is 4.2998 inches, that is,

$$E(X) = \int_0^\infty xf(x)dx = 4.2998 \ \textit{inches}.$$

We can also calculate confidence limits around the true value of the annual average rainfall. For example, we are at least 95% confident that the true annual average rainfall in the southern district is between 4.0579 and 4.5167 inches.

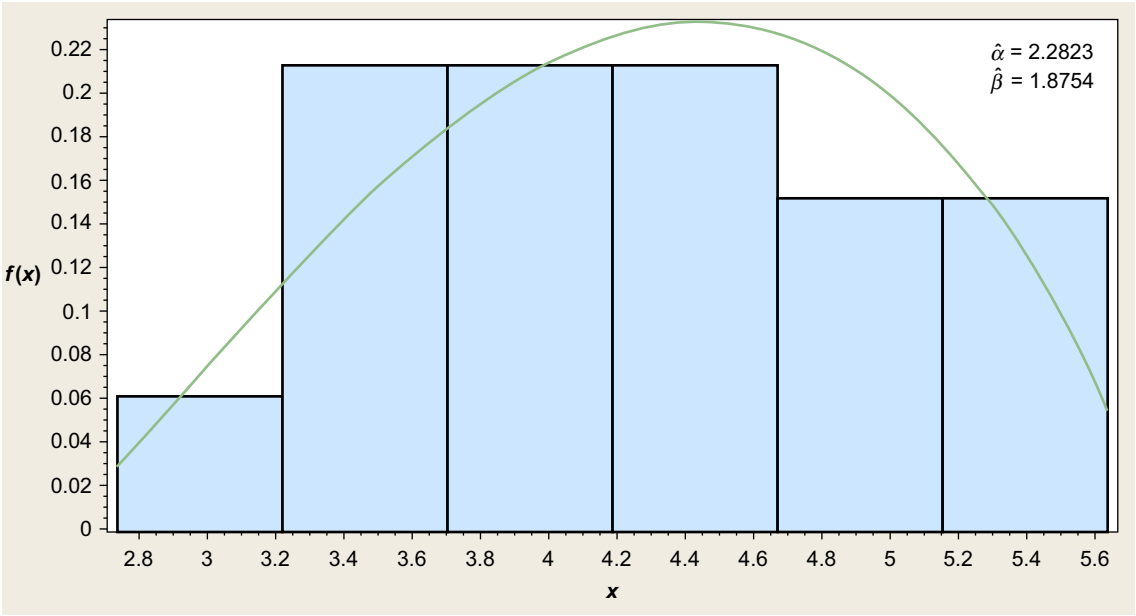

FIGURE 14.38 Beta probability density function for rainfall data.

14.8.6 Prostate cancer

In this application we will study the behavior of the cancerous tumor in prostate cancer patients. We shall use real prostate cancer data for white men from 1973 to 2007 from the SEER Program. The tumor size is the random variable of interest for 20,645 prostate cancer patients. Our primary objective is to identify the pdf that characterizes probabilistically the behavior of the cancerous tumor size in millimeters. From the initial structure of the histogram, Figure 14.39, we believe that the two-parameter Weibull pdf may fit the subject data. Thus, we set up our hypothesis to test our belief:

H_0 : the prostate cancerous tumor sizes follow the Weibull pdf

versus

H_a : the subject data do not follow the Weibull pdf.

After we apply the Kolmogorov−Smirnov, Anderson−Darling, and chi-square tests, all support the null hypothesis that the subject data follow the two-parameter Weibull pdf. The maximum likelihood estimates of the parameters α and β that drive the Weibull pdf are $\widehat{\alpha} = 0.8704$ and $\widehat{\beta} = 12.4403$.

Thus, the two-parameter Weibull pdf is given by:

$$f(x) = \frac{0.8704}{12.4403}\left(\frac{x}{12.4403}\right)^{-0.1296} \exp\left[-\left(\frac{x}{12.4403}\right)^{-0.8704}\right], \quad x \geq 0,$$

where x represents the size of the cancerous tumor in millimeters. The cumulative Weibull pdf is useful in obtaining various probabilities of the size of the tumor and is given by:

$$F(x) = P(X \leq x) = 1 - \exp\left[-\left(\frac{x}{12.4403}\right)^{-8704}\right], \quad x \geq 0.$$

Given in Fig. 14.39 is the Weibull pdf over the initial histogram along the cumulative pdf.

Thus, for an individual patient drawn at random from the subject population, we expect his cancer tumor size to be 13.341 mm, that is,

$$E(X) = \int_0^\infty x f(x) dx = 13.341 \text{ mm}.$$

Furthermore, we can calculate confidence limits around the true unknown size of the prostate tumor, that is, a 90% confidence interval for the true mean size is (0.410, 43.81). We can conclude that we are at least 90% certain that the true size of the tumor will be between 0.410 and 43.81 mm for an individual who falls in the subject population.

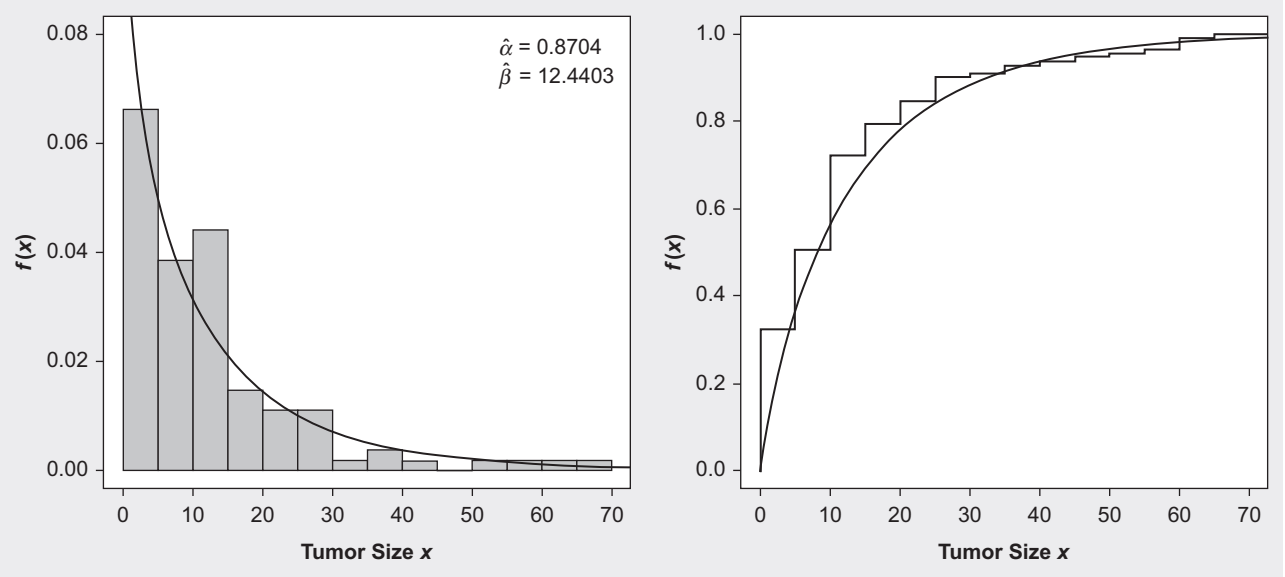

FIGURE 14.39 Weibull probability density function (*pdf*) and cumulative distribution function (*cdf*) for prostate tumor sizes.

14.8 Exercises

14.8.1. Global warming:
Carbon dioxide (CO_2) data in the United States are collected in two locations in inland Hawaii, Point Barrow and Mauna Lao.
These data are given at http://booksite.elsevier.com/9780124171138. Using the CO_2 data collected in Point Barrow from 1974 to 2004, perform the following analysis:
- **(a)** Construct a histogram of the data and interpret its visual behavior.
- **(b)** Apply the chi-square goodness-of-fit test to prove or disprove that the CO_2 data follow the exponential power probability distribution, using $\alpha = 0.05$.
- **(c)** If you have proven that the CO_2 data follow the exponential power pdf, proceed to calculate and interpret the expected value of the subject pdf.

14.8.2. Answer the same questions stated in Exercise 14.8.1 using the CO_2 data that were collected at Mauna Lao.

14.8.3. Rainfall data:
At http://booksite.elsevier.com/9780124171138 you will find the average yearly rainfall data in inches for the northern, central, and southern regions of the United States from 1975 to 2007. Using the northern region data, perform the following analysis:
- **(a)** Construct a histogram of the yearly average rainfall for the northern region. Does the histogram give you any visual indication of the type of pdf that the data follow?
- **(b)** Using the Kolmogorov−Smirnov goodness-of-fit test, verify if the subject data follow the normal pdf for $\alpha = 0.05$.
- **(c)** If you have proven that the data follow the normal pdf, what is the expected rainfall for a given year? Also, calculate the 95% confidence limits for the true average rainfall and interpret their meaning.
- **(d)** Calculate a P−P plot and interpret its visual meaning with respect to (b).

14.8.4. Using the average yearly rainfall from the central region of the United States, perform the same analysis as for the northern region and in place of the normal pdf use the gamma pdf.

14.8.5. Using the data given for the southern region of the United States, perform the same analysis as for the northern region, Exercise 14.8.3, with the normal pdf replaced by the beta pdf.

14.8.6. Hurricane Katrina:
Hurricane Katrina was the most devastating hurricane to hit the United States in the past 100 years. Katrina was an Atlantic-based category 5 hurricane that reached wind velocities of more than 160 mph. At http://booksite.elsevier.com/9780124171138 you will find 63 measurements of the wind velocity of Katrina. Using the subject data perform the following analysis:

(a) We believe that the wind velocity measurements of Hurricane Katrina follow the three-parameter Weibull pdf. Test this belief using:
  (i) the Kolmogorov−Smirnov goodness-of-fit test;
  (ii) the Anderson−Darling goodness-of-fit test;
  using $\alpha = 0.05$
(b) Discuss the results of (i) and (ii) above. What conclusion have you reached about our belief?
(c) If our belief is correct, write the complete form of the pdf that characterizes the behavior of the wind velocity of Hurricane Katrina.
(d) If in the future we experience a category 5 hurricane, what would the expected velocity of such a hurricane be?

14.8.7. With respect to Exercise 14.8.6, there is a group of scientists who believe that the Rayleigh pdf is a better fit of the wind speed measurements of Hurricane Katrina. Follow the same questions posed in Exercise 14.8.6 using the Rayleigh pdf. What do you conclude in comparing the results of Exercise 14.8.6 with those of Exercise 14.8.7?

14.8.8. National unemployment:
At http://booksite.elsevier.com/9780124171138 you will find the data for the annual average percentage of unemployment for the United States from 1957 to 2007. Using these data perform the following analysis:
(a) Construct a histogram of the data. Does this histogram convey any useful information concerning the behavior of the data?
(b) Using the goodness-of-fit test of your choice, can you identify the pdf that characterizes the behavior of the data, that is, the pdf that the subject data were drawn from using $\alpha = 0.05$.
(c) Once you have found the subject pdf of the unemployment data, calculate the expected value of the annual average percentage of unemployment rate.

14.8.9. Breast cancer:
At http://booksite.elsevier.com/9780124171138 we have the malignant breast tumor sizes in millimeters of 250 breast cancer patients. In the database, draw a random sample of tumor sizes of $n = 50$ breast cancer patients. For the 50 tumor sizes in millimeters perform the following analysis:
(a) Construct a histogram of the 50 tumor sizes. Discuss any visual information you might obtain concerning the possible pdf that characterizes the data behavior.
(b) Identify, if possible, a pdf that you believe may characterize the given data, using one or more of the goodness-of-fit tests, using $\alpha = 0.05$.
(c) If you were not able to identify the pdf, why not? If you were successful, identify completely the pdf, with appropriate parameter estimates.
(d) If you have identified correctly the pdf, calculate and interpret the expected value of the subject data.

14.8.10. At http://booksite.elsevier.com/9780124171138 we have the survival times (in years) of 250 breast cancer patients, that is, the age at which they died due to breast cancer. From this database, draw a random sample of $n = 50$ survival times. Use these survival times to perform the following analysis:
(a) Construct a histogram to possibly guide you in identifying the pdf of the subject data.
(b) Use any of the goodness-of-fit tests to search for the correct pdf that characterizes the behavior of the given survival times for $\alpha = 0.05$.
(c) State completely the pdf you have identified and discuss its usefulness in obtaining information about the subject data.
(d) Obtain the cumulative distribution function $F(t)$ of the pdf $f(t)$ you have found. If you take 1 minus the $F(t)$ you will obtain the survival function, $S(t)$, of the given data. That is, $S(t) = 1 - F(t)$. The survival function, $S(t)$, gives you the probability that a given patient drawn from the database of 250 breast cancer patients will survive a specified year.
(e) Write the survival function of the given data set and graph it, that is, $S(t)$ versus t. Discuss the useful information that the graph gives concerning breast cancer patients.

14.8.11. Lung cancer:
At http://booksite.elsevier.com/9780124171138 we have the malignant tumor sizes for male and female lung cancer patients. We also include the survival times of both genders, that is, the age in years at which they died due to lung cancer. From the male database draw a random sample of $n = 60$ malignant tumor sizes and perform the following analysis:
(a) Construct a histogram of the 60 measurements of the tumor sizes in millimeters.
(b) Let (a) guide you, if possible, in performing goodness-of-fit testing at $\alpha = 0.05$ to identify the best possible pdf that characterizes the probabilistic behavior of the tumor sizes.
(c) Write the pdf completely with appropriate parameter estimates and obtain and interpret its expected value.

14.8.12. Proceed to obtain a random sample of $n = 60$ from the female database and perform the same analysis as in (a)−(c) in Exercise 14.8.11.

14.8.13. Give a precise comparison of males and females for each of the analyses you performed in (a)−(c) of Exercises 14.8.11 and 14.8.12. Discuss your comparison findings.

14.8.14. In the lung cancer database we have also given information about the survival times of male and female lung cancer patients. Take a random sample of $n = 50$ of the survival times of male lung patients and proceed to perform the same analysis for the survival times as in Exercise 14.8.10.

14.8.15. Similar to Exercise 14.8.4, proceed to take a random sample of $n = 50$ of the survival times of female lung patients and perform the same analysis as you did for the male patients in Exercise 14.8.14.

14.8.16. Give a precise comparison of the analysis of the findings of male and female lung patients that you made in Exercises 14.8.14 and 14.8.15, respectively. Discuss your comparison findings.

14.8.17. Colon cancer:
At http://booksite.elsevier.com/9780124171138 we have the malignant tumor sizes of male and female colon cancer patients. From this database draw a random sample of $n = 50$ tumor sizes of the male colon cancer patients. Using these data proceed to perform the same analysis that you did for the lung cancer data in Exercise 14.8.11.

14.8.18. Proceed to draw a random sample of $n = 50$ from the female database that gives the malignant colon tumor size. Perform the same analysis for the females that you did for the males in Exercise 14.8.17.

14.8.19. In the colon cancer database we also give the survival times for both male and female patients. From the male database draw a random sample of $n = 60$ survival times and proceed to perform the same analysis as you did in Exercise 14.8.11.

14.8.20. From the survival times of female colon cancer patients draw a random sample of $n = 60$ and proceed with the same analysis that you did for the male patients in Exercise 14.8.19.

14.8.21. Give a precise comparison of the survival times analyses in (a)−(c) for males and females.

14.9 Conclusion

We have briefly discussed some of the real-world problems that arise in applied data analysis. However, this discussion is not exhaustive. There are various other special problems that can arise in applied data analysis. For example, if one or both of the sample sizes are small, it may be hard to detect violations of some of the assumptions. For small samples, violation of assumptions such as inequalities of variances is hard to discover. Also, for small sample sizes, possible outliers whose detection may be in doubt may have undue influence on the inferences. It is better to avoid such problems in the design stage of an experiment, when suitable sample sizes can be determined before we start collecting data.

Differences in distributional shapes can influence the testing procedures of two or more samples. In those cases, utilizing a transformation may settle that problem and may also promote normality as well as correcting the problem of inequality of variances. There are also many issues related to simulation that are discussed in Chapter 13 in the utilization of empirical methods—for instance, in the application of Markov chain Monte Carlo methods, the issues of burn-in, the choice of the correct proposal function, and convergence. These are beyond the scope of this book.

Combining the issues discussed in this chapter with the rest of the material in this textbook should give the student a good footing in the theory of statistics as well as the ability to deal with many real-world problems.