PRACTICAL STATISTICAL METHODS

A SAS Programming Approach

Lakshmi V. Padgett

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

PRACTICAL STATISTICAL METHODS

A SAS PROGRAMMING APPROACH

PRACTICAL STATISTICAL METHODS

A SAS PROGRAMMING APPROACH

Lakshmi V. Padgett

CRC Press
Taylor & Francis Group
Boca Raton London New York

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

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

© 2011 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20110701

International Standard Book Number-13: 978-1-4398-1254-9 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

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

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

To my father and

in loving memory of my mother

Contents

Preface xiii

1. Introduction 1
1.1 Types of Data 1
1.2 Descriptive Statistics/Data Summaries 2
1.3 Graphical and Tabular Representation 4
1.4 Population and Sample 12
1.5 Estimation and Testing Hypothesis 14
1.6 Normal Distribution 16
1.7 Nonparametric Methods 17
1.8 Some Useful Concepts 17

2. Qualitative Data 21
2.1 One Sample 21
2.1.1 Binary Data 21
2.1.2 *t* Categorical Responses 24
2.2 Two Independent Samples 25
2.2.1 Two Proportions 25
2.2.2 Odds Ratio and Relative Risk 30
2.2.3 Logistic Regression with One Dichotomous Explanatory Variable 31
2.2.4 Cochran–Mantel–Haenszel Test for a 2×2 Table 33
2.2.5 *t* Categorical Responses 35
2.3 Paired Two Samples 36
2.3.1 Binary Responses 36
2.3.2 *t* Categorical Responses 38
2.4 *k* Independent Samples 40
2.4.1 *k* Proportions 40
2.4.2 Logistic Regression When the Explanatory Variable Is Not Dichotomous 41
2.4.3 CMH Test 43
2.4.4 *t* Categorical Responses 44
2.5 Cochran's Test 45
2.6 Ordinal Data 46
2.6.1 Row Mean Score Test 46
2.6.2 Cochran–Armitage Test 48
2.6.3 Measures of Association 48
2.6.4 Ridit Analysis 49
2.6.5 Weighted Kappa 51

2.6.6 Ordinal Logistic Regression 52
2.6.6.1 Two Samples 52
2.6.6.2 k Samples 53

3. Continuous Normal Data 57
3.1 One Sample 57
3.2 Two Samples 62
3.2.1 Independent Samples 62
3.2.1.1 Means 62
3.2.1.2 Variances 66
3.2.2 Paired Samples 66
3.3 k Independent Samples 69
3.3.1 One-Way Analysis of Variance 69
3.3.1.1 Variance 78
3.3.2 Covariance Analysis 79
3.4 Multivariate Methods 81
3.4.1 Correlation, Partial, and Intraclass Correlation 85
3.4.2 Hotelling's T^2 90
3.4.2.1 One Sample 90
3.4.2.2 Two Samples 93
3.4.3 One-Way Multivariate Analysis of Variance 96
3.4.4 Profile Analysis 98
3.4.5 Discriminant Functions 101
3.4.6 Cluster Analysis 104
3.4.7 Principal Components 106
3.4.8 Factor Analysis 109
3.4.9 Canonical Correlation 111
3.5 Multifactor ANOVA 113
3.5.1 Crossed Factors 113
3.5.2 Tukey 1 df for Nonadditivity 115
3.5.3 Nested Factors 116
3.6 Variance Components 118
3.7 Split Plot Designs 120
3.8 Latin Square Design 123
3.9 Two-Treatment Crossover Design 126

4. Nonparametric Methods 131
4.1 One Sample 131
4.1.1 Sign Test 131
4.1.2 Wilcoxon Signed-Rank Test 133
4.1.3 Kolmogorov Goodness of Fit 134
4.1.4 Cox and Stuart Test 134
4.2 Two Samples 135
4.2.1 Wilcoxon–Mann–Whitney Test 136
4.2.2 Mood's Median Test 137

4.2.3 Kolmogorov–Smirnov 138
4.2.4 Equality of Variances 139
4.3 *k* Samples 141
4.3.1 Kruskal–Wallis Test 141
4.3.2 Median Test 143
4.3.3 Jonckheere Test 143
4.4 Transformations 144
4.5 Friedman Test 145
4.6 Association Measures 147
4.6.1 Spearman Rank Correlation 147
4.6.2 Kendall's Tau 149
4.6.3 Kappa Statistic 149
4.7 Censored Data 150
4.7.1 Kaplan–Meier Survival Distribution Function 151
4.7.2 Wilcoxon (Gehan) and Log-Rank Test 153
4.7.3 Life-Table (Acturial Method) 154

5. Regression 157
5.1 Simple Regression 158
5.2 Polynomial Regression 162
5.3 Multiple Regressions 164
5.3.1 Multicollinearity 168
5.3.2 Dummy Variables 169
5.3.3 Interaction 171
5.3.4 Variable Selection 171
5.4 Diagnostics 175
5.4.1 Outliers 177
5.4.2 Influential Observations 178
5.4.3 Durbin–Watson Statistic 179
5.5 Weighted Regression 180
5.6 Logistic Regression 181
5.6.1 Dichotomous Logistic Regression 181
5.6.2 Multinomial Logistic Model 185
5.6.3 Cumulative Logistic Model 188
5.7 Poisson Regression 190
5.8 Robust Regression 194
5.9 Nonlinear Regression 196
5.10 Piecewise Regression 198
5.11 Accelerated Failure Time (AFT) Model 199
5.12 Cox Regression 203
5.12.1 Proportional Hazards Model 203
5.12.2 Proportional Hazard Assumption 207
5.12.3 Stratified Cox Model 208
5.12.4 Time-Varying Covariates 209
5.12.5 Competing Risks 210

5.13 Parallelism of Regression Equations 213
5.14 Variance-Stabilizing Transformations 215
5.15 Ridge Regression 216
5.16 Local Regression (LOESS) 218
5.17 Response Surface Methodology: Quadratic Model 218
5.18 Mixture Designs and Their Analysis 221
5.19 Analysis of Longitudinal Data: Mixed Models 224

6. Miscellaneous Topics 227
6.1 Missing Data 227
6.2 Diagnostic Errors and Human Behavior 231
6.2.1 Introduction 231
6.2.2 Independent Samples 232
6.2.2.1 Two Independent Samples 232
6.2.2.2 *k* Independent Samples 234
6.2.3 Two Dependent Samples 236
6.2.4 Finding the Threshold for a Screening Variable 237
6.2.5 Analyzing Response Data with Errors 239
6.2.6 Responders' Anonymity 241
6.3 Density Estimation 242
6.3.1 Parametric Density Estimation 242
6.3.2 Nonparametric Univariate Density Estimation 243
6.3.3 Bivariate Kernel Estimator 245
6.4 Robust Estimators 248
6.5 Jackknife Estimators 250
6.6 Bootstrap Method 251
6.7 Propensity Scores 252
6.8 Interim Analysis and Stopping Rules 255
6.8.1 Stopping Rules 255
6.8.2 Conditional Power 256
6.9 Microarrays and Multiple Testing 257
6.9.1 Microarrays 257
6.9.2 Multiple Testing 258
6.10 Stability of Products 260
6.11 Group Testing 260
6.12 Correspondence Analysis 261
6.13 Classification Regression Trees 263
6.14 Multidimensional Scaling 265
6.15 Path Analysis 267
6.16 Choice-Based Conjoint Analysis 272
6.16.1 Availability Designs and Cross Effects 273
6.16.2 Pareto-Optimal Choice Sets 274
6.16.3 Mixture-Amount Designs 275

6.17 Meta-Analysis 276
6.17.1 Homogeneity of the Effect Sizes 277
6.17.2 Combining the *p*-Values 277

References and Selected Bibliography 279

Preface

This book discusses various statistical methods aimed at a diverse audience. A broad spectrum of methods is given that will be useful for researchers without much statistical background. Omitting mathematical details and complicated formulae, SAS programs are provided to carry out the necessary analyses and draw appropriate inferences for statistical problems that are widely used by most researchers. Similar types of programs can be used with other statistical packages, drawing similar conclusions. The book emphasizes the different, commonly used methodologies and interpretation of the data using SAS version 9.1.

Practical Statistical Methods: A SAS Programming Approach will assist researchers in all fields, including those in the pharmaceutical industry, the biological sciences, physical sciences, social sciences, market research, and psychology. Additionally, it will be a reference book for applied statisticians working with researchers in the fields indicated above. It will also be helpful as supplemental reading material for graduate students taking a statistical methods course for one or two semesters.

Fundamental statistical concepts are introduced in Chapter 1 and the methods used for quantitative data, continuous data following normal and nonnormal distributions, are discussed in Chapters 2 through 4. Regression methodology starting with simple linear regression and discussing logistic regression and PH regression are highlighted in Chapter 5. Chapter 6 briefly sketches miscellaneous topics such as propensity scores, cutoff points of a marker indicating the end point, misclassification errors, interim analysis, conditional power, bootstrap and jackknife, among others.

Lakshmi V. Padgett

1

Introduction

1.1 Types of Data

Data play a crucial part in statistical inference. The type of data one has, determines the type of statistical analysis needed. The methods of analyses performed on one type of data may not necessarily be applicable for other types of data. Further, one needs to have clean data, one without any outliers (observations that do not follow the pattern of the bulk of the data) and holes (chunks of missing observations).

There are *qualitative* and *quantitative* types of data.

Qualitative data include variables such as a persons' eye color, income level (low, middle, high), winning a medal (gold, silver, bronze), an individual having a particular disease or not, and age (<21, 21–35, 36–55, >55). Qualitative data consist of *nominal data* and *ordinal data*. Nominal data have no particular ordering like an individuals' eyecolor, brands of soft drinks, and brands of laptops. Ordinal data can be ranked by category such as income level, winning a medal, movie ratings, and students' letter grades in a course.

Data that can be measured or things that can be counted include age, weight, leaves on a tree, and the number of cars in a household. These types of data are called quantitative data. Quantitative data can be discrete or continuous. Usually all measuring variables are continuous and all counting variables are discrete, with the exception of monetary variable which is treated as continuous. Discrete variables include the number of children in a family, the number of deaths due to a disease at a given time, the number of leaves on a tree, and the number of cars in a household. Age, weight, height, cholesterol level, serum glucose, and blood pressure are examples of continuous variables. Depending on the precision one wants, the continuous variables such as weight can be 125 lbs or 125.1 lbs. The data for the continuous variables may look like discrete variables.

Continuous data can be converted to ordinal data based on the question one is trying to address. Thus, the age of an individual can be given in years, or it can be expressed categorically as a child (<13 years), teenager [13–18), adult [18–55), or senior citizen (≥55 years) depending on whether one is interested in trying to address the average age of people in a community or

the proportion of teenagers in a community. The "[" means inclusive of that particular observation and ")" means exclusive of that observation.

1.2 Descriptive Statistics/Data Summaries

Consider the age of 50 people who entered a clinical trial. We want to summarize these age groups by providing quantities that express the nature of the data. These quantities that represent the data are called *descriptive* or *summary statistics*.

All of the summary statistics can be obtained using PROC UNIVARIATE in SAS as follows:

```
Data a;
/***Input data***/
Input Age @@; Cards;
45 55 67 89 78 67 56 43 44 45 57 69 78 76 54 34 23
46 67 65 54 98 78 67 58 49 53 42 74 97 43 45 46 47
48 59 70 89 78 76 65 34 45 36 67 45 47 75 64 61
;
/***end of input data***/
proc univariate; var age;run;
```

The first part of the output which will be used to explain the descriptive statistics is

```
                    The UNIVARIATE Procedure
                          Variable: Age
                        Moments
N                 50                Sum weights       50
Mean              59.36 (a1)        Sum observations  2968
Std deviation     16.8713741 (a6)   Variance          284.643265
Skewness          0.32002628        Kurtosis          -0.3227663
Uncorrected SS    190128            Corrected SS      13947.52
Coeff. variation  28.4221262 (a7)   Std error mean    2.38597261

                  Basic Statistical Measures
        Location                        Variability
Mean       59.36000        Std deviation         16.87137
Median     57.50000 (a2)   Variance              284.64327
Mode       45.00000 (a3)   Range                 75.00000 (a9)
                           Interquartile range   25.00000 (a8)
```

Note: The mode displayed is the smallest of two modes with a count of 5.

```
                    ***
Quantile            Estimate
100% Max            98.0
99%                 98.0
95%                 89.0
90%                 78.0
75% Q3              70.0 (a5)
50% Median          57.5
25% Q1              45.0 (a4)
10%                 42.5
5%                  34.0
1%                  23.0
0% Min              23.0
```

The commonly used descriptive statistics are mean, median, mode, Q1, Q3, standard deviation, and the coefficient of variation. These statistics are described below:

Mean is the average of the data, and from our output (a1) is 59.36. If the data have two categories, one of them coded as "1" and the other "0," then the mean is the proportion of the category coded as "1."

Approximately 50% of the observations are not more than the *median*, and approximately 50% of the observations are not less than the median. From the output, (a2) represents the median and is 57.5. If the mean and median are close, the mean is used as a central tendency measure of the data. However, if the mean and median are farther apart, then the median is used as the focal point of the data.

Mode is that observation that is repeated the most. From the output, (a3) represents the mode and is 45.0. Sometimes you may have more than one mode. In SAS, the smallest of the modes is reported.

Approximately 25% of the observations are less than or equal to Q1, the *lower quartile*, and approximately 75% of the observations are greater than or equal to Q1. From the output, (a4) represents the lower quartile and is 45.0.

Approximately 75% of the observations are less than or equal to Q3, the *upper quartile*, and approximately 25% of the observations are greater than or equal to Q3. From the output, (a5) represents the upper quartile and is 70.0.

The spread of the data from the mean is indicated by the *standard deviation*. It is 16.87 and is shown by (a6) in our output. The standard deviation is always nonnegative. If the standard deviation is small, the data are clustered near the mean and if the standard deviation is large, the data are widely spread out from the mean. The square of the standard deviation is the variance and is given in the output.

The mean and standard deviation are given in the same units of measurement as the data. If data are collected in different scales or sources, the *coefficient of variation* provides a unit-free measure of the spread of the data. In a group of people with mean annual income of $20,000, a person with an

annual income of $21,000 is considered rich, whereas in a group of individuals with an annual income of $100,000, a person making $101,000 is not considered rich, although the squared deviations of the individuals from the group means are the same. Thus, if two groups of data have the same standard deviation with different means, the variability will have practical importance if it is considered in terms of the mean, and the coefficient of variation takes that aspect into consideration and is a better measure to compare the variabilities of groups of data. While comparing the variability of two variables expressed in different units of measurement, the coefficient of variation is also an appropriate measure. From the output, (a7) is the coefficient of variation and is 28.4.

The *interquartile range* for the data from the output at (a8) is 25 and this is Q3–Q1. Approximately 50% of the data is separated by 25 units starting at Q1 = 45 and ending at Q3 = 70. Some researchers indicate the interquartile range as (45, 70).

The *range* for the data from the output at (a9) is 75 and this is the difference between the maximum of all the observations and the minimum of all the observations. Some researchers indicate the range as (23, 98).

For further details, see Raghavarao (1988).

1.3 Graphical and Tabular Representation

Data can either be presented in a tabular form or can be plotted and displayed graphically. Figures sometimes help to get a clearer and a more concise picture of what the data represent. However, one needs to find a balance between figures and tabular forms of data representation.

There are many statistical packages that graph the data, including but not limited to SAS, S-Plus, JMP, and Minitab. Some packages have better graphical qualities, whereas others are more user-friendly. It is not in the scope of this book to evaluate these packages. Some of the common types of graphical displays include stem-and-leaf plots, histograms, bar charts, pie charts, line plots, box plots, scatter plots, and Kaplan–Meier (KM) curves. We will now describe them.

Stem-and-Leaf Plot is similar to the histogram. The advantage of using a stem-and-leaf plot is that we know the individual data points up to the indicated units in the stem-and-leaf plot, whereas in the histogram we only know the class the observation falls into, but not the individual value. Tukey (1977) also provided other graphical representations. Let us consider the age of the 50 people in a clinical trial discussed in Section 1.2.

Figure 1.1 depicts the stem-and-leaf plot.

The leaf is the unit value of the age and the stem is the tens of the age. Leaf 9 at stem 4 corresponds to observation 49. The counts of leaves from the top

Age (Leaf unit = 1)

	Stem	Leaf
1	2	3
4	3	4 4 6
19	4	2 3 3 4 5 5 5 5 5 6 6 7 7 8 9
(8)	5	3 4 4 5 6 7 8 9
23	6	1 4 5 5 7 7 7 7 7 9
13	7	0 4 5 6 6 8 8 8 8
4	8	9 9
2	9	7 8

FIGURE 1.1
Stem-and-leaf plot for age.

and bottom are shown on the left side. The median observation is in the stem where the count is shown within parentheses.

The *histogram* shown in Figure 1.2 is similar to the stem-and-leaf plot except that the information is condensed. The boxes are obtained by dividing the range of a numerically continuous variable, categorizing into several non-overlapping classes with no gaps and the classes may or may not be of equal

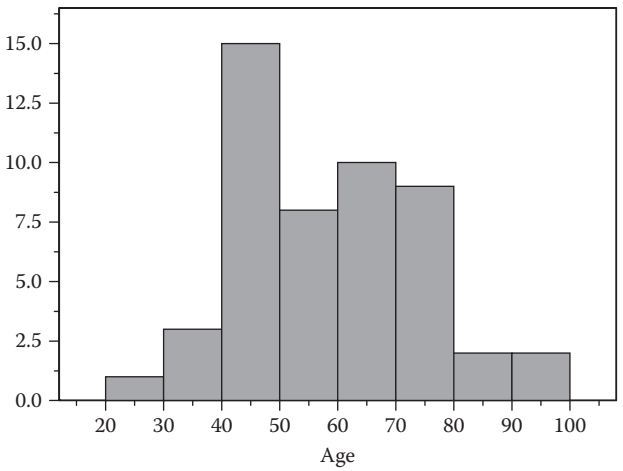

FIGURE 1.2
Histogram for age.

length. The area of the bar indicates the frequency in the corresponding class. In Figure 1.2, we form the classes [20, 30), [30–40), [40–50), [50–60), [60–70), [70–80), [80, 90), and [90, 100).

The *bar chart* is similar to a histogram, where the bars are separated from each other. In a histogram, all the bars are joined together indicating that the data are continuous, whereas in a bar chart, each bar stands separate, indicating that the data are discrete or qualitative. Qualitative or discrete data may be shown by the bars with the frequency indicating the bar height. The average temperature for the 12 months in a year in a city is shown by a bar diagram in Figure 1.3.

Figure 1.3 shows how temperatures in different months fluctuate and indicate the hottest and coldest months in the city.

Suppose one wants to compare the monthly temperatures in four cities for the first 3 months of a year. It can be graphically depicted by multiple bars as shown in Figure 1.4.

A *pie chart* is commonly used to represent frequency type of data on a categorical variable. Classes are mutually exclusive. Figure 1.5 represents a pie diagram of the percentage of hospitals that treated a given number of patients in an emergency room in an average week.

A pie chart can also be used to indicate the percentage of salary spent on housing, food, clothing, entertainment, and savings of a typical American family.

In the longitudinal data graph, the values (y-axis) are plotted across time (x-axis). A *line plot* is a graph joining all the points by lines. They can be plotted in a single graph for continuous data across time for different groups of

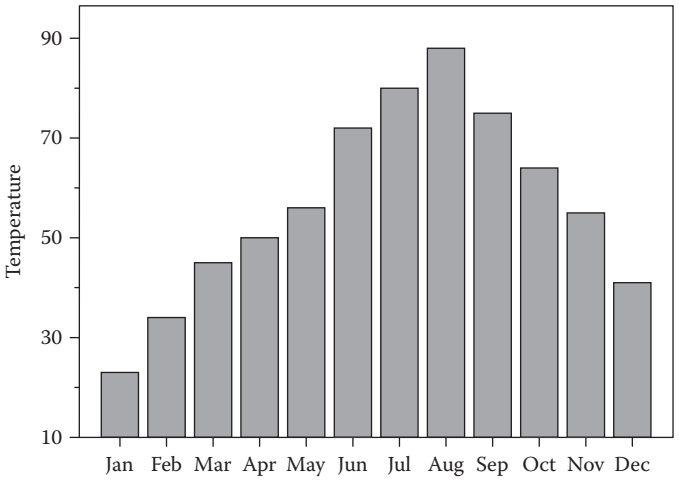

FIGURE 1.3
Bar chart of average temperature for a city.

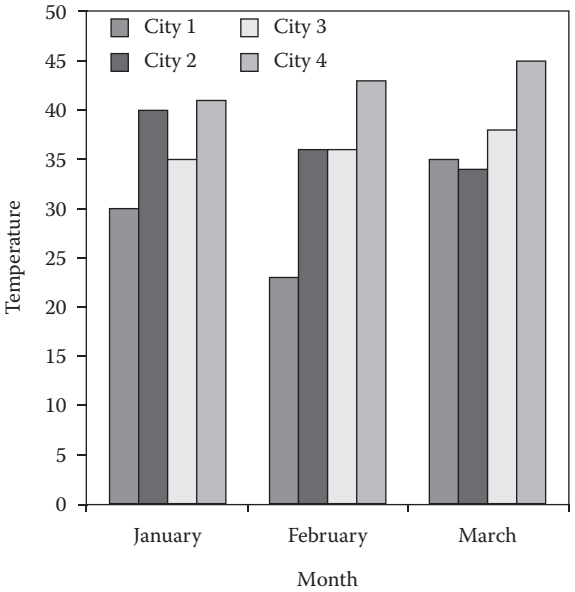

FIGURE 1.4
Bar chart of average temperature in the first 3 months for four cities.

interest. They can also be used to plot the mean or median temperature in a city by month as shown in Figure 1.6.

One can visualize the *box plot* given in Figure 1.7 as an extension of a histogram containing some summary statistics. It is also useful to detect outliers. The line in the middle of the box represents the median value and the dot represents the mean value. The first and third quartiles are the ends of the box. The lines at the end are the whiskers indicating the minimum and maximum observation. Potential outlier points can be indicated on the graph. For this purpose, we form inner fences $Q_1 - 1.5$ (IQR), $Q_3 + 1.5$ (IQR); and outer

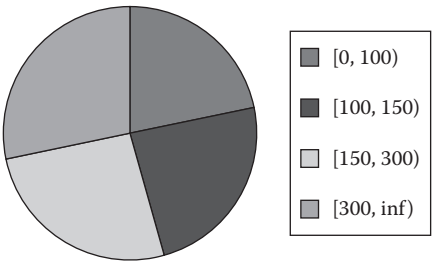

FIGURE 1.5
Pie chart of percentage of hospitals that treated a given number of patients.

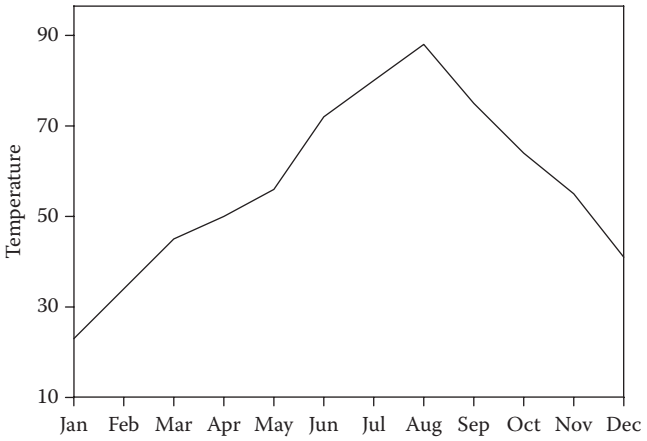

FIGURE 1.6
Line plot of average temperature in a city.

fences $Q_1 - 3$ (IQR), $Q_3 + 3$ (IQR), where IQR is the interquartile range. Observations between inner and outer fences are considered as possible outliers and beyond outer fences are definite outliers.

The *scatter plot* is used to observe an association between two continuous variables such as height and weight (see Figure 1.8). The closer the points are near a straight line, the more correlated are the variables. However, spread-out-points across the plane indicate less correlation of the variables.

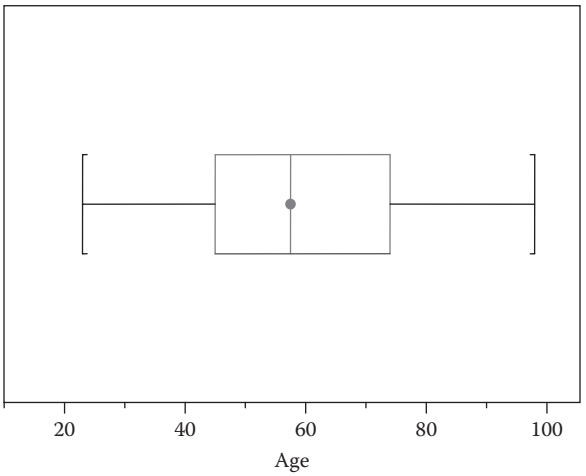

FIGURE 1.7
Box plot for age.

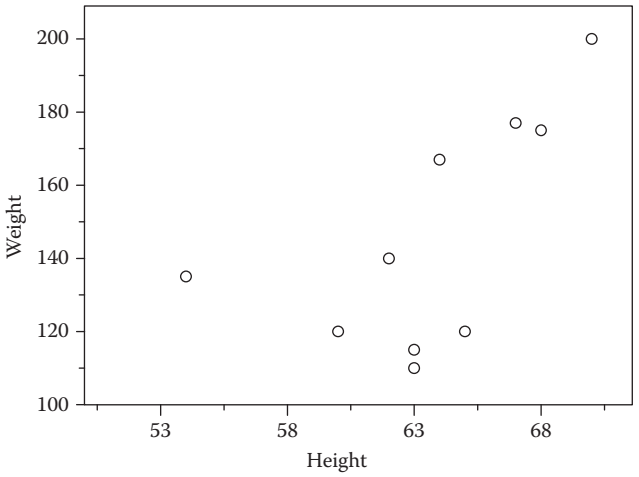

FIGURE 1.8
Scatter plot of height versus weight.

By looking at the scatter plot one detects outliers. It can also help in modeling the dependent variable on the y-axis in terms of the independent variable on the x-axis.

The *KM curve* is a step function indicating the chance of survival at different time points (see Figure 1.9). This is based on the data showing the time of an event like death. This is also used when the exact time of the event is not recorded, but is known that it exceeds some censored value.

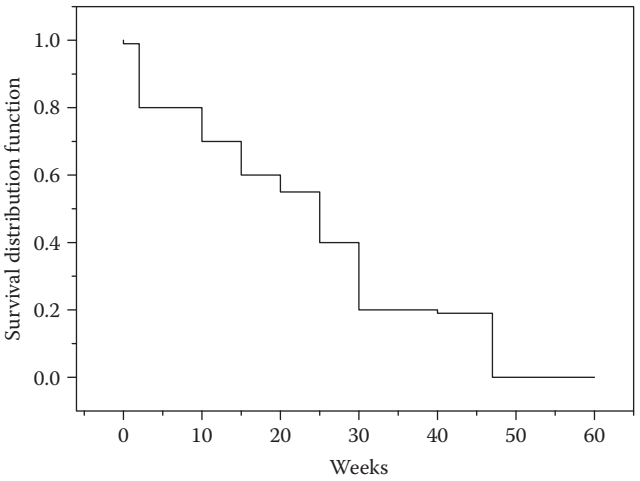

FIGURE 1.9
Survival function.

The histogram is a graphical representation of a frequency table. For the histogram chart displayed in Figure 1.2 one can use PROC FREQ in SAS and get the frequency table as follows:

```
data b;set a;
*Use age data from Sec 1.2; *Categories are defined by the
researcher;
if 20<=age<30 then AGECAT='[20,30)';
else if 30<=age<40 then AGECAT='[30,40)';
else if 40<=age<50 then AGECAT='[40,50)';
else if 50<=age<60 then AGECAT='[50,60)';
else if 60<=age<70 then AGECAT='[60,70)';
else if 70<=age<80 then AGECAT='[70,80)';
else if 80<=age<90 then AGECAT='[80,90)';
else if 90<=age<100 then AGECAT='[90,100)';
proc freq;tables agecat;run;
```

The SAS provides the frequency for the histogram categories in the following output:

The FREQ Procedure

AGECAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
[20,30)	1	2.00	1	2.00
[30,40)	3	6.00	4	8.00
[40,50)	15	30.00	19	38.00
[50,60)	8	16.00	27	54.00
[60,70)	10	20.00	37	74.00
[70,80)	9	18.00	46	92.00
[80,90)	2	4.00	48	96.00
[90,100)	2	4.00	50	100.00

From this output we see the range of categories in column 1 and the frequency in column 2. The percentage of observations in each category is shown in column 3. Cumulative frequency given in column 4 is the total frequency upto and including that category. Cumulative percentage given in column 5 is the cumulative frequency expressed as the percentage of the total number of observations. For example, in the [60, 70) age category, 10 in column 2 indicates that 10 individuals are aged 60 and above and less than 70; this number is 20% of the group and is shown by 20 in column 3. The number 37 in column 4 of this row indicates that 37 people are aged less than 70, and this number is 74% of the group and is shown by 74 in column 5.

When two variables are involved, we form a *contingency table* (cross tabulation). If age and systolic blood pressure are measured for 50 people, the

following contingency table summarizes the information by using a similar modified command as earlier:

```
Data a;
/***Input data***/
Input age syst @@; cards;
45 100 55 125 67 155 89 145 78 187 67 167
56 163 43 158 44 148 45 157 57 169 69 147
78 178 76 156 54 135 34 167 23 124 46 165
67 147 65 148 54 138 98 144 78 136 67 188
58 168 49 148 53 148 42 158 74 144 97 149
43 137 45 128 46 159 47 157 48 147 59 136
70 135 89 167 78 135 76 164 65 132 34 173
45 145 36 156 67 125 45 156 47 145 75 164
64 183 61 128
;
/***end of input data***/
data b;set a;
if syst<145 then SYSTD='SYST<145';
else if syst>=145 then SYSTD='SYST >= 145';
if 20<=age<30 then AGECAT='[20,30)';
else if 30<=age<40 then AGECAT='[30,40)';
else if 40<=age<50 then AGECAT='[40,50)';
else if 50<=age<60 then AGECAT='[50,60)';
else if 60<=age<70 then AGECAT='[60,70)';
else if 70<=age<80 then AGECAT='[70,80)';
else if 80<=age<90 then AGECAT='[80,90)';
else if 90<=age<100 then AGECAT='[90,100)';
proc freq;tables systd*agecat;run;
```

This provides the following contingency table:

```
Table of SYSTD by AGECAT

SYSTD AGECAT

Frequency ,
Percent   ,
Row Pct   ,
Col Pct   ,  [20,30) , [30,40) , [40,50) , [50,60) , [60,70) , [70,80) ,  [80,90) , [90,100) ,  Total

SYST<145 ,        1 ,       0 ,       3 ,       4 ,       3 ,       4 ,        0 ,        1 ,     16
         ,     2.00 ,    0.00 ,    6.00 ,    8.00 ,    6.00 ,    8.00 ,     0.00 ,     2.00 ,  32.00
         ,     6.25 ,    0.00 ,   18.75 ,   25.00 ,   18.75 ,   25.00 ,     0.00 ,     6.25 ,
         ,   100.00 ,    0.00 ,   20.00 ,   50.00 ,   30.00 ,   44.44 ,     0.00 ,    50.00 ,

SYST>=145 ,       0 ,       3 ,      12 ,       4 ,       7 ,       5 ,        2 ,        1 ,     34
          ,     0.00 ,    6.00 ,   24.00 ,    8.00 ,   14.00 ,   10.00 ,     4.00 ,     2.00 ,  68.00
          ,     0.00 ,    8.82 ,   35.29 ,   11.76 ,   20.59 ,   14.71 ,     5.88 ,     2.94 ,
          ,     0.00 ,  100.00 ,   80.00 ,   50.00 ,   70.00 ,   55.56 ,   100.00 ,    50.00 ,

Total             1        3       15        8       10        9         2         2      50
               2.00     6.00    30.00    16.00    20.00    18.00      4.00      4.00  100.00
```

From the output, let us look at [50, 60) age group and syst <145 cell. The first number 4 indicates that four people belong to this category. The

number 8 in the second row indicates the percentage of subjects who fall into this cell from all people. The number 25 in the third row is the percentage of people falling into this category out of people with syst <145, row category. The last row is 50 and indicates that 50% of people belong to this category out of people who are in the [50, 60) age group of the column category.

1.4 Population and Sample

Population is the entire group of data on which the experimenter or scientist wants to make a statistical inference. By the entire group we mean all people over 75 years, all households with two cars, all people having a particular disease, and so on. In a population we require measures to describe the characteristics we are interested in. These measures are called parameters. *Parameters* could be the proportion of people above 75 years who had a heart attack, the mean incubation period of a disease, and so forth.

Suppose we are interested in determining the proportion of people over 75 years who had a heart attack. It is very difficult to get all people over 75 years and ask them if they had a heart attack. So we get our information by taking a random *sample* from the population. The sample should represent the people aged over 75 years of different nationalities, different lifestyles, different socioeconomic status, and so forth. The summary measures calculated from the sample are called *statistics*.

Randomization allows one to unbiasedly make an inference on the population parameters based on sample estimates. If we are conducting a survey, then we need to make sure we randomly select our respondents to answer the survey; otherwise, we may be biasing the conclusions. The statistical procedures are only valid if the samples are randomly or probabilistically selected.

In clinical trials, by randomizing subjects to the treatment groups, treatment arms are distributed evenly and unbiasedly among the subjects. The experimental data cannot be used for statistical inferences if the subjects are not randomly assigned to treatments. One may also use stratification based on demographic variables or prognostic factors and randomly assign the treatments to the subjects in each strata. By stratifying one creates homogeneous blocks in order to compare treatment arms unbiasedly without being influenced by other characteristics of the subjects. However, one needs to keep the number of strata to a minimum. Many strata may result in having incomplete blocks (too few subjects in the blocks) and the study at the end may not be balanced for the key factors of interest, like the treatment group.

The easiest form of randomization includes rolling a die or looking at the table of random numbers and assigning treatment A to the odd numbers and treatment B to the even numbers to get the required number of subjects to each of the two treatments. The problem with this method is that it can lead to imbalances in the treatment groups. One advantage is that it is less likely for anyone to guess the treatment assignment. However, when one starts to add stratification variables, this randomization scheme may be inappropriate.

Block randomization is used when subjects are randomized into blocks. If a block has been completed, then the treatment arms will be completely balanced. The smaller the block size, the more likely the balancing among the treatment groups but easier for one to guess the treatment assignment; however, the larger the block size, the more likely the imbalance among the treatment groups but harder for one to guess the treatment assignment. One has to find a balance between these two scenarios.

Suppose a researcher wants to randomly assign three treatments at two sites, using age as a stratum (divided into two groups (>75, ≤75)), with 6 blocks of size 3 in each site and age classification.

From SAS, one can use PROC PLAN with which one can create a simple random list:

```
Data a;
proc plan seed=1; *test seed for dummy data;
factors site=2 ordered age=2 ordered block=6 ordered
trtgrp=3;  run;
```

This will provide the following output:

```
          The PLAN Procedure

Factor     Select    Levels     Order

Site         2         2       Ordered
Age          2         2       Ordered
Block        6         6       Ordered
trtgrp       3         3       Random

Site   Age     Block            -trtgrp-

1       1        1         1        3        2
                 2         1        3        2
                 3         2        3        1
                 4         1        3        2
                 5         3        2        1
                 6         1        2        3
        2        1         3        2        1
                 2         2        3        1
                 3         2        3        1
                 4         2        3        1
                 5         3        1        2
```

		6	3	1	2
2	1	1	2	1	3
		2	1	2	3
		3	1	3	2
		4	2	1	3
		5	2	1	3
		6	2	1	3
	2	1	2	1	3
		2	3	2	1
		3	2	1	3
		4	3	1	2
		5	3	1	2
		6	1	2	3

The above output can be explained as follows. In each site and age group, the subjects are blocked in sets of three people and they are given the indicated treatments. For example, the recruited 10, 11, and 12 subjects at Site 1 age group 1 are assigned treatments 1, 3, and 2, respectively based on block 4.

For a detailed discussion on randomization in clinical trials, refer to Efron (1971), Pocock and Simon (1975), Kalish and Begg (1985), Lachin et al. (1988), and Friedman et al. (1998).

1.5 Estimation and Testing Hypothesis

In scientific or social investigations, the major roles of statistics are as follows:

1. Estimate and test hypotheses about the parameters of interest based on sample data.
2. Find an appropriate model to estimate a dependent variable based on one or several independent variables, test the adequacy of the model, estimate the parameters involved in the model, and test hypotheses about the parameters.
3. Analyze the experimental data using appropriate techniques and determine the efficacy of the treatments.

While estimating parameters, we would like to see some desirable properties in the estimator like *unbiasedness*, *consistency*, and *minimum variance*. An estimator is unbiased if the average of the estimator, in repeated sampling or experimentation, is the parameter. In a single study the unbiased estimator need not be the parameter, but in repeated studies the average value of that estimator is the parameter. A consistent estimator for large samples approaches the parameter. We also want the estimator not to give widely

spread-out values in repeated studies. This property is the minimum variance of the estimator. The population mean and proportion are estimated in random sampling by the sample mean and sample proportion. These estimators possess the above properties. Usually, the estimates given by computer packages are based on the estimators with desirable properties. Note that the estimator gives the estimate based on the available data.

The estimate may be given as a single value called *point estimate*. Often, one likes to see an interval of possible values of the parameter, and such an interval is called a *confidence interval* (CI). The CI is associated with *confidence probability* and it can be interpreted as follows. Based on the CI formula, the interval can be calculated from a trial. When many trials are made, out of all those intervals, the actual parameter will be contained in (confidence probability) $\times$ 100% cases.

To test a hypothesis, we form *null* (H_0) and *alternative* (H_A) hypotheses, form a *test statistic* based on the observed data and hypothesis H_0, and find the chance (probability, *p-value*) of getting the present evidence or stronger evidence against H_0. If this chance is small, the evidence is improbable and as we are observing that evidence in the current data, we reject H_0. The test always aims to show that the current data are unlikely to be observed if H_0 is true.

The objective is not to establish H_0 from the data, but to find evidence against H_0 from the data, if possible. In view of this, all claims an investigator wants to make must be taken as alternative hypotheses and their complements are taken as null. If a test rejects (retains) H_0, it is not implied that H_0 is false (true). It is a decision the investigator makes based on the observed data. The truth of H_0 and the decision to reject or retain H_0 can be shown in the following table:

	Actually, H_0	
Decide, H_0	**True**	**False**
True	*Correct decision*	*Type II error*
False	*Type I error*	*Correct decision*

Ideally, the decision should be free of error and it is impossible to make an error-free decision. The decision to reject H_0 is devised controlling the probability of commiting Type I error at a prespecified value and minimizing the probability of committing Type II error. The prespecified probability of committing Type I error is called the *level of significance* and is denoted by α. If a decision to reject a hypothesis H_0 is made, either H_0 is false or the decision is one of those 100α% cases of rejection of H_0 when it is actually true. The *power* of the test is one minus the probability of commiting Type II error (β) and the sample size is usually determined to give a power of at least 80%. The commonly used test statistics are Z, t, χ^2, and F. The normal distribution is specified by its mean and standard deviation, and Z is the standardized normal distribution with mean zero and standard deviation one. The t and χ^2