Practical Multivariate Analysis
Sixth Edition

Abdelmonem Afifi, Susanne May, Robin A. Donatello, Virginia A. Clark

Contents

Preface

The first edition of this book appeared in 1984 under the title “Computer Aided Multivariate Analysis.” The title was chosen in order to distinguish it from other books that were more theoretically oriented. By the time we published the fifth edition in 2012, it was impossible to think of a book on multivariate analysis for scientists and applied researchers that is not computer oriented. We therefore decided at that time to change the title to Practical Multivariate Analysis to better characterize the nature of the book. Today, we are pleased to present the sixth edition.

We wrote this book for investigators, specifically behavioral scientists, biomedical scientists, and industrial or academic researchers, who wish to perform **multivariate statistical analyses** and understand the results. We expect the readers to be able to perform and understand the results, but also expect them to know when to ask for help from an expert on the subject. The book can either be used as a self-guided textbook or as a text in an applied course in multivariate analysis. In addition, we believe that the book can be helpful to many statisticians who have been trained in conventional mathematical statistics who are now working as statistical consultants and need to explain multivariate statistical concepts to clients with a limited background in mathematics.

We do not present mathematical derivations of the techniques; rather we rely on geometric and graphical arguments and on examples to illustrate them. The mathematical level has been deliberately kept low. While the derivations of the techniques are referenced, we concentrate on applications to real-life problems, which we feel are the ‘fun’ part of multivariate analysis. To this end, we assume that the reader will use a packaged software program to perform the analysis. We discuss specifically how each of four popular and comprehensive software packages can be used for this purpose. These packages are R, SAS, SPSS, and STATA. The book can be used, however, in conjunction with all other software packages since our presentation explains the output of most standard statistical programs.

We assume that the reader has taken a basic course in statistics that includes tests of hypotheses and covers one-way analysis of variance.

Approach of this book

We wrote the book in a modular fashion. Part One, consisting of six chapters, provides examples of studies requiring multivariate analysis techniques, discusses characterizing data for analysis, computer programs, data entry, data management, data clean-up, missing values, and transformations. It also includes a new chapter on graphics and data visualization and presents a rough guide to assist in the choice of an appropriate multivariate analysis. We included these topics since many investigators have more difficulty with these preliminary steps than with running the multivariate analyses themselves. Also, if these steps are not done with care, the results of the statistical analysis can be faulty.

In the rest of the chapters, we follow a standard format. The first four sections of each chapter include a discussion of when the technique is used, a data example, and the basic assumptions and concepts of the technique. In subsequent sections, we present more detailed aspects of the analysis. At the end of each chapter, we give a summary table showing which features are available in the four software packages. We also include a section entitled ‘What to watch out for’ to warn the reader about common problems related to data analysis. In those sections, we rely on our own experiences in consulting and those detailed in the literature to supplement the formal treatment of the subject.

Part Two covers regression analysis. Chapter 7 deals with simple linear regression and is included for review purposes to introduce our notation and to provide a more complete discussion of outliers and diagnostics than is found in some elementary texts. Chapters 8-10 are concerned with multiple linear regression. Multiple linear regression is used very heavily in practice and provides the foundation for understanding many concepts relating to residual analysis, transformations, choice of variables, missing values, dummy variables, and multicollinearity. Since these concepts are essential to a good grasp of multivariate analysis, we thought it useful to include these chapters in the book.

Chapters 11-18 might be considered the heart of multivariate analysis. They include chapters on discriminant analysis, logistic regression analysis, survival analysis, principal components analysis, factor analysis, cluster analysis, log-linear analysis and correlated outcomes regression. The multivariate analyses have been discussed more as separate techniques than as special cases of some general framework. The advantage of this approach is that it allows us to concentrate on explaining how to analyze a certain type of data from readily available computer programs to answer realistic questions. It also enables the reader to approach each chapter independently. We did include interspersed discussions of how the different analyses relate to each other in an effort to describe the ‘big picture’ of multivariate analysis.

How to use the book

We have received many helpful suggestions from instructors and reviewers on how to order these chapters for reading or teaching purposes. For example, one instructor uses the following order in teaching: principal components, factor analysis, and then cluster analysis, . Another prefers presenting a detailed treatment of multiple regression followed by logistic regression and survival analysis. Instructors and self-learning readers have a wide choice of other orderings of the material because the chapters are largely self contained.

What’s new in the Sixth Edition

During the nearly thirty six years since we wrote the first edition of this book, tremendous advances have taken place in the field of computing and software development. These advances have made it possible to quickly perform any of the multivariate analyses that were available only in theory at that time. They also spurred the invention of new multivariate analyses as well as new options for many of the standard methods. In this edition, we have taken advantage of these developments and made many changes as described below.

For each of the techniques discussed, we used the most recent software versions available and discussed the most modern ways of performing the analysis. In each chapter, we updated the references to today’s literature (while still including the fundamental original references). In terms of statistical software, we discontinued description of S-Plus because of the more wide-spread use of the similar package R. Also, we no longer include Statistica since it is largely not used by our intended readers.

In addition to the above-described modifications, we included comments to distinguish between exploratory and confirmatory analyses in Chapter 1 and throughout the book. We also expanded the discussion of missing values in Chapter 3 and added a discussion of literate programming and reproducible research.

As mentioned above, we added a new chapter (Chapter 4) on graphics and data visualization. In Chapter 9, we updated our discussion of variable selection and added a description of Lasso, a more recent method than the ones already included. In Chapter 10, we added a description of MICE, a multiple imputation approach for dealing with missing values. In Chapter 18, we added a description of the generalized estimating equations (GEE) method for handling correlated data and compared it to the mixed model approach. Finally, in each chapter we updated and/or expanded the summary table of the options available in the four statistical packages to make it consistent with the most recent software versions.

Data sets used for examples and problems are described throughout the book as needed and summarized in Appendix A. Two web sites are also available. The first one is the CRC web site: `http://www.crcpress.com/product/isbn/9781138702226`. From this site, you can download all the data sets used in the book by clicking on the Downloads/Updates tab. The other web site that is available to all readers is: `https://stats.idre.ucla.edu/other/examples/pma6`. This site, developed by the UCLA Institute for Digital Research and Education (IDRE), includes the data sets in the formats of various statistical software packages available in the links included in the Appendix A part of the table of contents in that web page. It also includes illustrations of examples in most chapters, complete with code for three of the four software packages used in the book. Please note that the current site is done for the 5th edition and it is hoped that it will be updated for the sixth edition. We encourage readers to obtain data from either web site and frequently refer to the solutions given in the UCLA web site for practice.

Acknowledgements

We would like to express our appreciation to our colleagues and former students and staff that helped us over the years, both in the planning and preparation of the various editions. These include our colleagues Drs. Carol Aneshensel, Roger Detels, Robert Elashoff, Ralph Frerichs, Mary Ann Hill, and Roberta Madison. Our former students include Drs. Stella Grosser, Luohua Jiang, Jack Lee, Steven Lewis, Tim Morgan, Leanne Streja and David Zhang. Our former staff includes Ms. Dorothy Breininger, Jackie Champion, and Anne Eiseman. In addition, we would like to thank Ms. Meike Jantzen, and Mr. Jack Fogliasso for their help with the references and typesetting.

We also thank Rob Calver and Lara Spieker from CRC Press for their very capable assistance in the preparation of the sixth edition.

We especially appreciate the efforts of the staff of UCLA Institute for Digital Research and Education in putting together the UCLA web site of examples from the book (referenced above).

Our deep gratitude goes to our spouses Marianne Afifi, Bruce Jacobson, Ian Donatello and Welden Clark for their patience and encouragement throughout the stages of conception, writing, and production of the book. Special thanks go to Welden Clark for his expert assistance and troubleshooting of earlier electronic versions of the manuscript.

Abdelmonem Afifi
Susanne May
Robin Donatello
Virginia A. Clark

Authors' Biographies

Abdelmonem Afifi, Ph.D., has been Professor of Biostatistics in the School of Public Health, University of California, Los Angeles (UCLA) since 1965, and served as the Dean of the School from 1985 until 2000. His research includes multivariate and multilevel data analysis, handling missing observations in regression and discriminant analyses, meta-analysis, and model selection. Over the years, he taught well-attended courses in biostatistics for Public Health students and clinical research physicians, and doctoral-level courses in multivariate statistics and multilevel modeling. He has authored many publications in statistics and health related fields, including two widely used books (with multiple editions) on multivariate analysis. He received several prestigious awards for excellence in teaching and research.

Susanne May, Ph.D., is a Professor in the Department of Biostatistics at the University of Washington in Seattle. Her areas of expertise and interest include clinical trials, survival analysis, and longitudinal data analysis. She has more than 20 years of experience as a statistical collaborator and consultant on health related research projects. In addition to a number of methodological and applied publications, she is a coauthor (with Drs. Hosmer and Lemeshow) of *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Dr. May has taught courses on introductory statistics, clinical trials, and survival analysis.

Robin A. Donatello, Dr. P.H., is an Associate Professor in the Department of Mathematics and Statistics and the Developer of the Data Science Initiative at California State University, Chico. Her areas of interest include applied research in the Public Health and Natural Science fields. She has expertise in data visualization, techniques to address missing and erroneous data, implementing reproducible research workflows, computational statistics and Data Science. Dr. Donatello teaches undergraduate and graduate level courses in statistical programming, applied statistics, and data science.

Virginia A. Clark, Ph. D., was professor emerita of Biostatistics and Biomathematics at UCLA. For 27 years, she taught courses in multivariate analysis and survival analysis, among others. In addition to this book, she is coauthor of four books on survival analysis, linear models and analysis of variance, and survey research as well as an introductory book on biostatistics. She published extensively in statistical and health science journals.

Part I

Preparation for Analysis

Chapter 1

What is multivariate analysis?

1.1 Defining multivariate analysis

The expression **multivariate analysis** is used to describe analyses of data that are multivariate in the sense that numerous observations or variables are obtained for each individual or unit studied. In a typical survey 30 to 100 questions are asked of each respondent. In describing the financial status of a company, an investor may wish to examine five to ten measures of the company's performance. Commonly, the answers to some of these measures are interrelated. The challenge of disentangling complicated interrelationships among various measures on the same individual or unit and of interpreting these results is what makes multivariate analysis a rewarding activity for the investigator. Often results are obtained that could not be attained without multivariate analysis.

In the next section of this chapter several studies are described in which the use of multivariate analysis is essential to understanding the underlying problem. Section 1.3 provide a rational for making a distinction between confirmatory and exploratory analyses. Section 1.4 gives a listing and a very brief description of the multivariate analysis techniques discussed in this book. Section 1.5 then outlines the organization of the book.

1.2 Examples of multivariate analyses

The studies described in the following subsections illustrate various multivariate analysis techniques. These are used later in the book as examples.

Depression study example

The data for the depression study have been obtained from a complex, random, multiethnic sample of 1000 adult residents of Los Angeles County. The study was a **panel** or **longitudinal** design where the same respondents were interviewed four times between May 1979 and July 1980. About three-fourths of the respondents were re-interviewed for all four interviews. The field work for the survey was conducted by professional interviewers from the Institute for Social Science Research at the University of California in Los Angeles.

This research is an epidemiological study of depression and help-seeking behavior among free-living (noninstitutionalized) adults. The major objectives are to provide estimates of the prevalence and incidence of depression and to identify causal factors and outcomes associated with this condition. The factors examined include demographic variables, life events stressors, physical health status, health care use, medication use, lifestyle, and social support networks. The major instrument used for classifying depression is the Depression Index (CESD) of the National Institute of Mental Health, Center of Epidemiological Studies. A discussion of this index and the resulting prevalence of depression in this sample is given in Frerichs et al. (1981).

The longitudinal design of the study offers advantages for assessing causal priorities since the time sequence allows us to rule out certain potential causal links. Nonexperimental data of this type cannot directly be used to establish causal relationships, but models based on an explicit theoretical framework can be tested to determine if they are consistent with the data. An example of such model testing is given in Aneshensel and Frerichs (1982).

Data from the first time period of the depression study are described in Chapter 3. Only a subset of the factors measured on a subsample of the respondents is included in this book's web site in order to keep the data set easily comprehensible. These data are used several times in subsequent chapters to illustrate some of the multivariate techniques presented in this book.

Parental HIV study

The data from the parental HIV study have been obtained from a clinical trial to evaluate an intervention given to increase coping skills (Rotheram-Borus et al., 2001). The purpose of the intervention was to improve behavioral, social, and health outcomes for parents with HIV/AIDS and their children. Parents and their adolescent children were recruited from the New York City Division of Aids Services (DAS). Adolescents were eligible for the study if they were between the ages of 11 and 18 and if the parents and adolescents had given informed consent. Individual interviews were conducted every three months for the first two years and every six months thereafter. Information obtained in the interviews included background characteristics, sexual behavior, alcohol and drug use, medical and reproductive history, and a number of psychological scales.

A subset of the data from the study is available on this book's web site. To protect the identity of the participating adolescents we used the following procedures. We randomly chose one adolescent per family. In addition, we reduced the sample further by choosing a random subset of the original sample. Adolescent case numbers were assigned randomly without regard to the original order or any other numbers in the original data set.

Data from the baseline assessment will be used for problems as well as to illustrate various multivariate analysis techniques.

Northridge earthquake study

On the morning of January 17, 1994 a magnitude 6.7 earthquake centered in Northridge, CA awoke Los Angeles and Ventura County residents. Between August 1994 and May 1996, 1830 residents were interviewed about what happened to them in the earthquake. The study uses a telephone survey lasting approximately 48 minutes to assess the residents' experiences in and responses to the Northridge earthquake. Data from 506 residents are included in the data set posted on the book web site, and described in Appendix A.

Subjects were asked where they were, how they reacted, where they obtained information, whether their property was damaged or whether they experienced injury, and what agencies they were in contact with. The questionnaire included the Brief Symptom Inventory (BSI), a measure of psychological functioning used in community studies, and questions on emotional distress. Subjects were also asked about the impact of the damage to the transportation system as a result of the earthquake. Investigators not only wanted to learn about the experiences of the Southern California residents in the Northridge earthquake, but also wished to compare their findings to similar studies of the Los Angeles residents surveyed after the Whittier Narrows earthquake on October 1, 1987, and Bay Area residents interviewed after the Loma Prieta earthquake on October 17, 1989.

The Northridge earthquake data set is used in problems at the end of several chapters of the book to illustrate a number of multivariate techniques. Multivariate analyses of these data include, for example, exploring pre- and post-earthquake preparedness activities as well as taking into account several factors relating to the subject and the property (Nguyen et al., 2006).

Bank loan study

The managers of a bank need some way to improve their prediction of which borrowers will successfully pay back a type of bank loan. They have data from the past on the characteristics of persons to whom the bank has lent money and the subsequent record of how well the person has repaid the loan. Loan payers can be classified into several types: those who met all of the terms of the loan, those who eventually repaid the loan but often did not meet deadlines, and those who simply defaulted. They also have information on age, sex, income, other indebtedness, length of residence, type of residence, family size, occupation, and the reason for the loan. The question is, can a simple rating system be devised that will help the bank personnel improve their prediction rate and lessen the time it takes to approve loans? The methods described in Chapter 12 and Chapter 13 can be used to answer this question.

Lung function study

The purpose of this lung function study of chronic respiratory disease is to determine the effects of various types of smog on lung function of children and adults in the Los Angeles area. Because they could not randomly assign people to live in areas that had different levels of pollutants, the investigators were very concerned about the interaction that might exist between the locations where persons chose to live and their values on various lung function tests. The investigators picked four areas of quite different types of air pollution and measured various demographic and other responses on all persons over seven years old who live there. These areas were chosen so that they are close to an air-monitoring station.

The researchers took measurements at two points in time and used the change in lung function over time as well as the levels at the two periods as outcome measures to assess the effects of air pollution. The investigators had to do the lung function tests by using a mobile unit in the field, and much effort went into problems of validating the accuracy of the field observations. A discussion of the particular lung function measurements used for one of the four areas can be found in Detels et al. (1975). In the analysis of the data, adjustments must be made for sex, age, height, and smoking status of each person.

Over 15,000 respondents have been examined and interviewed in this study. The data set is being used to answer numerous questions concerning effects of air pollution, smoking, occupation, etc. on different lung function measurements. For example, since the investigators obtained measurements on all family members seven years old and older, it is possible to assess the effects of having parents who smoke on the lung function of their children (Tashkin et al., 1984). Studies of this type require multivariate analyses so that investigators can arrive at plausible scientific conclusions that could explain the resulting lung function levels.

This data set is described in Appendix A. Lung function and associated data for nonsmoking families for the father, mother, and up to three children ages 7–17 are available from the book's web site.

School data set

The school data set is a publicly available data set that is provided by the National Center for Educational Statistics. The data come from the National Education Longitudinal Study of 1988 (called NELS:88). The study collected data beginning with 8th graders and conducted initial interviews and four follow-up interviews which were performed every other year. The data used here contain only initial interview data. They represent a random subsample of 23 schools with 519 students out of more than a thousand schools with almost twenty five thousand students. Extensive documentation of all aspects of the study is available at the following web site: `http://nces.ed.gov/surveys/NELS88/`. The longitudinal component of NELS:88 has been used to investigate change in students' lives and school-related and other outcomes. The focus on the initial interview data provides the opportunity to examine associations between school and student-related factors and students' academic performance in a cross-sectional manner. This type of analysis will be illustrated in Chapter 18.

1.3 Exploratory versus confirmatory analyses

A crucial component for most research studies and analyses is the testing of hypotheses. For some types of studies, hypotheses are specified in detail prior to study start (a priori) and then remain unchanged. This is typically the case, e.g., for clinical trials and other designed experiments. For other types of studies, some hypotheses might be specified in advance while others are generated only after study start and potentially after reviewing some or all of the study data. This is often the case for observational studies. In this section, we make a distinction between two conceptually different approaches to analysis and reporting based on whether the primary goal of a study is to *confirm* prespecified hypotheses or to *explore* hypotheses that have not been prespecified.

The following is a motivating example provided by Fleming (2010). He describes an experience where he walked into a **maternity ward** (when they still had such) while visiting a friend who had just given birth. He noticed that there were 22 babies, but only 2 of one gender while the other 20 were of the other gender. As a statistician, he dutifully calculated the p-value for the likelihood of seeing such (or worse) imbalance if in truth there are 50% of each. The two-sided p-value turns out to be 0.0001, indicating a very small likelihood (1 in 10,000) of such or more extreme imbalance to be observed if in truth there are 50% of each. This is an example of where the hypothesis was generated after seeing the data. We will call such hypotheses *exploratory*.

Following Fleming, researchers might want to go out and test an exploratory hypothesis in another setting or with new data. In the example above, one might want to go to another maternity ward to collect further evidence of a strong imbalance in gender distribution at birth. Imagine that in a second (*confirmatory*) maternity ward there might be exactly equal numbers for each gender (e.g. 11 boys and 11 girls). Testing the same hypothesis in this setting will not yield any statistically significant difference from the presumed 50%. Nevertheless, one might be tempted to simply combine the two studies. A corresponding two-sided p-value remains statistically significant (p-value < 0.01).

The above example might appear silly, because few researchers will believe that the distribution of gender at birth (without human interference) is very different from 50%. Nevertheless, there are many published research articles which test and present the results for hypotheses that were generated by looking at data and noticing ‘unusual’ results. Without a clear distinction between whether hypotheses were specified a priori or not, it is difficult to interpret the p-values provided.

Results from confirmatory analyses provide much stronger evidence than results from exploratory analyses. Accordingly, interpretation of results from confirmatory analyses can be stated using much stronger language than interpretation of results from exploratory analyses. Furthermore, results from exploratory analyses should not be combined with results from confirmatory analyses (e.g. in meta analyses), because the **random high bias** (Fleming, 2010) will remain (albeit attenuated). To avoid random high bias when combining data or estimates from multiple studies only data/estimates from confirmatory analyses should be combined. However, this requires clear identification of whether confirmatory or exploratory analysis were performed for each individual study and/or analysis.

Many authors have pointed out that the medical literature is replete with studies that cannot be reproduced (Breslow, 1999; Munafò et al., 2017). As argued by Breslow (1999), **reproducibility** of studies, and in particular epidemiologic studies, can be improved if hypotheses are specified a priori and the nature of the study (exploratory versus confirmatory) is clearly specified.

Throughout this book, we distinguish between the two approaches to multivariate analyses and presentations of results and provide examples for each.

1.4 Multivariate analyses discussed in this book

In this section a brief description of the major multivariate techniques covered in this book is presented. To keep the statistical vocabulary to a minimum, we illustrate the descriptions by examples.

Simple linear regression

A nutritionist wishes to study the effects of early calcium intake on the bone density of postmenopausal women. She can measure the bone density of the arm (radial bone), in grams per square centimeter, by using a noninvasive device. Women who are at risk of hip fractures because of too low a bone density will tend to show low arm bone density also. The nutritionist intends to sample a group of elderly churchgoing women. For women over 65 years of age, she will plot calcium intake as a teenager (obtained by asking the women about their consumption of high-calcium foods during their teens) on the horizontal axis and arm bone density (measured) on the vertical axis. She expects the radial bone density to be lower in women who had a lower calcium intake. The nutritionist plans to fit a simple linear regression equation and test whether the slope of the regression line is zero. In this example a single outcome factor is being predicted by a single predictor factor.

Simple linear regression as used in this case would not be considered multivariate by some statisticians, but it is included in this book to introduce the topic of multiple regression.

Multiple linear regression

A manager is interested in determining which factors predict the dollar value of sales of the firm's personal computers. Aggregate data on population size, income, educational level, proportion of population living in metropolitan areas, etc. have been collected for 30 areas. As a first step, a multiple linear regression equation is computed, where dollar sales is the outcome variable and the other factors are considered as candidates for predictor variables. A linear combination of the predictors is used to predict the outcome or response variable.

Discriminant function analysis

A large sample of initially disease-free men over 50 years of age from a community has been followed to see who subsequently has a diagnosed heart attack. At the initial visit, blood was drawn from each man, and numerous other determinations were made, including body mass index, serum cholesterol, phospholipids, and blood glucose. The investigator would like to determine a linear function of these and possibly other measurements that would be useful in predicting who would and who would not get a heart attack within ten years. That is, the investigator wishes to derive a classification (discriminant) function that would help determine whether or not a middle-aged man is likely to have a heart attack.

Logistic regression

An online movie streaming service has classified movies into two distinct groups according to whether they have a high or low proportion of the viewing audience when shown. The company also records data on features such as the length of the movie, the genre, and the characteristics of the actors. An analyst would use logistic regression because some of the data do not meet the assumptions for statistical inference used in discriminant function analysis, but they do meet the assumptions for logistic regression. From logistic regression we derive an equation to estimate the probability of capturing a high proportion of the target audience.

Poisson regression

In a health survey, middle school students were asked how many visits they made to the dentist in the last year. The investigators are concerned that many students in this community are not receiving adequate dental care. They want to determine what characterizes how frequently students go to the dentist so that they can design a program to improve utilization of dental care. Visits per year are count data and Poisson regression analysis provides a good tool for analyzing this type of data. Poisson regression is covered in the logistic regression chapter.

Survival analysis

An administrator of a large health maintenance organization (HMO) has collected data for a number of years on length of employment in years for their physicians who are either family practitioners or internists. Some of the physicians are still employed, but many have left. For those still employed, the administrator can only know that their ultimate length of employment will be greater than their current length of employment. The administrator wishes to describe the distribution of length of employment for each type of physician, determine the possible effects of factors such as gender and location of work, and test whether or not the length of employment is the same for two specialties. Survival analysis, or event history analysis (as it is often called by behavioral scientists), can be used to analyze the distribution of time to an event such as quitting work, having a relapse of a disease, or dying of cancer.

Principal components analysis

An investigator has made a number of measurements of lung function on a sample of adult males who do not smoke. In these tests each man is told to inhale deeply and then blow out as fast and as much as possible into a spirometer, which makes a trace of the volume of air expired over time. The maximum or forced vital capacity (FVC) is measured as the difference between maximum inspiration and maximum expiration. Also, the amount of air expired in the first second (FEV1), the forced mid-expiratory flow rate (FEF 25–75), the maximal expiratory flow rate at 50% of forced vital capacity (V50), and other measures of lung function are calculated from this trace. Since all these measures are made from the same flow–volume curve for each man, they are highly interrelated. From past experience it is known that some of these measures are more interrelated than others and that they measure airway resistance in different sections of the airway.

The investigator performs a principal components analysis to determine whether a new set of measurements called principal components can be obtained. These principal components will be linear functions of the original lung function measurements and will be uncorrelated with each other. It is hoped that the first two or three principal components will explain most of the variation in the original lung function measurements among the men. Also, it is anticipated that some operational meaning can be attached to these linear functions that will aid in their interpretation. The investigator may decide to do future analyses on these uncorrelated principal components rather than on the original data. One advantage of this method is that often fewer principal components are needed than original variables. Also, since the principal components are uncorrelated, future computations and explanations can be simplified.

Factor analysis

An investigator has asked each respondent in a survey whether he or she strongly agrees, agrees, is undecided, disagrees, or strongly disagrees with 15 statements concerning attitudes toward inflation. As a first step, the investigator will do a factor analysis on the resulting data to determine which statements belong together in sets that are uncorrelated with other sets. The particular statements that form a single set will be examined to obtain a better understanding of attitudes toward inflation. Scores derived from each set or factor will be used in subsequent analyses to predict consumer spending.

Cluster analysis

Investigators have made numerous measurements on a sample of patients who have been classified as being depressed. They wish to determine, on the basis of their measurements, whether these patients can be classified by type of depression. That is, is it possible to determine distinct types of depressed patients by performing a cluster analysis on patient scores on various tests?

Unlike the investigator studying men who do or do not get heart attacks, these investigators do not possess a set of individuals whose type of depression can be known before the analysis is performed. Nevertheless, the investigators want to separate the patients into unique groups and to examine the resulting groups to see whether distinct types do exist and, if so, what their characteristics are.

Log-linear analysis

An epidemiologist in a medical study wishes to examine the interrelationships among the use of substances that are thought to be risk factors for disease. These include four risk factors where the answers have been summarized into categories. The risk factors are smoking tobacco (yes at present, former smoker, never smoked), drinking (yes, no), marijuana use (yes, no), and other illicit drug use (yes, no). Previous studies have shown that people who drink are more apt than nondrinkers to smoke cigarettes, but the investigator wants to study the associations among the use of these four substances simultaneously.

Correlated outcomes regression

A health services researcher is interested in determining the hospital-related costs of appendectomy, the surgical removal of the appendix. Data are available for a number of patients in each of several hospitals. Such a sample is called a **clustered sample** since patients are clustered within hospitals. For each operation, the information includes the costs as well as the patient's age, gender, health status and other characteristics. Information is also available on the hospital, such as its number of beds, location and staff size. A multiple linear regression equation is computed, where cost is the outcome variable and the other factors are considered as candidates for predictor variables. As in multiple linear regression, a linear combination of the predictors is used to predict the outcome or response variable. However, adjustments to the analysis must be made to account for the clustered nature of the sample, namely the possibility that patients within any one hospital may be more similar to each other than to patients in other hospitals. Since the outcomes within a given hospital are correlated, the researcher plans to use correlated outcomes regression to analyze the data.

1.5 Organization and content of the book

This book is organized into two major parts. Part One (Chapters 1–6) deals with data entry, preparation, visualization, screening, missing values, transformations, and decisions about likely choices for analysis. Part Two (Chapters 7–18) deals with regression analysis.

Chapters 2–6 are concerned with data preparation and the choice of what analysis to use. First, **variables** and how they are classified are discussed in Chapter 2. The next chapter concentrates on the practical problems of getting data into the computer, handling nonresponse, data management, getting rid of erroneous values, and preparing a useful codebook. Visualization techniques are discussed in Chapter 4. The next chapter deals with checking assumptions of normality and independence. The features of computer software packages used in this book are discussed. The choice of appropriate statistical analyses is discussed in Chapter 6.

Readers who are familiar with handling data sets on computers could skip some of these initial chapters and go directly to Chapter 7. However, formal course work in statistics often leaves an investigator unprepared for the complications and difficulties involved in real data sets. The material in Chapters 2–6 was deliberately included to fill this gap in preparing investigators for real world data problems.

For a course limited to multivariate analysis, Chapters 2–6 can be omitted if a carefully prepared data set is used for analysis. The depression data set, presented in Chapter 3, has been modified to make it directly usable for multivariate data analysis, but the user may wish to subtract one from the variables 2, 31, 33, and 34 to change the values to zeros and ones. Also, the lung function data, the lung cancer data, and the parental HIV data are briefly described in Appendix A. These data, along with the data in Table 9.1 and Table 16.1, are available on the web from the publisher. See Appendix A or the preface for the exact web site address.

In Chapters 7–18 we follow a standard format. The topics discussed in each chapter are given, followed by a discussion of when the techniques are used. Then the basic concepts and formulas are explained. Further interpretation, and data examples with topics chosen that relate directly to the techniques, follow. Finally, a summary of the available computer output that may be obtained from four statistical software packages is presented. We conclude each chapter with a discussion of pitfalls to avoid and alternatives to consider when performing the analyses described.

As much as possible, we have tried to make each chapter self-contained. However, Chapters 11 and 12, on discriminant analysis and logistic regression, are somewhat interrelated, as are Chapters 14 and 15, covering principal components and factor analysis.

References for further information on each topic are given in each chapter. Most of the references do require more mathematics than this book, but special emphasis can be placed on references that include examples. If you wish primarily to learn the concepts involved in multivariate techniques and are not as interested in performing the analysis, then a conceptual introduction to multivariate analysis can be found in Kachigan (1991). Everitt and Dunn (2001) provide a highly readable introduction also. For a concise description of multivariate analysis see Manly (2016).

We believe that the best way to learn multivariate analysis is to do it on data that you are familiar with. No book can illustrate all the features found in computer output for a real-life data set. Learning multivariate analysis is similar to learning to swim: you can go to lectures, but the real learning occurs when you get into the water.

Chapter 2

Characterizing data for analysis

2.1 Variables: their definition, classification, and use

In performing multivariate analysis, the investigator deals with numerous variables. In this chapter, we define what a variable is in Section 2.2. Section 2.3 presents a method of classifying variables that is sometimes useful in multivariate analysis since it allows one to check that a commonly used analysis has not been missed. Section 2.4 explains how variables are used in analysis and gives the common terminology for distinguishing between the two major uses of variables. Section 2.5 includes some examples of classifying variables and Section 2.6 discusses other characteristics of data and references exploratory data analysis.

2.2 Defining statistical variables

The word **variable** is used in statistically oriented literature to indicate a characteristic or property that is possible to measure. When we measure something, we make a numerical model of the thing being measured. We follow some rule for assigning a number to each level of the particular characteristic being measured. For example, the height of a person is a variable. We assign a numerical value to correspond to each person's height. Two people who are equally tall are assigned the same numeric value. On the other hand, two people of different heights are assigned two different values. Measurements of a variable gain their meaning from the fact that there exists unique correspondence between the assigned numbers and the levels of the property being measured. Thus two people with different assigned heights are not equally tall. Conversely, if a variable has the same assigned value for all individuals in a group, then this variable does not convey useful information to differentiate individuals in the group.

Physical measurements, such as height and weight, can be measured directly by using physical instruments. On the other hand, properties such as reasoning ability or the state of depression of a person must be measured indirectly. We might choose a particular intelligence test and define the variable "intelligence" to be the score achieved on this test. Similarly, we may define the variable "depression" as the number of positive responses to a series of questions. Although what we wish to measure is the degree of depression, we end up with a count of yes answers to some questions. These examples point out a fundamental difference between direct physical measurements and abstract variables.

Often the question of how to measure a certain property can be perplexing. For example, if the property we wish to measure is the cost of keeping the air clean in a particular area, we may be able to come up with a reasonable estimate, although different analysts may produce different estimates. The problem becomes much more difficult if we wish to estimate the benefits of clean air.

On any given individual or thing we may measure several different characteristics. We would then be dealing with several variables, such as age, height, annual income, race, sex, and level of depression of a certain individual. Similarly, we can measure characteristics of a corporation, such as various financial measures. In this book we are concerned with analyzing data sets consisting of measurements on several variables for each individual in a given sample. We use the symbol P to denote the number of variables and the symbol N to denote the number of **individuals, observations, cases**, or **sampling units**.

2.3 Stevens's classification of variables

In the determination of the appropriate statistical analysis for a given set of data, it is useful to classify variables by type. One method for classifying variables is by the degree of sophistication evident in the way they are measured. For example, we can measure the height of people according to whether the top of their head exceeds a mark on the wall; if yes, they are tall; and if no, they are short. On the other hand, we can also measure height in centimeters or inches. The latter technique is a more sophisticated way of measuring height. As a scientific discipline advances, the measurement of the variables used in it tends to become more sophisticated.

Various attempts have been made to formalize variable classification. A commonly accepted system is that proposed by Stevens (1955). In this system, measurements are classified as **nominal, ordinal, interval,** or **ratio**. In deriving his classification, Stevens characterized each of the four types by a transformation that would not change a measurement's classification. In the subsections that follow, rather than discuss the mathematical details of these transformations, we present the practical implications for data analysis.

As with many classification schemes, Stevens's system is useful for some purposes but not for others. It should be used as a general guide to assist in characterizing the data and to make sure that a useful analysis is not overlooked. However, it should not be used as a rigid rule that ignores the purpose of the analysis or limits its scope (Velleman and Wilkinson, 1993).

Nominal variables

With **nominal variables** each observation belongs to one of several distinct categories. The categories are not necessarily numerical, although numbers may be used to represent them. For example, "sex" is a nominal variable. An individual's gender is either male or female. We may use any two symbols, such as M and F, to represent the two categories. In data analysis, numbers are used as the symbols since many computer programs are designed to handle only numerical symbols. Since the categories may be arranged in any desired order, any set of numbers can be used to represent them. For example, we may use 0 and 1 to represent males and females, respectively. We may also use 1 and 2 to avoid confusing zeros with blanks. Any two other numbers can be used as long as they are used consistently.

An investigator may rename the categories, thus performing a numerical operation. In doing so, the investigator must preserve the uniqueness of each category. Stevens expressed this last idea as a "basic empirical operation" that preserves the category to which the observation belongs. For example, two males must have the same value on the variable "sex," regardless of the two numbers chosen for the categories. Table 2.1 summarizes these ideas and presents further examples. Nominal variables with more than two categories, such as race or religion, may present special challenges to the multivariate data analyst. Some ways of dealing with these variables are presented in Chapter 8.

Ordinal variables

Categories are used for **ordinal variables** as well, but there also exists a known order among them. For example, in the Mohs Hardness Scale, minerals and rocks are classified according to ten levels of hardness. The hardest mineral is diamond and the softest is talc (Pough, 1998).

Any ten numbers can be used to represent the categories, as long as they are ordered in magnitude. For instance, the integers 1–10 would be natural to use. On the other hand, any sequence of increasing numbers may also be used. Thus, the basic empirical operation defining ordinal variables is whether one observation is greater than another. For example, we must be able to determine whether one mineral is harder than another. Hardness can be tested easily by noting which mineral

Table 2.1: *Stevens's measurement system*

Type of measurement	Basic empirical operation	Examples
Nominal	Determine equality of categories	Company names Race Religion Soccer players' numbers
Ordinal	Determine greater than or less than (ranking)	Hardness of minerals Socioeconomic status Rankings of wines
Interval	Determine equality of differences between levels	Temperature in degrees Fahrenheit Calendar dates
Ratio	Determine equality of ratios of levels	Height Weight Density Difference in time

can scratch the other. Note that for most ordinal variables there is an underlying continuum being approximated by artificial categories. For example, in the above hardness scale fluorite is defined as having a hardness of 4, and calcite, 3. However, there is a range of hardness between these two numbers not accounted for by the scale.

Often investigators classify people, or ask them to classify themselves, along some continuum (see Luce and Narens, 1987). For example, a physician may classify a patient's disease status as none = 1, mild = 2, moderate = 3, and severe = 4. Clearly, increasing numbers indicate increasing severity, but it is not certain that the difference between not having an illness and having a mild case is the same as between having a mild case and a moderate case. Hence, according to Stevens's classification system, this is an ordinal variable.

Interval variables

An **interval variable** is a variable in which the differences between successive values are always the same. For example, the variable "temperature," in degrees Fahrenheit, is measured on the interval scale since the difference between 12° and 13° is the same as the difference between 13° and 14° or the difference between any two successive temperatures. In contrast, the Mohs Hardness Scale does not satisfy this condition since the intervals between successive categories are not necessarily the same. The scale must satisfy the basic empirical operation of preserving the equality of intervals.

Ratio variables

Ratio variables are interval variables with a natural point representing the origin of measurement, i.e., a natural zero point. For instance, height is a ratio variable since zero height is a naturally defined point on the scale. We may change the unit of measurement (e.g., centimeters to inches), but we would still preserve the zero point and also the ratio of any two values of height. Temperature is not a ratio variable since we may choose the zero point arbitrarily, thus not preserving ratios.

There is an interesting relationship between interval and ratio variables. The difference between two interval variables is a ratio variable. For example, although time of day is measured on the interval scale, the length of a time period is a ratio variable since it has a natural zero point.

Other classifications

Other methods of classifying variables have also been proposed. Many authors use the term **categorical** to refer to nominal and ordinal variables where categories are used.

We mention, in addition, that variables may be classified as discrete or continuous. A variable is called **continuous** if it can take on any value in a specified range. Thus the height of an individual may be 70 or 70.4539 inches. Any numerical value in a certain range is a conceivable height.

A variable that is not continuous is called **discrete**. A discrete variable may take on only certain specified values. For example, counts are discrete variables since only zero or positive integers are allowed. In fact, all nominal and ordinal variables are discrete. Interval and ratio variables can be continuous or discrete. This latter classification carries over to the possible distributions assumed in the analysis. For instance, the normal distribution is often used to describe the distribution of continuous variables.

Statistical analyses have been developed for various types of variables. In Chapter 6 a guide to selecting the appropriate descriptive measures and multivariate analyses will be presented. The choice depends on how the variables are used in the analysis, a topic that is discussed next.

2.4 How variables are used in data analysis

The type of data analysis required in a specific situation is also related to the way in which each variable in the data set is used. Variables may be used to measure outcomes or to explain why a particular outcome resulted. For example, in the treatment of a given disease a specific drug may be used. The **outcome variable** may be a discrete variable classified as “cured” or “not cured.” The outcome variable may depend on several characteristics of the patient such as age, genetic background, and severity of the disease. These characteristics are sometimes called **explanatory** or **predictor variables**. Equivalently, we may call the outcome the **dependent variable** and the characteristics the **independent variable**. The latter terminology is very common in statistical literature. This choice of terminology is unfortunate in that the “independent” variables do not have to be statistically independent of each other. Indeed, these independent variables are usually interrelated in a complex way. Another disadvantage of this terminology is that the common connotation of the words implies a causal model, an assumption not needed for the multivariate analyses described in this book. In spite of these drawbacks, the widespread use of these terms forces us to adopt them.

In other situations the dependent or outcome variable may be treated as a continuous variable. For example, in household survey data we may wish to relate monthly expenditure on cosmetics per household to several explanatory or independent variables such as the number of individuals in the household, their gender, and the household income.

In some situations the roles that the various variables play are not obvious and may also change, depending on the question being addressed. Thus a data set for a certain group of people may contain observations on their sex, age, diet, weight, and blood pressure. In one analysis, we may use weight as a dependent or outcome variable with height, sex, age, and diet as the independent or predictor variables. In another analysis, blood pressure might be the dependent or outcome variable, with weight and other variables considered as independent or predictor variables.

In certain exploratory analyses all the variables may be used as one set with no regard to whether they are dependent or independent. For example, in the social sciences a large number of variables may be defined initially, followed by attempts to combine them into a smaller number of summary variables. In such an analysis the original variables are not classified as dependent or independent. The summary variables may later be used either as outcome or predictor variables. In Chapter 6 multivariate analyses described in this book will be characterized by the situations in which they apply according to the types of variables analyzed and the roles they play in the analysis.

2.5 Examples of classifying variables

In the depression data example several variables are measured on the nominal scale: sex, marital status, employment, and religion. The general health scale is an example of an ordinal variable. Income and age are both ratio variables. No interval variable is included in the data set. A partial listing and a codebook for this data set are given in Chapter 3.

One of the questions that may be addressed in analyzing these data is "Which factors are related to the degree of psychological depression of a person?" The variable "cases" may be used as the dependent or outcome variable since an individual is considered a case if his or her score on the depression scale exceeds a certain level. "Cases" is an ordinal variable, although it can be considered nominal because it has only two categories. The independent or predictor variable could be any or all of the other variables (except ID and measures of depression). Examples of analyses without regard to variable roles are given in Chapters 14 and 15 using the variables C_1 to C_{20} in an attempt to summarize them into a small number of components or factors.

Sometimes, Stevens's classification system is difficult to apply, and two investigators could disagree on a given variable. For example, there may be disagreement about the ordering of the categories of a socioeconomic status variable. Thus the status of blue-collar occupations with respect to the status of certain white-collar occupations might change over time or from culture to culture. So such a variable might be difficult to justify as an ordinal variable, but we would be throwing away valuable information if we used it as a nominal variable. Despite these difficulties, Stevens's system is useful in making decisions on appropriate statistical analysis, as will be discussed in Chapter 6.

2.6 Other characteristics of data

Data are often characterized by whether the measurements are accurately taken and are relatively error free, and by whether they meet the assumptions that were used in deriving statistical tests and confidence intervals. Often, an investigator knows that some of the variables are likely to have observations that have errors. If the effect of an error causes the numerical value of an observation to not be in line with the numerical values of most of the other observations, these extreme values may be called **outliers** and should be considered for removal from the analysis. But other observations may not be accurate and still be within the range of most of the observations. Data sets that contain a sizeable portion of inaccurate data or errors are called "dirty" data sets.

Special statistical methods have been developed that are resistant to the effects of dirty data. Other statistical methods, called robust methods, are insensitive to departures from underlying model assumptions. In this book, we do not present these methods but discuss finding outliers and give methods of determining if the data meet the assumptions. For further information on statistical methods that are well suited for dirty data or require few assumptions, see Hoaglin et al. (2000); Schwaiger and Opitz (2003), or Fox and Long (1990).

2.7 Summary

In this chapter statistical variables were defined. Their types and the roles they play in data analysis were discussed. Stevens's classification system was described. These concepts can affect the choice of analyses to be performed, as will be discussed in Chapter 6.

2.8 Problems

2.1 Classify the following types of data by using Stevens's measurement system: decibels of noise level, father's occupation, parts per million of an impurity in water, density of a piece of bone, rating of a wine by one judge, net profit of a firm, and score on an aptitude test.

2.2 In a survey of users of a walk-in mental health clinic, data have been obtained on sex, age, household roster, race, education level (number of years in school), family income, reason for coming to the clinic, symptoms, and scores on screening examination. The investigator wishes to determine what variables affect whether or not coercion by the family, friends, or a governmental agency was used to get the patient to the clinic. Classify the data according to Stevens's measurement system. What would you consider to be possible independent variables? Dependent variables? Do you expect the dependent variables to be independent of each other?

2.3 For the chronic respiratory study data described in Appendix A, classify each variable according to Stevens's scale and according to whether it is discrete or continuous. Pose two possible research questions and decide on the appropriate dependent and independent variables.

2.4 Repeat problem 2.3 for the lung cancer data set described in Table 13.1.

2.5 From a field of statistical application (perhaps your own field of specialty), describe a data set and repeat the procedures described in Problem 2.3.

2.6 If the RELIG variable described in Table 3.4 of this text was recoded 1 = Catholic, 2 = Protestant, 3 = Jewish, 4 = none, and 5 = other, would this meet the basic empirical operation as defined by Stevens for an ordinal variable?

2.7 Give an example of nominal, ordinal, interval, and ratio variables from a field of application you are familiar with.

2.8 Data that are ordinal are often analyzed by methods that Stevens reserved for interval data. Give reasons why thoughtful investigators often do this.

2.9 The Parental HIV data set described in Appendix A includes the following variables: job status of mother (JOBMO, 1=employed, 2=unemployed, and 3=retired/disabled) and mother's education (EDUMO, 1=did not complete high school, 2=high school diploma/GED, and 3=more than high school). Classify these two variables using Stevens's measurement system.

2.10 Give an example from a field that you are familiar with of an increased sophistication of measuring that has resulted in a measurement that used to be ordinal now being interval.

Chapter 3

Preparing for data analysis

3.1 Processing data so they can be analyzed

Once the data are available from a study there are still a number of steps that must be undertaken to get them into shape for analysis. This is particularly true when multivariate analyses are planned since these analyses are often done on large data sets. In this chapter we provide information on topics related to data processing.

Section 3.2 describes the statistical software packages used in this book. Note that several other statistical packages offer an extensive selection of multivariate analyses. In addition, almost all statistical packages and even some of the spreadsheet programs include at least multiple regression as an option.

The next topic discussed is data entry (Section 3.3). Data collection is often performed using computers directly via Computer Assisted Personal Interviewing (CAPI), Audio Computer Assisted Self Interviewing (ACASI), via the Internet, or via phone apps. For example, SurveyMonkey and Google Forms are free and commercially available programs that facilitate sending and collecting surveys via the Internet. Nonetheless, paper and pencil interviews or mailed questionnaires are still a form of data collection. The methods that need to be used to enter the information obtained from paper and pencil interviews into a computer depend on the size of the data set. For a small data set there are a variety of options since cost and efficiency are not important factors. Also, in that case the data can be easily screened for errors simply by visual inspection. But for large data sets, careful planning of data entry is necessary since costs are an important consideration along with getting a data set for analysis that is as error-free as possible. Here we summarize the data input options available in the statistical software packages used in this book and discuss some important options.

Section 3.4 covers combining and updating data sets. The operations used and the options available in the various packages are described. Initial discussion of missing values, outliers, and transformations is given and the need to save results is stressed.

Section 3.5 discusses methods to conduct research in a reproducible manner and the importance of documenting steps taken during data preparation and analysis in a manner that is human-readable. Finally, in Section 3.6 we introduce a multivariate data set that will be widely used in this book and summarize the data in a codebook.

We want to stress that the procedures discussed in this chapter can be time consuming and frustrating to perform when large data sets are involved. Often the amount of time used for data entry, editing, and screening can far exceed that used on statistical analyses. It is very helpful to either have computer expertise yourself or have access to someone you can get advice from occasionally. Of note, our definition of large data sets includes data sets such as those publicly available from the Centers for Disease Control and Prevention (CDC). More complicated issues arise when data sets that are much larger (in the order of terabytes). These arise, e.g., with genetic data or internet data bases. Such sets do not fall within the scope of this book.

3.2 Choice of a statistical package

There is a wide choice of statistical **software packages** available. Many packages, however, are quite specialized and do not include many of the multivariate analyses given in this book. For example, there are statistical packages that are aimed at particular areas of application or give tests for exact statistics that are more useful for other types of work. In choosing a package for multivariate analysis, we recommend that you consider the statistical analyses listed in Table 6.2 and check whether the package includes them.

In some cases the statistical package is sold as a single unit and in others you purchase a basic package, but you have a choice of additional programs so you can buy what you need. Some programs require yearly license fees, others are free.

Ease of use

Some packages are easier to use than others, although many of us find this difficult to judge–we like what we are familiar with. In general, the packages that are simplest to use have two characteristics. First, they have fewer options to choose from and these options are provided automatically by the program with little need for programming by the user. Second, they use the "point and click" method known as graphical user interface (GUI) for choosing what is done rather than requiring the user to write out statements. However, many current point and click programs do not leave the user with an audit trail of what choices have been made.

On the other hand, software programs with extensive options have obvious advantages. Also, the use of written statements (or *commands*) allows you to have a written record of what you have done. Such a record makes it easier to re-run programs and to facilitate reproducibility. The record of the commands used can be particularly useful in large-scale data analyses that extend over a considerable period of time and involve numerous investigators. Still other programs provide the user with a programming language that allows the users great freedom in what output they can obtain.

Packages used in this book

In this book, we make specific reference to four general-purpose statistical software packages (listed in alphabetical order): R v3.5, SAS v9.4, SPSS v25, and Stata v15

R is a language created for performing statistical analyses and provides rich data visualization capabilities. The user writes the language expressions that are read and immediately executed by the program. This process allows the user to write a function, run it, see what happens, and then use the result of the function in a second function. R is a free and open source program where much of the added functionality comes from external contributed packages that must be installed. The CRAN task views aim to provide some guidance on which packages are relevant for tasks related to a certain topic. It is important to note that R is typically used through an integrated development environment (IDE) program called RStudio (RStudio Team, 2015). There are numerous books written on writing programs in R for different areas of application; for example, see Matloff (2011), Hothorn and Everitt (2014), Cotton (2013), Maindonald and Braun (2010), Muenchen (2011) or any of the topic-specialized books in The R Series of textbooks from CRC. Because it is a free program the developers point out that it "comes with ABSOLUTELY NO WARRANTY".

The SAS philosophy is that the user should string together a sequence of procedures to perform the desired analysis. Some data management and analysis features are available via point and click operations (SAS/ASSIST and SAS/EG). In addition to large volumes of manuals for SAS, numerous texts have been written on using SAS; for example, see Khattree and Naik (1999), Der and Everitt (2014), Delwiche and Slaughter (2012), Marasinghe and Koehler (2018), or Freund and Littell (2000).

SPSS was originally written for survey applications. It offers a number of comprehensive programs and users can choose specific options that they desire. It provides excellent data entry programs and data manipulation procedures. It can be used either by clicking through the file menu system or by writing and executing commands. In addition to the manuals, books such as the ones by Abu-Bader (2010) or Green and Salkind (2016) are available.

Stata is similar to SAS and R in that an analysis consists of a sequence of commands with their own options. Analyses can also be performed via the file menu system with the option to save the commands generated. Features are available to easily log and rerun an analysis. Alternatively, a GUI can be used to select and run analyses. It also includes numerous data management features, a very rich set of graphics options, and a growing set of community-contributed commands for specialized tasks. These are presented through publications such as the Stata Journal. Several books are available which discuss statistical analysis using Stata; see Lalanne and Mesbah (2016), Hamilton (2012), Hills and Stavola (2012), or Cleves et al. (2016) among others.

Since R, SAS and Stata primarily are used by writing and executing a series of commands, effort is required to learn these languages. However, doing so provides the user with a highly versatile programming tool for statistical computing.

When you are learning to use a package for the first time, there is no substitute for reading the on-line HELP, manuals, or texts that present examples. However, at times the sheer number of options presented in these programs may seem confusing, and advice from an experienced user may save you time. Many programs offer default options, and it often helps to use these when you run a program for the first time. In this book, we frequently recommend which options to use. On-line HELP is especially useful when it is programmed to offer information needed for the part of the program you are currently using (context sensitive). Links to the websites for the four software programs discussed in this book can be found in the UCLA web site cited in the preface. Despite the fact that we are providing tables which summarize the commands for the discussed statistical analysis techniques and comment on some in the text, this book is not intended to provide instructions on how to use the statistical software packages. For software specific instructions, the reader is referred to the printed or on-line manuals and books dedicated to that purpose.

There are numerous statistical packages and programming languages that offer statistical packages or modules that are not included in this book. We have tried to choose those that offer a wide range of multivariate techniques.

For information on other packages, you can refer to the statistical computing software review sections of *The American Statistician* or journals in your own field of interest.

3.3 Techniques for data entry

Appropriate techniques for entering data for analysis depend mainly on the size of the data set and the form in which the data set is stored. As discussed below, all statistical packages can use data in a spreadsheet (or rectangular) format. Each column represents a specific variable and each row has the data record for a case or observation. The variables are in the same order for each case. For example, for the depression data set given later in this chapter, looking only at the first three variables and four cases, we have

ID	Sex	Age
1	2	68
2	1	58
3	2	45
4	2	50

where for the variable “sex,” 1 = male and 2 = female, and “age” is given in years.

Typically each row represents an individual case. What is needed in each row depends on the unit of analysis for the study. By unit of analysis, we mean what is being studied in the analysis. If the individual is the unit of analysis, as it usually is, then the data set just given is in a form suitable for analysis. Another situation is when the individuals belong to one household, and the unit of analysis is the household but data have been obtained from several individuals in the household.

Alternatively, for a company, the unit of analysis may be a sales district and sales made by different salespersons in each district are recorded. Data sets given in the last two examples are called hierarchical or clustered data sets and their form can get to be quite complex. Some statistical packages have limited capacity to handle hierarchical data sets. In other cases, the investigator may have to use a relational database package such as Access to first get the data set into the rectangular or spreadsheet form used in the statistical package.

As discussed below, either one or two steps are involved in data entry. The first one is entering the data into the computer if data have been collected on paper. Another typical step is to transfer the data to the desired statistical package.

Data entry

Before entering data in most statistical, spreadsheet, or database management packages, the investigator first names the file where the data are stored, states how many variables will be entered, names the variables, and provides information on these variables. Note that in the example just given we listed three variables which were named for easy use later. The file could be called "depress." Statistical packages commonly allow the user to designate the format and type of variable, e.g., numeric or alphabetic, calendar date, or categorical. They allow you to specify missing value codes, the length of each variable, and the placement of the decimal points. Each program has slightly different features so it is critical to read the appropriate online HELP statements or manual, particularly if a large data set is being entered.

The two commonly used formats for data entry are the **spreadsheet** and the **form**. By spreadsheet, we mean the format given previously where the columns are the variables and the rows the cases. This method of entry allows the user to see the input from previous records, which often gives useful clues if an error in entry is made. The spreadsheet method is very commonly used, particularly when all the variables can be seen on the screen without scrolling.

With the form method, only one record, the one being currently entered, is on view on the screen. There are several reasons for using the form method. An entry form can be made to look like the original data collection form so that the data entry person sees data in the same place on the screen as it is in the collection form. A large number of variables for each case can be seen on a computer monitor screen and they can be arranged in a two-dimensional array, instead of just the one-dimensional array available for each case in the spreadsheet format. Flipping pages (screens) in a display may be simpler than scrolling left or right for data entry. Short coding comments can be included on the screen to assist in data entry. Also, if the data set includes alphabetical information such as short answers to open-ended questions, then the form method is preferred.

The choice between these two formats can be a matter of personal preference, but in general the spreadsheet is used for data sets with a small or medium number of variables and the form is used for a larger number of variables and for studies requiring detailed records of data entry and potential data changes. In some cases a scanner can be used to enter the data and then an optical character recognition program converts the image to the desired text and numbers.

To make the discussion more concrete, we present the features given in a specific data entry package. The SPSS data entry program provides a good mix of features that are useful in entering large data sets. It allows either spreadsheet or form entry and switching back and forth between the two modes. In addition to the features already mentioned, SPSS provides what is called "skip and fill." In medical studies and surveys, it is common that if the answer to a certain question is no, a series of additional questions can then be skipped. For example, subjects might be asked if they ever smoked, and if the answer is yes they are asked a series of questions on smoking history. But if they answer no, these questions are not asked and the interviewer skips to the next section of the interview. The skip-and-fill option allows the investigator to specify that if a person answers no, the smoking history questions are automatically filled in with specified values and the entry cursor moves to the start of the next section. This saves a lot of entry time and possible errors.

Another feature available in many packages is range checking. Here the investigator can enter upper and lower values for each variable. If the data entry person enters a value that is either lower than the low value or higher than the high value, the data entry program provides a warning. For example, for the variable "sex," if an investigator specifies 1 and 2 as possible values and the data entry person hits a 3 by mistake, the program issues a warning. This feature, along with input by forms or spreadsheet, is available also in SAS.

Each software has its own set of features and the reader is encouraged to examine them before entering medium or large data sets, to take advantage of them.

Mechanisms of entering data

Data can be entered for statistical computation from different sources. We will discuss four of them.

1. entering the data along with the program or procedure statements for a batch-process run;
2. using the data entry features of the statistical package you intend to use;
3. entering the data from an outside file which is constructed without the use of the statistical package;
4. importing the data from another package using the operating system such as Windows or MAC OS.

Of note, the above data entry approaches are typically considered acceptable for observational data and/or for small studies. However, some studies, such as most clinical trials, require much more stringent data quality controls and procedures to track over the course of the study including when and by whom data are entered or changed. For initial data entry for such studies data are often entered twice potentially by two different individuals. This is referred to as double data entry and the resulting files are compared to identify data entry errors. Further details can be found in many clinical trial books and articles such as Friedman et al. (2015), and Piantadosi (2017).

The first of the four methods listed above can only be used with a limited number of programs which use program or procedure statements, for example R, SAS or Stata. It is only recommended for very small data sets that are not going to be used very many times. For example, a SAS data set called "depress" could be made by stating:

```
data depress;
   input id sex age;
   cards;
1   2   68
2   1   58
3   2   45
4   2   50
:
run;
```

Similar types of statements can be used for the other programs which use the spreadsheet type of format.

The disadvantage of this type of data entry is that there are only limited editing features available to the person entering the data. No checks are made as to whether or not the data are within reasonable ranges for this data set. For example, all respondents were supposed to be 18 years old or older, but there is no automatic check to verify that the age of the third person, who was 45 years old, was not erroneously entered as 15 years. Another disadvantage is that the data set disappears after the program is run unless additional statements are made. In small data sets, the ability to save the data set, edit typing, and have range checks performed is not as important as in larger data sets.

The second strategy is to use the data entry package or system provided by the statistical program you wish to use. This is always a safe choice as it means that the data set is in the form required by the program and no data transfer problems will arise. Table 3.1 summarizes the built-in data entry

Table 3.1: *Built-in data entry features of the statistical packages*

	R	SAS	SPSS	Stata
Spreadsheet entry	Yes	Yes	Yes	Yes
Form entry	No	Yes	Yes	No
Range check	User	Yes	Yes	No
Logical check	User	Yes	Yes	No
Skip and fill	User	Use SCL	Yes	No
Verify mode	No	No	Yes	No

features of the four statistical packages used in this book. Note that for SAS, `PROC COMPARE` can be used to verify data obtained by double data entry. In general, as can be seen in Table 3.1, SPSS and SAS have extensive data entry features.

The third method is to enter the data into a secondary program such as a spreadsheet or data management program, or a form-based data entry program, and then import it into your statistical software package of choice.

The advantage of this method is that an available program that you are familiar with can be used to enter the data. Excel and Google sheets provides entry in the form of a spreadsheet and is widely available. Access allows entry using forms and provides the ability to combine different data sets. Once the data sets are combined in Access, it is straightforward to transfer them to Excel. Google Forms and other survey software such as Qualtrics and SurveyMonkey allow data entry using forms from that can be distributed online or through email.

Many of the statistical software packages import Excel files. In addition, many of the statistical packages also allow the user to import data from other statistical packages. For example, R will import SAS, SPSS, and Stata data files, but SPSS will import only SAS and Stata data files. Many software packages also allow you to export data in a format designed for use in another statistical package. For example SPSS can not read R data files directly, but R can export SPSS data files directly, or it can write a plain text ASCII file (mentioned next) which can be read into any statistical software program. One suggestion is to first check the manual or HELP for the statistical package you wish to use to see which types of data files it can import.

A widely used transfer method is to create an ASCII file from the data set. ASCII (American Standard Code for Information Interchange) files are more commonly known as **plain text files** and can be created by almost any spreadsheet, data management, or word processing program. Instructions for reading ASCII files are given in the statistical packages. The disadvantage of transferring ASCII files is that typically only the data are transferred, and variable labels and information concerning the variables have to be reentered into the statistical package. This is a minor problem if there are not too many variables. If this process appears to be difficult, or if the investigators wish to retain the variable labels, then they can run a special-purpose program such as STAT/TRANSFER that will copy data files created by a wide range of spread sheet, data base and statistical software programs and put them into the right format for access by other spread sheet, database and statistical programs.

Finally, if the data entry program and the statistical package both use the Windows operating system, then three methods of transferring data may be considered depending on what is implemented in the programs. First, the data in the data entry program may be highlighted and moved to the statistical package using the usual copy and paste options. Second, dynamic data exchange (DDE) can be used to transfer data. Here the data set in the statistical package is dynamically linked to the data set in the entry program. If you correct a variable for a particular case in the entry program, the identical change is made in the data set in the statistical package, Third, object linking and embedding (OLE) can be used to share data between a program used for data entry and statistical analysis. Here also the data entry program can be used to edit the data in the statistical program. The investigator can activate the data entry program from within the statistical package.

If you have a very large data set to enter, it is often sensible to use a professional data entering service. A good service can be very fast and can offer different levels of data checking and advice on which data entry method to use. But whether or not a professional service is used, the following suggestions may be helpful for data entry.

1. Whenever possible, code information in numbers not letters.
2. Code information in the most detailed form you will ever need. You can use the statistical program to aggregate the data into coarser groupings later. For example, it is better to record age as the exact age at the last birthday rather than to record the ten-year age interval into which it falls.
3. The use of range checks or maximum and minimum values can eliminate the entry of extreme values but they do not guard against an entry error that falls within the range. If minimizing errors is crucial then the data can be entered twice into separate data files. One data file can be subtracted from the other and the resulting nonzeros examined. Alternatively, some data entry programs have a verify mode where the user is warned if the first entry does not agree with the second one or have special commands to allow for comparison of two data sets.
4. If the data are stored on a personal computer, then backup copies should be made on an external storage device, such as an external hard drive, or on a cloud storage system such as Box, Dropbox or Google Drive or similar systems. Backups should be updated regularly as changes are made in the data set.
5. For each variable, use a code to indicate missing values. Avoid using potential observed values (such as -9 or 99) for coding missing data. Most programs have their own way to indicate missing values (such as "." or "NA"). The manuals or HELP statements should be consulted so that you can match what they require with what you do.
6. When entering data representing calendar dates, be consistent across all records and use a standard representation such as ISO 8601. This looks like YYYY-MM-DD, which means a four digit year, 2 digit day and 2 digit month, in that order with each separated by hyphens. Example: 2018-07-01 is July 1st, 2018. Times have a similar ISO8601 convention of hh:mm:ss, which can be read as the two digit hour using the 24 hour clock system, two digit minutes, 2 digit seconds, separated by a colon. Consult the HELP manual for your software program for more information on how the program handles date and time formats.

To summarize, there are three important considerations in data entry: accuracy, cost, and ease of use of the data file. Whichever system is used, the investigator should ensure that the data file is free of typing errors, that time and money are not wasted, and that the data file is readily available for future data management and statistical analysis.

3.4 Organizing the data

Prior to statistical analysis, it is often necessary to make some changes in the data set. Table 3.2 summarizes the common options in the programs described in this book.

Combining data sets

Combining data sets is an operation that is commonly performed. For example, in biomedical studies, data may be taken from medical history forms, a questionnaire, and laboratory results for each patient. These data for a group of patients need to be combined into a single rectangular data set where the rows are the different patients and the columns are the combined history, questionnaire, and laboratory variables. In longitudinal studies of voting intentions, the questionnaire results for each respondent must be combined across time periods in order to analyze change in voting intentions of an individual over time. There are essentially two steps in this operation. The first is sorting on some key variable (given different names in different packages) which must be included in both of the separate data sets to be merged. Usually this key variable is an identification or ID variable (case number). The second step is combining the separate data sets side-by-side, matching the correct records with the correct person using the key variable. Sometimes one or more of the data items are missing for an individual. For example, in a longitudinal study it may not be possible to locate a respondent for one or more of the interviews. In such a case, a symbol or symbols indicating missing values will be inserted into the spaces for the missing data items by the program. This is done so that you will end up with a rectangular data set or file, in which information for an individual is put into the proper row, and missing data are so identified.

Data sets can be combined in the manner described above in R by using the `merge` function with the `by` argument to specify the name of the matching variable (such as ID). This function has additional `by.x` or `by.y` arguments that can be used for more complex situations (see the help file). A popular user-written package called `dplyr` contains `join` statements that behave similar to `merge`. The function `cbind` is available to “bind columns” of data, but it can yield unexpected results if the numbers of data points or data types are different in the data sets to be merged. It is also left up to the analyst to be certain that the rows in each data set are listed in the same order.

Combining data sets in SAS can be done by using the `MERGE` statement followed by a `BY` statement and the variable(s) to be used to match the records. The data must first be sorted by the values of the matching variable, say ID. An `UPDATE` statement can also be used to add variables to a master file. In SPSS, you use the `JOIN MATCH` command followed by the data files to be merged if you are certain that the cases are already listed in precisely the same order and each case is present in all the data files. Otherwise, you first sort the separate data files on the key variable and use the `JOIN MATCH` command followed by the `BY` key variable.

In Stata, you `use` the first data file and then use a `merge m:m` key variable `using` the second data file statement. The `m:m` component specifies that either 1 or many records with the same key variable value are merged.

If you have knowledge of the Structured Query Language (SQL) programming language, it is useful to know that both SAS and R have the ability to process SQL queries. Consult your chosen package’s help documentation to learn more about these methods.

In any case, it is highly desirable to list (view) the individual data records to determine that the merging was done in the manner that you intended. If the data set is large, then only the first and last 25 or so cases need to be listed to see that the results are correct. If the separate data sets are expected to have missing values, you need to list sufficient cases so you can see that missing records are correctly handled.

Another common way of combining data sets is to put one data set at the end of another data set. This process is referred to as **concatenation**. For example, an investigator may have data sets that are collected at different places and then combined together. In an education study, student records could be combined from two high schools, with one simply placed at the bottom of the other set.

Concatenation is done using the `rbind` function in R, and `PROC APPEND` in SAS. In SPSS the `JOIN` command with the keyword `ADD` can be used to combine cases from two to five data files, and in Stata the `append` command is used.

It is also possible to update the data files with later information using the editing functions of the package. Thus, a single data file can be obtained that contains the latest information. This option can also be used to replace data that were originally entered incorrectly.

When using a statistical package that does not have provision for merging data sets, it is recommended that a spreadsheet program be used to perform the merging and then, after a rectangular data file is obtained, the resulting data file can be transferred to the desired statistical package. In general, the newer spreadsheet programs have excellent facilities for combining data sets side-by-side or for adding new cases.

Missing values

There are two types of missing data. The first type occurs when no information is obtained from a case, individual, or sampling unit. This type is called **unit nonresponse**. For example, in a survey it may be impossible to reach a potential respondent or the subject may refuse to answer. In a biomedical study, records may be lost or a laboratory animal may die of unrelated causes prior to measuring the outcome. The second type of nonresponse occurs when the case, individual, or sampling unit is available but yields incomplete information. For example, in a survey the respondent may refuse to answer questions on income or only fill out the first page of a questionnaire. Busy physicians may not completely fill in a medical record. This type of nonresponse is called **item nonresponse**. In general, the more control the investigator has of the sampling units, the less apt unit nonresponse or item nonresponse is to occur. In surveys the investigator often has little or no control over the respondent, so both types of nonresponse are apt to happen. For this reason, much of the research on handling nonresponse has been done in the survey field and the terminology used reflects this emphasis.

The seriousness of either unit nonresponse or item nonresponse depends mainly on the magnitude of the nonresponse and on the characteristics of the nonresponders. If the proportion of nonresponse is very small, it is seldom a problem and if the nonresponders can be considered to be a random sample of the population then it can be ignored (see Section 10.2 for a more complete classification of nonresponse). Also, if the units sampled are highly homogeneous then most statisticians would not be too concerned. For example, some laboratory animals have been bred for decades to be quite similar in their genetic background. In contrast, people in most major countries have very different backgrounds and their opinions and genetic makeup can vary greatly.

When only unit nonresponse occurs, the data gathered will look complete in that information is available on all the variables for each case. Suppose in a survey of students 80% of the females respond and 60% of the males respond and the investigator expects males and females to respond differently to a question (X). If in the population 55% are males and 45% are females, then instead of simply getting an overall average of responses for all the students, a weighted average could be reported. For males $w_1 = .55$ and for females $w_2 = .45$. If $\overline{X}_1$ is the mean for males and $\overline{X}_2$ is the mean for females, then a weighted average could be computed as

$$\overline{X} = \frac{\sum w_i \overline{X}_i}{\sum w_i} = \frac{w_1 \overline{X}_1 + w_2 \overline{X}_2}{w_1 + w_2}$$

Another common technique is to assign each observation a weight and the weight is entered into the data set as if it were a variable. Observations are weighted more if they come from subgroups that have a low response rate. This weight may be adjusted so that the sum of the weights equals the sample size. When weighting data, the investigator is assuming that the responders and nonresponders in a subgroup are similar.

In this book, we do not discuss such weighted analyses in detail. A more complete discussion of using weights for adjustment of unit nonresponse can be found in Groves et al. (2002) or Little and Rubin (2002). Several types of weights can be used and it is recommended that the reader consider the various options before proceeding. The investigator would need to obtain information on the units in the population to check whether the units in the sample are proportional to the units in the population. For example, in a survey of professionals taken from a listing of society members if the sex, years since graduation, and current employment information is available from both the listing of the members and the results of the survey, these variables could be used to compute subgroup weights.

The data set should also be screened for item nonresponse. As will be discussed in Section 10.2, most statistical analyses require complete data on all the variables used in the analysis. If even one variable has a missing value for a case, that case will not be used. Most statistical packages provide programs that indicate how many cases were used in computing common univariate statistics such as means and standard deviations (or report how many cases were missing). Thus it is simple to find which variables have few or numerous missing values.

Some programs can also indicate how many missing values there are for each case. Other programs allow you to transpose or flip your data file so the rows become the columns and the columns become the rows (Table 3.2). Thus the cases and variables are switched as far as the statistical package is concerned. The number of missing values by case can then be found by computing the univariate statistics on the transposed data. Examination of the pattern of missing values is important since it allows the investigator to see if it appears to be distributed randomly or only occurs in some variables. Also, it may have occurred only at the start of the study or close to the end.

Once the pattern of missing data is determined, a decision must be made on how to obtain a complete data set for analysis. For a first step, most statisticians agree on the following guidelines.

1. If a variable is missing in a very high proportion of cases, then that variable could be deleted, but this could represent a limitation to the study that might need to be noted.
2. If a case is missing many variables that are crucial to your analysis, then that case could be deleted. If a substantial proportion of cases have this issue, this could be a problem with the generalizability of the analysis results.

You should also carefully check if there is anything special about the cases that have numerous missing data as this might give you insight into problems in data collection. It might also give some insight into the population to which the results actually apply. Likewise, a variable that is missing in a high proportion of the respondents may be an indication of a special problem. Following the guidelines listed previously can reduce the problems in data analysis but it will not eliminate the problems of reduced efficiency due to discarded data or potential bias due to differences between the data that are complete and the grossly incomplete data. For example, this process may result in a data set that is too small or that is not representative of the total data set. That is, the missing data may not be missing completely at random (see Section 10.2). In such cases, you should consider methods of imputing (or filling-in) the missing data (see Section 10.2 and the books by Rubin (2004); Little and Rubin (2002); Schafer (1997); Molenberghs and Kenward (2007), or Laaksonen (2018).

Item nonresponse can occur in two ways. First, the data may be missing from the start. In this case, the investigator enters a code for missing values at the time the data are entered into the computer. One option is to enter a symbol that the statistical package being used will automatically recognize as a missing value. For example, a period, an asterisk (*), or a blank space may be recognized as a missing value by some programs. Commonly, a numerical value is used that is outside the range of possible values. For example, for the variable "sex" (with 1 = male and 2 = female) a missing code could be 9. A string of 9s is often used; thus, for the weight of a person 999 could be used as a missing code.

That value should then be replaced within the software program, using an appropriate command, by that program's specific missing code.For example, using SAS one could state

```
if sex = 9, then sex = . ;
```

Similar statements are used for the other programs. The reader should check the manual for the precise statement. We recommend against the use of missing value codes that could potentially be observed values. Otherwise results of the statistical analyses can be misleading or nonsensical because some actual observations could be considered missing.

If the data have been entered into a spreadsheet program, then it is recommended to leave the cells with missing data blank. Most statistical packages will recognize a blank value as missing.

The second way in which values can be considered missing is if the data values are beyond the range of the stated maximum or minimum values. For example, if the age of a respondent is entered as 167 and it is not possible to determine the correct value, then the 167 should be replaced with a missing value code so an obviously incorrect value is not used.

Further discussion of the types of missing values and of ways of handling item nonresponse in data analysis is given in Section 10.2. Here, we will briefly mention one simple method.

The replacement of missing values with the mean value of that variable is a common option in statistical software packages and is the simplest method of imputation. This method results in underestimation of the variances and covariances that are subsequently used in many analyses. Thus we do not recommend the use of this method.

Detection of outliers

Outliers are observations that appear inconsistent with the remainder of the data set (Barnett and Lewis, 1994). One method for determining outliers has already been discussed, namely, setting minimum and maximum values. By applying these limits, extreme or unreasonable outliers are prevented from entering the data set.

Often, observations are obtained that seem quite high or low but are not impossible. These values are the most difficult ones to cope with. Should they be removed or not? Statisticians differ in their opinions, from "if in doubt, throw it out" to the point of view that it is unethical to remove an outlier for fear of biasing the results. The investigator may wish to eliminate these outliers from the analyses but report them along with the statistical analysis. Another possibility is to run the analyses twice, both with the outliers and without them, to see if they make an appreciable difference in the results. Most investigators would hesitate, for example, to report rejecting a null hypothesis if the removal of an outlier would result in the hypothesis not being rejected. We recommend that whatever decision is made regarding outliers that such decision and potentially its consequences are made transparent and are justified in any report or publication.

A review of formal tests for detection of outliers is given in Barnett and Lewis (1994). To make the formal tests you usually must assume normality of the data. Some of the formal tests are known to be quite sensitive to nonnormality and should only be used when you are convinced that this assumption is reasonable. Often an alpha level of 0.10 or 0.15 is used for testing if it is suspected that outliers are not extremely unusual. Smaller values of alpha can be used if outliers are thought to be rare.

The data can be examined one variable at a time by using histograms and box plots if the variable is measured on the interval or ratio scale. A questionable value would be one that is separated from the remaining observations. For nominal or ordinal data, the frequency of each outcome can be noted. If a recorded outcome is impossible, it can be declared missing. If a particular outcome occurs only once or twice, the investigator may wish to consolidate that outcome with a similar one. We will return to the subject of outliers in connection with the statistical analyses starting in Chapter 7, but mainly the discussion in this book is not based on formal tests.

Transformations of the data

Transformations are commonly made either to create new variables with a form more suitable for analysis or to achieve an approximate normal distribution. Here we discuss the first possibility. Transformations to achieve approximate normality are discussed in Chapter 5.

Transformations to create new variables can either be performed as a step in organizing the data or can be included later when the analyses are being performed. It is recommended that they be done as a part of organizing the data. The advantage of this is that the new variables are created once and for all, and sets of instructions for running data analysis from then on do not have to include the data transformation statements. This results in shorter sets of instructions with less repetition and chance for errors when the data are being analyzed. This is almost essential if several investigators are analyzing the same data set.

One common use of transformations occurs in the analysis of questionnaire data. Often the results from several questions are combined to form a new variable. For example, in studying the effects of smoking on lung function it is common to ask first a question such as:

Have you ever smoked cigarettes? yes___
or no___

If the subjects answer no, they skip a set of questions and go on to another topic. If they answer yes, they are questioned further about the amount in terms of packs per day and length of time they smoked (in years). From this information, a new pack–year variable is created that is the number of years times the average number of packs. For the person who has never smoked, the answer is zero. Transformation statements are used to create the new variable.

Each package offers a slightly different set of transformation statements, but some general options exist. The programs allow you to select cases that meet certain specifications using IF statements. Here for instance, if the response is no to whether the person has ever smoked, the new variable should be set to zero. If the response is yes, then pack–years is computed by multiplying the average amount smoked by the length of time smoked. This sort of arithmetic operation is provided for and the new variable is added to the end of the data set. Variables that are generated or calculated based on other originally collect information are often called **calculated variables**.

Additional options include taking means of a set of variables or the maximum value of a set of variables. For example a survey may ask about the total number of times in the past month an individual has used marijuana, cocaine or LSD in separate questions, but a researcher may be more interested in simply the total number of times in the past month that an individual has used *any* of these drugs.

Another common arithmetic transformation involves simply changing the numerical values coded for a nominal or ordinal variable. For example, for the depression data set, sex was coded male = 1 and female = 2. In some of the analyses used in this book, we have recoded that to male = 0 and female = 1 by simply subtracting one from the given value.

Saving the results

After the data have been screened for missing values and outliers, and transformations made to form new variables, the results are saved in a master file that can be used for analysis. We recommend that a copy or copies of this master file be made on an external storage device such as a CD or USB drive, or in a cloud storage service so that it can be stored outside the computer. A summary of decisions made in data screening and transformations used should be stored with the master file. Enough information should be stored so that the investigator can later describe what steps were taken in organizing the data.

If the steps taken in organizing and preparing the data were performed by typing commands in a statistical programming language, it is recommended that a copy of these commands be recorded in a file and stored along with the data sets. The file containing these commands is commonly referred to as a **code file** or **script file**. Then, should the need arise, the manipulation can be redone by simply editing the code file instructions rather than completely recreating them. We discuss the importance of this process further in the next section on reproducibility.

If results are saved interactively (point and click), then it is recommended that multiple copies be saved along the way until you are perfectly satisfied with the results and that a memo facility or other program be used to document your steps. Some packages such as SPSS and Stata do give you the code that is a result of executing a series of file menu commands. Figure 3.1 summarizes the steps taken in data entry and data management.

3.5 Reproducible research and literate programming

Reproducibility is the ability for any researcher to take the same data set and run the same set of software program instructions as another researcher and achieve the same results (Patil et al., 2016). This process allows others to verify existing findings and build upon them. As data sets become larger and more complex, requiring more sophisticated and computationally intensive anal-

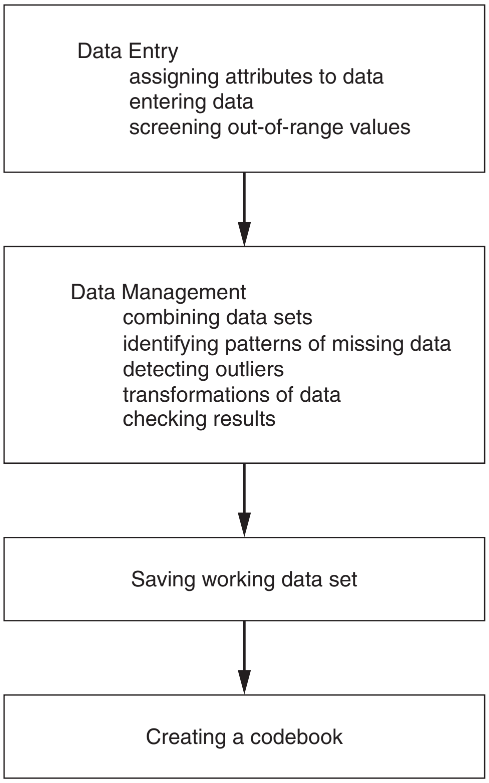


Figure 3.1: *Preparing Data for Statistical Analysis*

ysis methods, the need to provide sufficient information to enable reproducibility of results becomes more important.

The goal is to create an exact record of what was done to a data set to produce a specific result. To achieve reproducibility, we believe that three things must be present:

1. The un-processed data are connected directly to software code file(s) that perform data preparation techniques such as the ones discussed in this chapter and in Chapter 5.
2. The processed data are connected directly to other software code file(s) that perform the analyses.
3. All data and code files are self-contained such that they could be given to another researcher to execute the code commands on a separate computer and achieve the same results as the original author.

Incorporating reproducible research techniques into your workflow not only provides a benefit to the analysts and their collaborators, but also the scientific community in general. In addition, some scientific journals are requesting that authors publish their code and data along with the manuscript (Loder and Groves, 2015; Sturges et al., 2015; Piwowar et al., 2007; Gandrud, 2015).

It is important to note that sometimes making research data available to the scientific community or beyond needs to be done in such a way that confidentiality of participants are protected. Procedures to protect confidentiality might be interfering with exact replication. For more information on methods to protect biological and social data we refer readers to Hauser et al. (2010), and on the ethics on sharing medical data see Hollis (2016).

Literate programming is the programming paradigm introduced by Knuth (1984) intended to present an explanation of the code being written in a natural language such as English. The scientist explains the logic of the program or analysis process in the natural language, with small code snippets included at each step. The code snippets create a full set of instructions that can be executed or compiled to produce a result, such as a data analysis report or a series of data preprocessing steps.

Imagine you are tasked with analyzing a large data set that includes a lot of data-preprocessing, statistical analyses, and creating graphics. You could process the data using a combination of manual and menu driven edits and produce tables and figures individually and copy them into a word processing program where you write up the results in paragraph form.

Now imagine that you find out that there were additional errors in the original data, or that additional data records are now available and need to be included in the analysis. You are now faced with completely repeating all the effort to process the data, conduct statistical analyses and create the report.

Practicing reproducible research techniques using literate programming tools allows such major updates to be a simple matter of re-compiling all coded instructions using the updated data set. The effort then is reduced to a careful review and update of any written results.

Literate programming tools such as those listed in Table 3.3 use additional markup languages such as **Markdown** or LaTeX to create formatted documents with section headers, bold and italicized words, tables and graphics with built-in captions in a streamlined manner that is fully synchronized with the code itself. The author writes the text explanations, interpretations, and code in the statistical software program itself, and the program will execute all commands and combine the text, code and output all together into a final dynamic document.

For details on how to use literate programming tools such as those listed in Table 3.3 we refer the reader to references such as Xie (2015) and Leisch and R-Core (2017) for programming in R, Lenth and Højsgaard (2007) for programming in SAS, and Haghish (2016a) and Haghish (2016b) for programming in Stata.

For a more general discussion on tools, practices and guidelines and platforms to conduct reproducible research see Stodden et al. (2014). For a detailed and open source guide to enhancing reproducibility in scientific results and writing, see Martinez et al. located at https://ropensci.github.io/reproducibility-guide/

3.6 Example: Depression study

In this section we discuss a data set that will be used in several succeeding chapters to illustrate multivariate analyses. The **depression study** itself is described in Chapter 1.

The data given here are from a subset of 294 observations randomly chosen from the original 1000 respondents sampled in Los Angeles. This subset of observations is large enough to provide a good illustration of the statistical techniques but small enough to be manageable. Only data from the first time period are included. Variables are chosen so that they would be easily understood and would be sensible to use in the multivariate statistical analyses described in Chapters 7–18.

The codebook, the variables used, and the data set are described below.

Codebook

In multivariate analysis, the investigator often works with a data set that has numerous variables, perhaps hundreds of them or more. An important step in making the data set understandable is to create a written codebook that can be given to all the users. The codebook should contain a description of each variable and the variable name given to each variable for use in the statistical package. Some statistical packages have limits on the length of the variable names so that abbreviations are used. Often blank spaces are not allowed, so dashes or underscores are included. Some statistical packages reserve certain words that may not be used as variable names. The variables should be listed in the same order as they are in the data file. The codebook serves as a guide and record for all users of the data set and as documentation needed to interpret results.

Table 3.4 contains a codebook for the depression data set. In the first column the variable number is listed, since that is often the simplest way to refer to the variables in the computer. A variable name is given next, and this name is used in later data analysis. These names were chosen to be eight characters or less so that they could be used by all the statistical programs at the time when the data set was created (early 1980's). This eight character limitation on variable names is no longer a restriction. It is helpful to choose variable names that are easy to remember and are descriptive of the variables, but short to reduce space in the display.

Finally a description of each variable is given in the last column of Table 3.4. For nominal or ordinal data, the numbers used to code each answer are listed. For interval or ratio data, the units used are included. Note that income is given in thousands of dollars per year for the household; thus an income of 15 would be $15,000 per year. Additional information that is sometimes given includes the number of cases that have missing values, how missing values are coded, the largest and smallest value for that variable, simple descriptive statistics such as frequencies for each answer for nominal or ordinal data, and means and standard deviations for interval or ratio data. Additional columns could be used to add information regarding the variables (e.g., if there were changes in the variables for different versions of the data collection instrument), or to indicate whether the variable is recorded as collected or whether and how it was calculated based on other variables. We note that one package (Stata) can produce a codebook for its users that includes much of the information just described.

Depression variables

The 20 items used in the depression scale are variables 9–28 and are named C1, C2,..., C20. (The wording of each item is given later in the text, in 14.2.) Each item was written on a card and the respondent was asked to tell the interviewer the number that best describes how often he or she felt or behaved this way during the past week. Thus respondents who answered item C2, "I felt depressed," could respond 0–3, depending on whether this particular item applied to them rarely or none of the time (less than 1 day: 0), some or little of the time (1–2 days: 1), occasionally or a moderate amount of the time (3–4 days: 2), or most or all of the time (5–7 days: 3).

Most of the items are worded in a negative fashion, but items C8–C11 are positively worded. For example, C8 is "I felt that I was as good as other people." For positively worded items the scores are **reversed**: that is, a score of 3 is changed to be 0, 2 is changed to 1, 1 is changed to 2, and 0 is changed to 3. In this way, when the total score of all 20 items is obtained by summation of variables C1–C20, a large score indicates a person who is depressed. This sum is the 29th variable, named CESD (short for: Center for Epidemiological Studies Depression).

Persons whose CESD score is greater than or equal to 16 are classified as depressed since this value is the common cutoff point used in the literature (Aneshensel and Frerichs, 1982). These persons are given a score of 1 in variable 30, the CASES variable. The particular depression scale employed here was developed for use in community surveys of noninstitutionalized respondents (Comstock and Helsing, 1977; Radloff, 1977).

Data set

As can be seen by examining the codebook given in Table 3.4 demographic data (variables 2–8), depression data (variables 9–30), and general health data (variables 32–37) are included in this data set. Variable 31, DRINK, was included so that it would be possible to determine if an association exists between drinking and depression.

The actual data for the first 30 of the 294 respondents included here are listed in Table 3.5. The rest of the data set, along with the other data sets used in this book, are available on the CRC Press and UCLA web sites (see Appendix A).

3.7 Summary

In this chapter we discussed the steps necessary before statistical analysis can begin. The first of these is the decision of what computer and software packages to use. Once this decision is made, data entry and organizing the data can be started.

Note that investigators often alter the order of these operations. For example, some prefer to check for missing data and outliers and to make transformations prior to combining the data sets. This is particularly true in analyzing longitudinal data when the first data set may be available well before the others. This may also be an iterative process in that finding errors may lead to entering new data to replace erroneous values. Again, we stress saving the results on some other external or cloud storage device after each set of changes.

Four statistical packages — R, SAS, SPSS, and Stata — were noted as the packages used in this book. In evaluating a package it is often helpful to examine the data entry and data manipulation features they offer. The tasks performed in data entry and organization are often much more difficult and time consuming than running the statistical analyses, so a package that is easy and intuitive to use for these operations is a real help. If the package available to you lacks needed features, then you may wish to perform these operations in one of the spreadsheet or relational database packages and then transfer the results to your statistical package.

3.8 Problems

3.1 Enter the data set given in Table 9.1, Chemical companies' financial performance (Section 9.3), using a data entry program of your choice. Make a codebook for this data set.

3.2 Using the data set entered in the previous problem, delete the P/E variable for the Dow Chemical company and D/E for Stauffer Chemical and Nalco Chemical in a way appropriate for the statistical package you are using. Then, use the missing value features in your statistical package to find the missing values and replace them with an imputed value.

3.3 Transfer or read in a data set that was entered into a spreadsheet program into your statistical software package.

3.4 Describe the person in the depression data set who has the highest total CESD score.

3.5 For the statistical package you intend to use, describe how you would add data from three more time periods for the same subjects to the depression data set.

3.6 Combine the results from the following two questions into a single variable that measures the total number of days the individual has been sick during the time period.: This would allow one variable to be used for analysis involving this data.

a. Have you been sick during the last two weeks?

Yes, go to b. ___
No ___

b. How many days were you sick? ___

3.7 Consistency checks are sometimes performed to detect possible errors in the data. If a data set included information on sex, age, and use of contraceptive pill, describe a consistency check that could be used for this data set.

3.8 In the Parental HIV data set, the variable LIVWITH (who the adolescent was living with) was coded 1=both parents, 2=one parent, and 3=other. Transform the data so it is coded 1=one parent, 2=two parents, and 3=other using the features available in the statistical package or spreadsheet program you are using.

3.9 From the variables ACUTEILL and BEDDAYS described in Table 3.4, create a single variable that takes on the value 1 if the person has been both bedridden and acutely ill in the last two months and that takes on the value 0 otherwise.

Table 3.2: *Data management features of the statistical packages.*

	R*	SAS	SPSS	Stata
Merging data sets	`merge`, `join`	MERGE	MATCH FILES	merge
Adding data sets	`rbind`, `cbind`	PROC APPEND, SET	ADD FILES	append
Hierarchical data sets	`reshape`, reshape2, tidyr	Write multiple OUTPUT statements RETAIN	CASESTOVARS	reshape, frlink
Transpose data	`t`	PROC TRANSPOSE	FLIP	xpose
Missing value imputation	mice	PROC MI, PROC MIANALYZE	MULTIPLE IMPUTATION	mi
Calendar dates	`as.Date` chron, lubridate	INFORMAT	FORMATS	dates

*`Monospace font` denotes the function name. Normal font denotes a user written package containing functions to perform the specified task

Table 3.3: *Literate Programming Tools*

Software	Addons or Packages
R	RMarkdown, Sweave, knitr
SAS	SASWeave
SPSS	
Stata	MarkDoc, Ketchup, Weaver, Dyndoc, Markdown

Table 3.4: *Codebook for depression data*

Variable number	Variable name	Description
1	ID	Identification number from 1 to 294
2	SEX	1 = male; 2 = female
3	AGE	Age in years at last birthday
4	MARITAL	1 = never married; 2 = married; 3 = divorced; 4 = separated; 5 = widowed
5	EDUCAT	1 = less than high school; 2 = some high school; 3 = finished high school; 4 = some college; 5 = finished bachelor's degree; 6 = finished master's degree; 7 = finished doctorate
6	EMPLOY	1 = full time; 2 = part time; 3 = unemployed; 4 = retired; 5 = houseperson; 6 = in school; 7 = other
7	INCOME	Thousands of dollars per year
8	RELIG	1 = Protestant; 2 = Catholic; 3 = Jewish; 4 = none; 5 = other
9–28	C1–C20	"Please look at this card and tell me the number that best describes how often you felt or behaved this way during the past week." 20 items from depression scale (already reflected; see text) 0 = rarely or none of the time (less than 1 day); 1 = some or a little of the time (1–2 days); 2 = occasionally or a moderate amount of the time (3–4 days); 3 = most or all of the time (5–7 days)
29	CESD	Sum of C1–20; 0 = lowest level possible; 60 = highest level possible
30	CASES	0 = normal; 1 = depressed, where depressed is CESD$\geq$16
31	DRINK	Regular drinker? 1 = yes; 2 = no
32	HEALTH	General health? 1 = excellent; 2 = good; 3 = fair; 4 = poor
33	REGDOC	Have a regular doctor? 1 = yes; 2 = no
34	TREAT	Has a doctor prescribed or recommended that you take medicine, medical treatments, or change your way of living in such areas as smoking, special diet, exercise, or drinking? 1 = yes; 2 = no
35	BEDDAYS	Spent entire day(s) in bed in last two months? 0 = no; 1 = yes
36	ACUTEILL	Any acute illness in last two months? 0 = no; 1 = yes
37	CHRONILL	Any chronic illness in last year? 0 = no; 1 = yes

Table 3.5: *Depression data for the first 30 respondents*

OBS	ID	SEX	AGE	MARITAL	EDUC	EMPLOY	INCOME	RELIG	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	CESD	CASES	DRINK	HEALTH	REGDOC	TREAT	BEDDAYS	ACUTEILL	CHRONILL
1	1	2	68	5	2	4	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	0	0	1
2	2	1	58	3	4	1	15	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	4	0	1	1	1	1	0	0	1
3	3	2	45	2	3	1	28	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	4	0	1	2	1	1	0	0	0
4	4	2	50	3	3	3	9	1	0	0	0	0	1	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	5	0	2	1	1	2	0	0	1
5	5	2	33	4	3	1	35	1	0	0	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	6	0	1	1	1	1	1	1	0
6	6	1	24	2	3	1	11	1	0	0	0	0	0	0	0	0	1	0	0	1	2	0	0	2	1	0	0	0	7	0	1	1	1	1	0	1	1
7	7	2	58	2	2	5	11	1	2	1	1	2	1	0	0	2	2	0	0	0	0	0	3	0	0	0	0	1	15	0	2	3	1	1	0	1	1
8	8	1	22	1	3	1	9	1	0	1	2	0	2	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	10	0	2	1	2	2	0	1	0
9	9	2	47	2	3	4	23	2	0	1	1	0	0	3	0	0	0	0	0	3	0	3	2	3	0	0	0	0	16	1	1	4	1	1	1	0	1
10	10	1	30	2	2	1	35	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2	0	0	0
11	11	2	20	1	2	3	25	4	0	0	1	0	1	2	1	0	0	1	0	1	2	2	1	1	2	3	0	0	18	1	1	2	1	2	0	0	0
12	12	2	57	2	3	2	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	4	0	2	2	1	1	1	1	1
13	13	1	39	2	2	1	28	1	1	1	0	0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	1	1	8	0	1	3	1	1	0	1	0
14	14	2	61	5	3	4	13	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	4	0	1	1	1	1	0	1	0
15	15	2	23	2	3	1	15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	0	2	1	8	0	1	1	1	2	0	0	0
16	16	2	21	1	2	1	6	1	1	1	2	0	1	1	1	1	2	2	0	1	1	2	1	1	1	2	0	0	21	1	1	3	1	1	1	0	1
17	17	2	23	1	4	1	8	1	3	3	2	3	3	3	2	2	3	2	2	2	1	2	3	2	0	1	0	3	42	1	1	1	2	2	1	1	0
18	18	2	55	4	2	3	19	1	1	0	1	1	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	6	0	2	3	1	1	1	1	1
19	19	2	26	1	6	1	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	1	1	0
20	20	1	64	5	2	4	9	4	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	2	1	2	0	0	0
21	21	2	44	1	3	1	6	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	3	0	1	1	1	1	0	0	1
22	22	2	25	2	3	1	35	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	4	0	1	2	1	1	0	1	1
23	23	2	72	5	3	4	7	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	1	2	1	1	0	0	1
24	24	2	61	2	3	1	19	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	0	4	0	2	3	1	1	0	0	1
25	25	2	43	3	3	1	6	1	0	0	0	0	1	0	1	2	1	0	0	1	0	1	1	2	0	0	0	0	10	0	2	3	1	1	0	0	1
26	26	2	52	2	2	5	19	2	1	2	1	0	1	0	0	0	0	0	0	1	1	0	3	2	0	0	0	0	12	0	1	3	1	1	0	0	0
27	27	2	23	2	3	5	13	1	0	0	0	0	0	0	0	3	0	0	0	0	1	1	0	0	0	1	0	0	6	0	2	2	1	2	0	1	0
28	28	1	73	4	2	4	5	2	0	1	2	0	2	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	9	0	1	3	1	1	0	0	1
29	29	2	34	2	3	2	19	2	0	2	2	0	1	0	2	1	1	1	1	2	3	2	3	3	2	0	0	2	28	1	1	2	1	2	0	0	0
30	30	2	34	2	3	1	20	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	2	1	2	0	0	1

Chapter 4

Data Visualization

4.1 Introduction

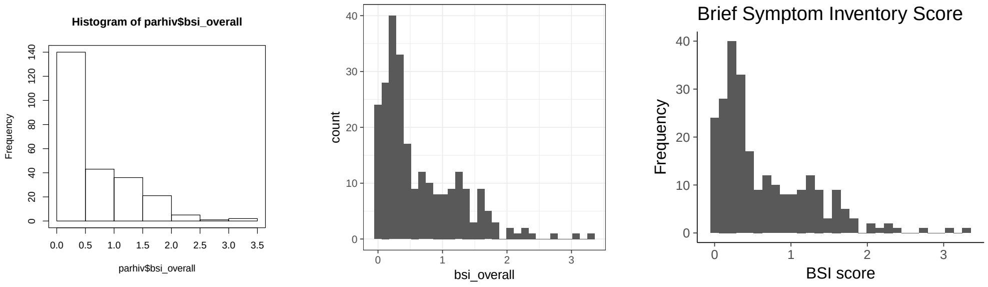
Visualizing data is one of the most important things we can do to become familiar with the data. There are often features and patterns in the data that cannot be uncovered with summary statistics alone. There tends to be two forms in which data can be presented; Summary tables are used for comparing exact values between groups for example, and plots for conveying trends and patterns when exact numbers are not always necessary to convey a story. This chapter introduces a series of plot types for both categorical and continuous data. We start with visualizations for a single variable only (univariate), then combinations of two variables (bivariate), and lastly a few examples and discussion of methods for exploring relationships between more than two variables (multivariate). Additional graphs designed for a specific analysis setting are introduced as needed in other chapters of this book.

This chapter uses several data sets described in Appendix A. Specifically, we use the parental HIV and the depression data sets to demonstrate different visualization techniques. Almost all graphics in this chapter are made using R, with section 4.5 containing a discussion of graphical capabilities to create these graphs in other statistical software programs.

There are three levels of visualizations that can be created, with examples shown in Figure 4.1a, b and c.

- **For your eyes only (4.1a):** Made by the analyst, for the analyst, these plots are quick and easy to create, using the default options without any annotation or context. These graphs are meant to be looked at once or twice for exploratory analysis in order to better understand the data.
- **For an internal report (4.1b):** Some chosen plots are then cleaned up to be shared with others, for example in a weekly team meeting or to be sent to co-investigators participating in the study. These plots need to be capable of standing on their own, but can be slightly less than perfect. Axis labels, titles, colors, annotations and other captions are provided as needed to put the graph in context.
- **For publication or external report (4.1c):** These are meant to be shared with other stakeholders such as the public, your collaborator(s) or administration. Very few plots make it this far. These plots should have all the “bells and whistles” as they appear in formal reports, and are often saved to an external file of a specific size or file type, with high resolution. For publication in most printed journals and books, figures typically need to be in black and white (possibly grayscale).

Along with having the audience in mind, it is important to give thought to the purpose of the chart. “The effectiveness of any visualization can be measured according to how well it fulfills the tasks it was designed for.” (A. Cairo, personal communication, Aug 9, 2018).



(**a**) *For your eyes only* (**b**) *For an internal report* (**c**) *For publication*

Figure 4.1: *Three levels of graphic quality and completeness using histograms as examples. The size of each plot above are the same, the inclusion and formatting of titles and axes will impact size of the plotting region.*

4.2 Univariate Data

This section covers how to visualize a single variable or characteristic. We start with plots for categorical data, then cover plots for continuous data. Visualization is one of the best methods to identify univariate outliers, skewness, low frequencies in certain categories and/or other oddities in the distribution of the data.

4.2.1 Categorical Data

Categorical (nominal or ordinal) data are summarized by reporting the count, or frequency, of records in the data set that take on the value for each category of the variable of interest. (See Section 2.3 for a review of data type classifications.) Common methods to display the counts of categorical data include tables, dot plots, and pie charts. The subsections below discuss and demonstrate each of these types.

Tables

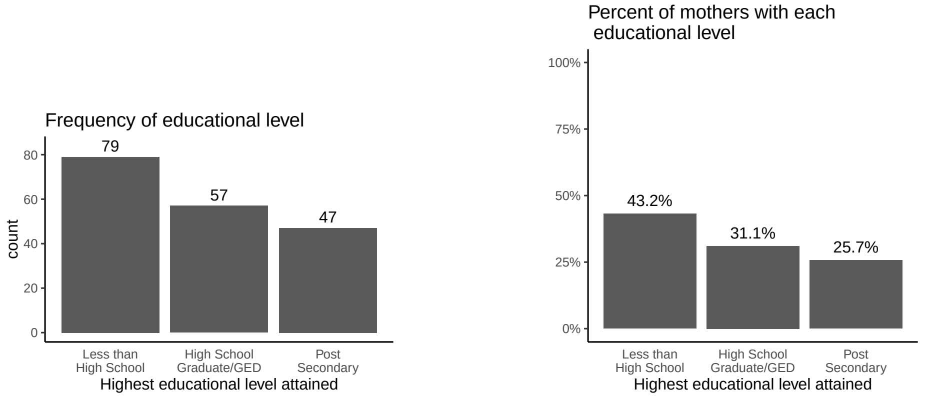
A **table** is the most common way to organize and display summary statistics of a categorical variable using just numbers. Tables should show both the frequency (N) and the percent for each category. That way readers can compare relative group sizes and the overall magnitude of data at the same time. Some software packages, such as SPSS, will automatically display percentages and generate a total row for frequency tables, others packages such as R require a follow up commands such as `prop.table` for percentages or `addmargins` for the total.

Table 4.1: *Education level among mothers with HIV*

Education Level	N	Percent
Less than High School	79	43.2%
High School Graduate/GED*	57	31.1%
Post Secondary	47	25.7%

*GED: General Education Development, an alternative
to a High School Diploma.

Table 4.1 shows that about a quarter (47, 25.7%) of mothers in the parental HIV data set have post-secondary school education level.



(a) *A bar chart comparing frequencies*

(b) *A bar chart comparing percentages*

Figure 4.2: *Two bar charts showing the distribution of highest level of education attained*

Bar Charts

A **bar chart** (Figure 4.2) takes these frequencies and draws bars for each category (shown along the horizontal axis) where the heights of the bars are determined by the frequencies seen in the table (Figure 4.2a). A reasonable modification is to put the percentages on the vertical axis (Figure 4.2b). This is a place to be cautious however. Some programs by default will exclude the missing data before calculating percentages, so the percentages shown are for available data. Other programs will display a bar for the missing category and display the percentages out of the full data set. For either choice it is advised that the analyst understand what the denominator is.

The ordering of categories is important for readability. Nearly all statistical software packages will set the automatic factor ordering to alphabetical, or according to the numerical value that is assigned to each category. If the data are ordinal, tables and plots should read left to right along with that ordering, such as the educational level example in Figure 4.2. Sometimes there is a partial ordering such as years of high school education and then different degrees that are not necessarily easy to summarize in years (can one year of vocational school be considered equivalent to one year of college or one year of community college?). In these situations it is left to the researcher to decide on how to define and justify the order of the categories using subject matter expertise.

Cleveland dot plot

Bars use a lot of ink, and the width of the bar is typically meaningless. **Cleveland dot plots** (Cleveland, 1993) provide an alternative method to display the frequencies using less ink and space than bar charts (Figure 4.6). This is especially helpful when there are a large number of categories to visualize. We use marital status as an example. Because it is a nominal variable (in contrast to the previous ordinal variable examples), these summary data are best displayed in descending order of frequency.

An important note is that Cleveland dot plots plot summary data, not raw data. After summarizing the data, such as calculating the frequency of records per category, we now only have one data point per category to plot. Examples of plots that include each individual data point are shown later in this chapter. There are numerous ways to depict data using dots on a graph. We attempt to be clear in our explanation of each plot discussed in this chapter, however naming conventions of various "dot plots" are not universally consistent. For example, the type of plot shown in Figure 4.6 is referred to as a dot plot with example R code using the function `dotchart` by Kabacoff (2015).

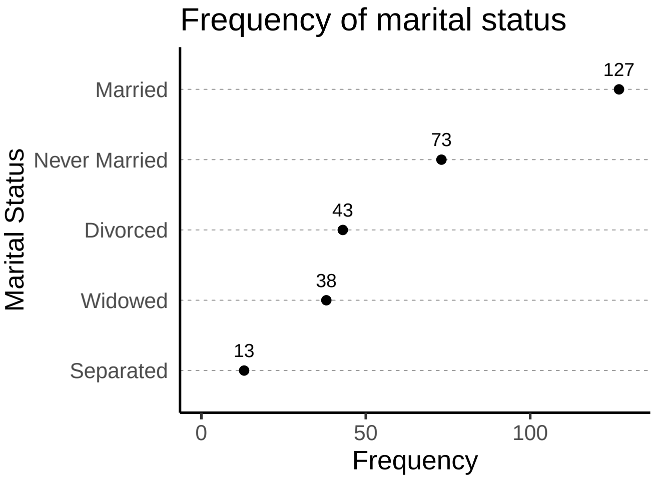


Figure 4.3: *A Cleveland dot plot of marital status*

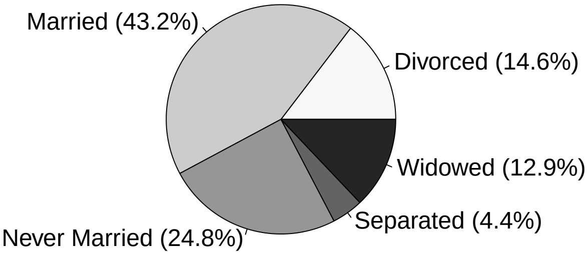


Figure 4.4: *A fully labeled pie chart of marital status. Note that percentages may not always add up to 100% due to rounding*

Pie Charts

Each wedge of a **pie chart** (Figure 4.4) contains an internal angle indicating the relative proportion of records in that category. However, human eyes cannot distinguish between angles that are close in size as well as they can distinguish between heights of bars (or lines or dots). As the number of categories increases, a necessary component to make a pie chart interpretable is having labels with names and percentages for each wedge. Depending on the defaults of the software program, the segments may start either at the 12 o'clock position or the 3 o'clock position.

4.2.2 Continuous Data

Continuous data by definition can take on infinite possible values, so the above plots that display frequencies of records within a finite number of categories do not apply here unless the continuous data are categorized into distinct groups (e.g. income brackets). To visualize continuous data, we need to display the actual value or the distribution of the data points directly. Common plot types include: stem-leaf plots, stripcharts, histograms, density graphs, boxplots, and violin plots. So, what do these plots depict and how are they generated?

Stem-leaf plots

The **stem-leaf plot** (Tukey, 1972) demonstrates how numbers placed on a line can describe the shape of the distribution of observed values and provide a listing of all individual observations in the same plot. Since this type of graphic includes each individual data point, the usefulness and readability diminishes as the number of data points increases. Figure 4.5 displays the values of age for individuals in the depression data set.

```
1 | 8888899999
2 | 00000011111122222222233333333333444444444
2 | 5555556666666667777888889999
3 | 00000011111222222222233333444444444
3 | 555566666677777889
4 | 000001222222222333333344
4 | 555566677777788889999
5 | 00000111111222233444
5 | 55556667777778888888999999999
6 | 000000011111222233444
6 | 555556667788889
7 | 000001112233444
7 | 5778899
8 | 011233333
8 | 9
```

Figure 4.5: *A stem-leaf plot of the individual age in the depression data set*

Because stem-leaf plots display the value of every observation in the data set, the data values can be read directly. The first row displays data from 15 to 19 years of age, or, the second half of the 10s place. Note that this study enrolled only adults, so the youngest possible age is 18. There are five 18 year olds and five 19 year olds in the data set. From this plot one can get an idea of how the data are distributed and know the actual values (of ages in this example). The second row displays data on ages between 20 and 24, or, the first half of the 20s. The third row displays data on ages between 25 and 29, or, the second half of the 20s, and so forth.

Stripcharts

Another type of plot where the value of of every observation in the data set is represented on the graph called a **stripchart**. Figure 4.6 depicts the age of an individual in the depression data set as a single dot. The points here have been *jittered* (where equal values are moved slightly apart from each other) to avoid plotting symbols on top of each other and thus making them difficult or impossible to identify.

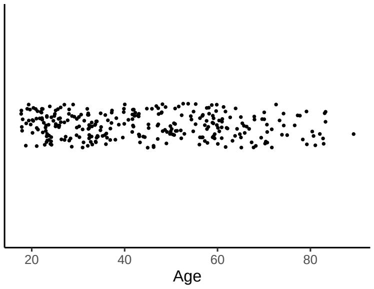
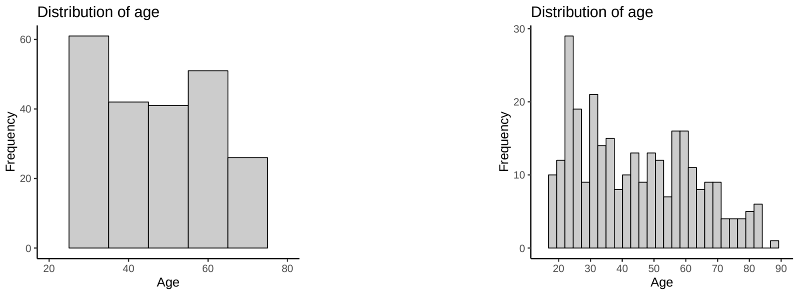


Figure 4.6: *A stripchart of the individual age in the depression data set*

Here we offer more cautionary words due to similar sounding plot names. Some authors may refer to this type of plot as a dotplot, a one-way dot, one-dimensional scatterplot, or a stripplot. In



(a) *A histogram with 8 bins* **(b)** *Using bins of width range/30*

Figure 4.7: *Histograms displaying the distribution of age in the depression data set*

some programs, dotplots differ from stripcharts in that in dotplots are created where the width of the dot is determined by a binning algorithm similar to the ones used to calculate the widths of the bars in a histogram, or (Wilkinson, 1999). As the the defaults of each software program, command or function may vary, we encourage the reader to refer to the help manual for their chosen program for more information.

Often we are not interested in the individual values of each data point; rather we want to examine the distribution of the data or summary measures of the distribution. Example questions might be: where is the majority of the data? Does the distribution look symmetric around some central point? Around what values do the bulk of the data lie? For example, the distribution of ages of individuals in the depression data set ranges from 18 to 89, is slightly skewed with a right tail, unimodal, with a mean around 45.

Histograms

Rather than showing the value of each observation, we often prefer to think of the value as belonging to a *bin*, or an interval. The heights of the bars in a **histogram** display the frequencies of values that fall into those bins. For example, if we grouped the ages of individuals in the depression data set into 10 year age bins, the frequency table looks like this:

(15,25]	(25,35]	(35,45]	(45,55]	(55,65]	(65,75]	(75,85]	(85,95]
57	61	42	41	51	26	15	1

In this table, the notation (15,25] indicates that this bin includes values that are between 15 and 25, including 25 but excluding 15, and so forth.

To create a histogram, the values of a continuous variable are plotted on the horizontal axis, with the height of the bar for each bin equal to the frequency of data within that bin (Figure 4.7a). The main difference between a histogram and a bar chart is that bar charts plot a categorical variable on the horizontal axis, so the vertical bars are separated. The horizontal axis of a histogram is continuous, so the bars touch each other, and thus there is no gap between bins because the bins represent directly adjacent categories.

The choice of the size of each bin can highlight or hide some features in the data. If there is no scientifically motivated or otherwise prespecified bin size, we might start with the default value for the chosen statistical software package, as we did in Figure 4.7b, and then adjust as necessary. Figure 4.7b displays the same data on ages in the depression data set using the default value of range/30 chosen by the ggplot2 package in R. For this example the range of ages is 18 to 89, thus the binwidth is $\frac{(89-18)}{30} = 2.4$. This choice of bin width shows the most frequent age at around 23, unlike Figure 4.7a where it appears to be over 25.

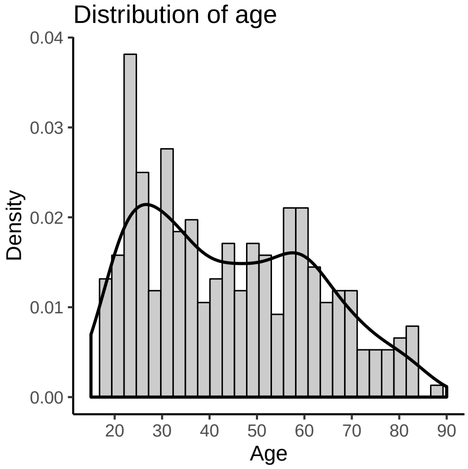


Figure 4.8: *Density plot for the distribution of age in the depression data set*

Kernel density plots

Instead of plotting bars for each bin, we can sometimes get a better (or different) idea of the true shape of the distribution by creating a **kernel density plot**. The kernel density is a function, $f(x)$, that is generated from the data set, similar to a histogram. Density plots differ from histograms in that this function is a smooth continuous function, not a stepwise discrete function that creates bars with flat tops. See Everitt and Skrondal (2010) for more information on how the kernel density is calculated.

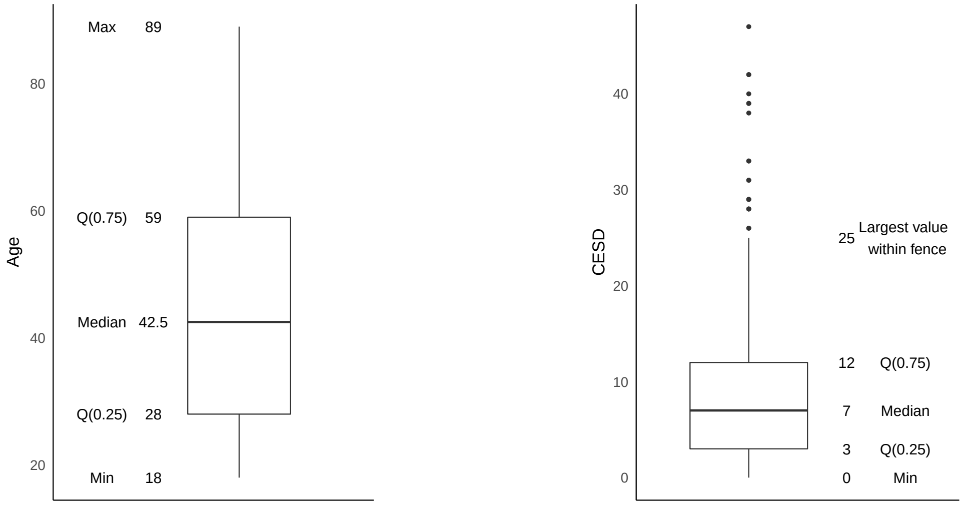
Figure 4.8 shows that the density line smooths out the multitude of peaks and valleys in the histogram, providing a better idea of the general shape of the data. Notice that the vertical axis on a **density plot** is no longer the frequency or count, but the value of the kernel density. While the density curve in Figure 4.8 is overlaid on top of the histogram with the smaller bin width, the first peak of the density plot is around 25, which is more representative of Figure 4.7a which uses the wider bin size. This highlights the importance of looking at multiple types of graphics to fully understand the distribution of the data.

Boxplots and Violin plots

A **boxplot** (also called **box-whisker** plot) display the five number summary (Min, $Q(0.25)$, Median, $Q(0.75)$, Max) in graphical format, where $Q(0.25)$ indicates that 25% of the data are equal to or below this value. The data are arranged in ascending order and then separated into four equal sized groups, i.e., same number of data points are in each of the four sections of a boxplot (Figure 4.9a).

The box outlines the middle 50%, or the interquartile-range ($IQR = Q(0.75) - Q(0.25)$) of the data, and the horizontal lines (whiskers) extend from the 1st quartile ($Q(0.25)$) down to the minimum value, and upwards from the third quartile $Q(0.75)$ to the maximum value. This means that in Figure 4.9a, the same number of individuals in the depression data set are between the 10 year span of ages between 18 and 28, as there are between the 30 year span of ages between 59 and 89.

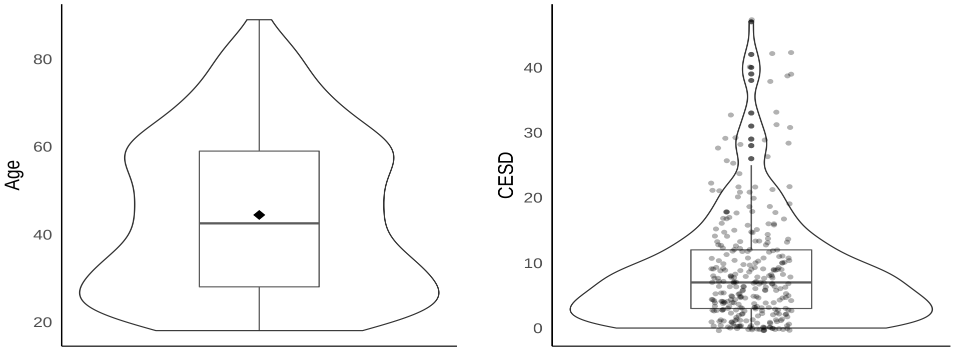
Some statistical packages plot the **modified boxplot** by default. We first define the **fences**, i.e., $Q(0.25) - 1.5 * IQR$ and $Q(0.75) + 1.5 * IQR$. Then, in the modified boxplot, the whiskers do not extend all the way out to the maximum and minimum, but out to the data points that are just inside the fences as calculated by the $1.5 * IQR$ rule. In the modified boxplots, outliers are typically denoted as points or dots outside the fences. For example, consider the continuous measure of depression, variable CESD (Center for Epidemiological Studies Depression) in Figure 4.9b. Here the upper



(a) *An unadorned boxplot of age*

(b) *A modified boxplot of CESD*

Figure 4.9: *An unadorned, and a modified boxplot*



(a) *Add a violin plot & the mean*

(b) *Add a violin plot & jittered points*

Figure 4.10: Boxplot enhancements

whisker extends to 25, the maximum value inside $1.5 * IQR$. The points above 25 are considered potential outliers. Some researchers choose to extend whiskers out to $Q(0.05)$ and $Q(0.95)$. See section for more information on how these values are calculated and used. Additions that make boxplots much more informative are displayed in Figure 4.10:

1) adding the mean as a point (Figure 4.10a).
2) adding a **violin** plot to show the density (reflected around the mid-line of the boxplot, Figure 4.10a). Violin plots are not commonly used, but they can be very informative in that they can display the shape of the kernel density in the same graph as the boxplot.
3) adding the data points directly as jittered dots (Figure 4.10b).

Numerous other modifications of this type of plot and other methods to visualize univariate continuous data have been proposed (among others, box-percentile plots by Esty and Banfield, 2003, extending whiskers to specified quantiles by Cleveland, 1985; Reimann et al., 2008, and “mountain” plots by Goldstein, 1996.)

4.3 Bivariate Data

Next we introduce graphical methods to explore relationships between two variables. Many of the same plotting types such as boxplots and histograms introduced for univariate exploration will be used again here.

4.3.1 Categorical versus Categorical Data

To compare the distribution of one categorical variable across levels of another categorical variable, primarily tables are created. Tables come under several names including **cross-tabulation**, **contingency** tables and **two-way** tables. Table 4.2 displays the frequency of gender by education level for individuals in the depression data set. The value in each cell is the number of records in the data set with that combination of factor levels. For example, there are 4 males with less than a high-school (HS) degree, 26 females who have completed a Bachelor’s (BS) degree, and 8 males who have completed a master’s (MS) degree.

Table 4.2: *Two-way frequency table of gender by educational level*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD	Total
Male	4	19	39	18	17	8	6	111
Female	1	42	75	30	26	6	3	183
Total	5	61	114	48	43	14	9	294

When group sizes are not comparable, it is more informative to compare percents instead of frequencies. There are three types of percentages that can be calculated, with each one having its own purpose. Table 4.3 displays a table of **cell percents**, where the denominator is the entire sample. There are 13.3% of all respondents in this data set who are male and have graduated high school. Table 4.4 displays the **row percents**, where the denominator is the row total. Relatively more males than females completed a four year degree: 15.3% of males completed a BS degree, compared to 14.2% of females. Table 4.5 displays the **column percents**, where the denominator is the column total. The majority of PhD graduates were male; 66.7% of the PhD graduates were male and 33.3% female.

Table 4.3: *Cell percents: Percent out of the entire data set*

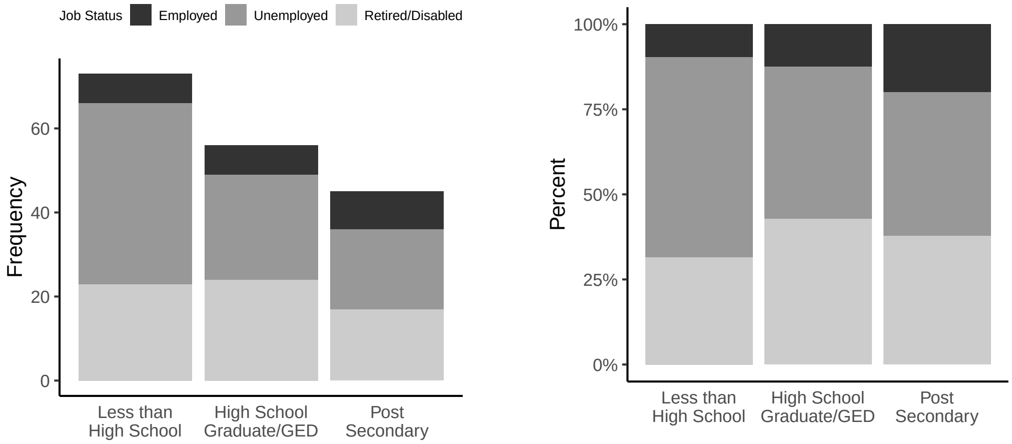
	<HS	Some HS	HS Grad	Some college	BS	MS	PhD	Total
Male	1.4	6.5	13.3	6.1	5.8	2.7	2.0	37.8
Female	0.3	14.3	25.5	10.2	8.8	2.0	1.0	62.2
Total	1.7	20.7	38.8	16.3	14.6	4.8	3.1	100.0

Table 4.4: *Row percents: Percent of educational level within each gender*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD	Total
Male	3.6	17.1	35.1	16.2	15.3	7.2	5.4	100.0
Female	0.5	23.0	41.0	16.4	14.2	3.3	1.6	100.0

Table 4.5: *Column percents: Percent of gender within each educational level*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD
Male	80.0	31.1	34.2	37.5	39.5	57.1	66.7
Female	20.0	68.9	65.8	62.5	60.5	42.9	33.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0



(a) *Frequency on the vertical axis*

(b) *Percents on the vertical axis*

Figure 4.11: *Distribution of current job status within highest education attained in the parental HIV data set*

Bar Charts

To visually compare the distribution of one categorical variable within levels of another categorical variable, we return to bar charts. Figure 4.11 compares the distribution of job status within the highest educational level attained using the parental HIV data set.

Stacked bar charts can be informative when plotting percentages instead of counts. Figure 4.11b shows how the proportion of observations in each job status category compare across each level of highest educational level attained. This plot is created by plotting column percentages, so that all percents within a column add up to 100%. The group of respondents whose highest education level is post secondary has the highest proportion of employed respondents compared to the other two educational level groups.

The default for some software programs, such as R, is a stacked bar chart as shown in Figure 4.11. For few categories this option could be acceptable. However, consider the proportion of those with HS/GED who are unemployed, is it bigger or smaller than the percent of those with post secondary degrees who are currently unemployed? It is difficult to tell in a stacked bar chart, but much easier to see the difference with the bars placed side by side (Figure 4.12).

Alternatively, stacked bar charts could be displayed in a tabular manner with spaces between the bars. This spacing allows for easier comparison within, and across categories. Figure 4.13 provides an example of this method using community-contributed Stata command `tabplot` (Cox, 2016).

The Cleveland dot plot can also be done across groups. Figure 4.14 demonstrates a slight variation where the dot is placed at the end of the solid line, instead of on a reference line as in Figure 4.6. In some fields this type of plot is referred to as a **lollipop** plot. This is also the first demonstration of **paneling**, where the data for each level of the grouping variable are set apart from the other levels using a rectangular border or frame. This method helps to visually separate the groups.

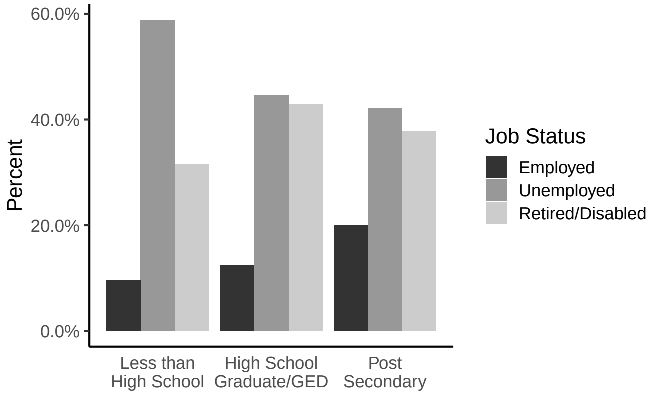


Figure 4.12: *Side by side bar chart depicting the percent of current job status within highest education level attained in the parental HIV data set*

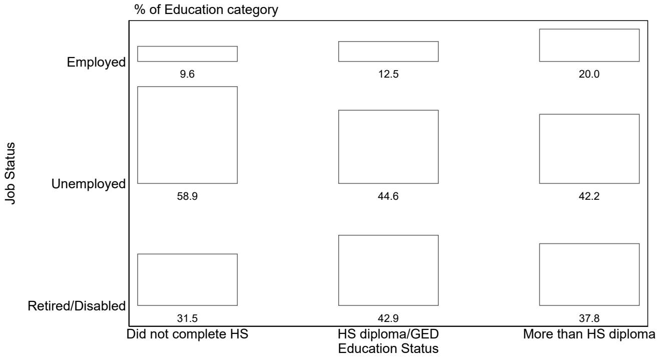


Figure 4.13: *Tabular bar chart comparing job status and educational level*

Another way the Cleveland dot plot can be used is to highlight differences in frequencies between two groups. Figure 4.15 shows the difference in the frequency of males and females from the Parental HIV data, within each of the mothers job status categories. There are more males than females with unemployed mothers, but the difference in counts between genders is less than 10. There are about 20 more females than males with mothers who are retired or disabled.

Mosaic plot

Bar plots and dot plots show either the row or the column percents of a bivariate comparison. They compare the distribution of one categorical variable within levels of a second categorical variable. **Mosaic plots** provide a graphical method to compare the association between two categorical variables.

Figure 4.16 compares job status to educational level by visualizing the cell proportions as area of a square. The heights of the boxes correspond to the marginal distribution of educational level, and the widths of the boxes correspond to the marginal distribution of job status. The area of each smaller rectangle is proportional to the percent of data with that combination of levels. Using Table 4.6 as a numerical reference, 4% of responses in the parental HIV data set have a GED and are employed, whereas 24.7% have less than a HS education and are currently unemployed. This may

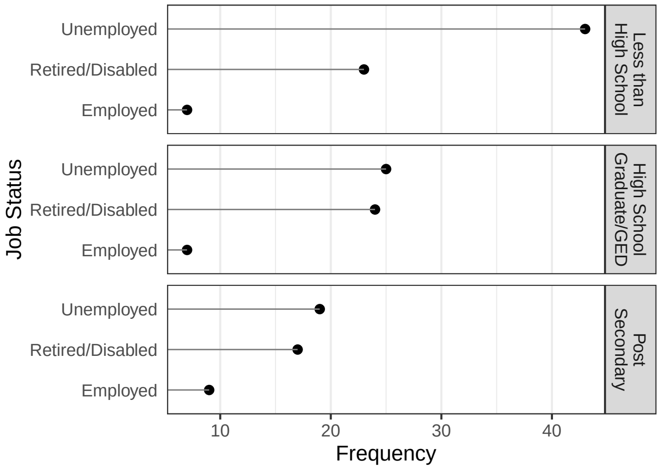


Figure 4.14: *Cleveland dot plot demonstrating the frequency of job status within highest education attained*

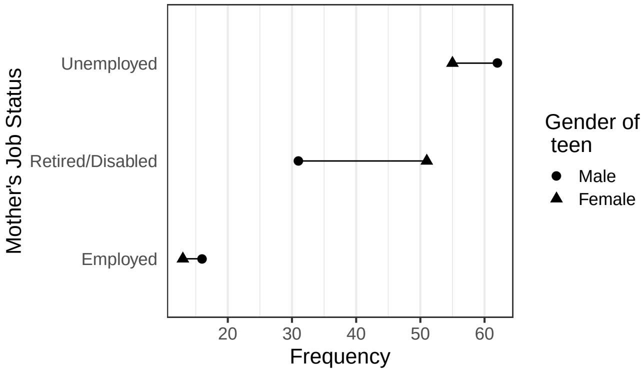


Figure 4.15: *Cleveland dot plot of the differences in frequency of males and females within the mother's job status*

seem like a high proportion of unemployment, but recall that these data were collected in the early nineties, where there were very limited treatment options for HIV positive individuals.

Table 4.6: *Cell percentages for the combination of educational level and job status*

	Employed	Unemployed	Retired/Disabled
Less than High School	4.0	24.7	13.2
High School Graduate/GED	4.0	14.4	13.8
Post Secondary	5.2	10.9	9.8

4.3.2 Continuous versus Categorical Data

When comparing the distribution of a continuous variable across levels of a categorical variable, the same types of plots seen for a single continuous variable can be used including histograms, density plots, boxplots and violin plots.

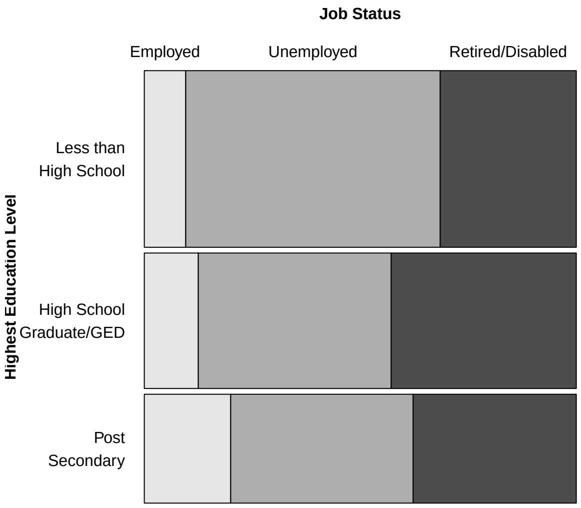
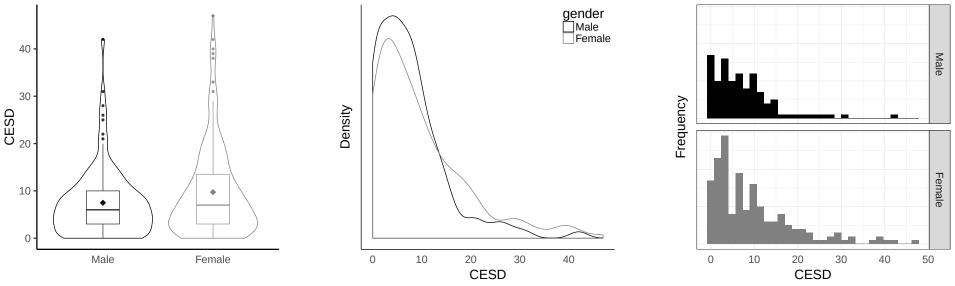


Figure 4.16: *Mosaic plot comparing job status and educational level*

Figure 4.17 demonstrates how to plot the distribution within each group side by side (a), overlaying plots onto the same plotting grid (b), or to create a grid of **panels** (c) with one group per panel. It is very important to use a shared or common axis when comparing conditional distributions across groups.



(a) *Side by side*

(b) *Overlapping*

(c) *Paneled*

Figure 4.17: *Three methods to compare the distribution of the continuous variable CESD across levels of the categorical variable gender*

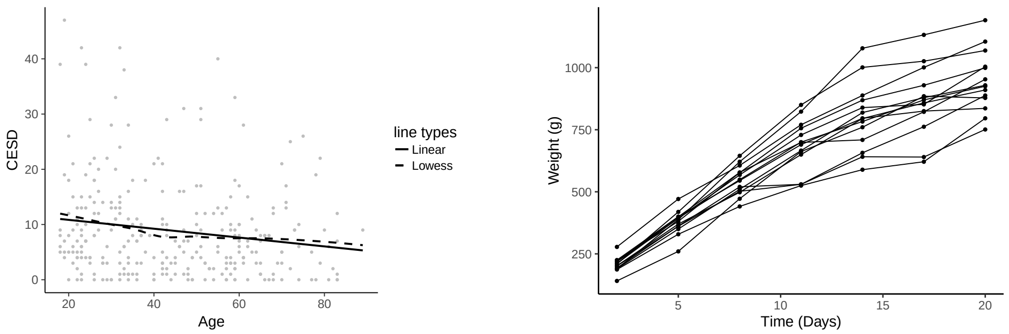
4.3.3 Continuous versus Continuous Data

The most common method of visualizing the relationship between two continuous variables is the **scatterplot** (Figure 4.18a). Lines are often added to help see the trend in the data points. The two most common best fit methods are the straight line (shown as a solid line) and the **lowess** smoother line (shown as a dashed line). In this plot the points have been colored grey to place the emphasis on the lines. These two methods are both regression techniques discussed in Chapter 7.

Line plots

Line plots connect the points with a line. This is typical for time series and in **profile plots** where the goal is to track the data on an individual or a population over time. One line is plotted per individual. For data sets with a larger number of individuals this process can create an unreadable plot. We suggest plotting data on a random subset of individuals to explore the data or create multiple such plots for subsets if feasible.

Figure 4.18b uses the mice data set described in Appendix A where the weights of mice were measured periodically for about a month. The mice grew almost at the same rate until about 8 days, and then started to separate due to individual and treatment characteristics. This particular type of plot is also known as a **spaghetti plot** or a **growth curve** because it typically presents a measure of growth over time. We use this plot again in Section 18.7 when discussing how to analyze longitudinal data.



(a) *Scatterplot of age against CESD, with a dashed lowess line and solid best fit linear line*

(b) *Weight over time for 14 mice*

Figure 4.18: *Two types of scatterplots for examining the relationship between two continuous variables. Points in the scatterplot (a) are not connected to other points whereas in the line plot (b) points from the same mouse are connected with a line*

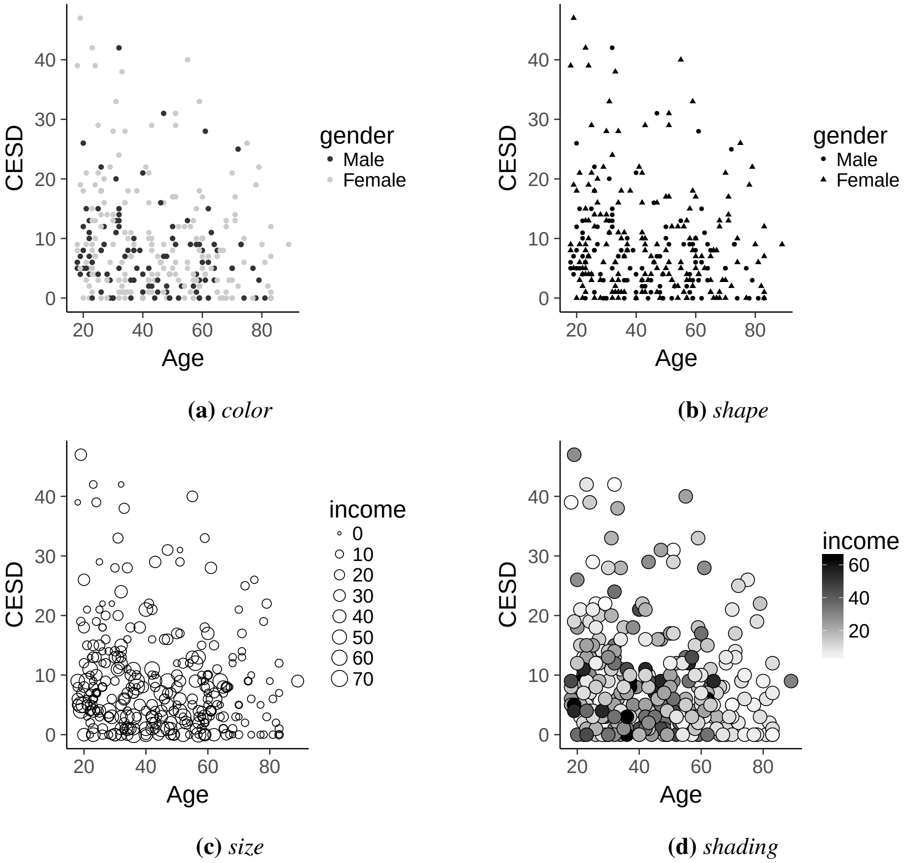
4.4 Multivariate Data

The techniques of applying colors, shadings, positioning and paneling of data from multiple groups to visualize bivariate relationships can be extended to visualize relationships among more than two variables simultaneously.

Figure 4.19 demonstrates how a third dimension can be added onto a bivariate scatterplot by changing the (a) color or (b) shape of the points according to the level of a third categorical variable, or by changing the (c) size or (d) fill shade of the points according to a continuous variable. For example, plot (a) allows us to see that the points in the low range of age with high CESD score are primarily female (grey dots), and plot (d) that those in the higher income levels (darker shades) tend to have a CESD score below 10.

There are many other ways to examine a multivariate relationship. Even on each of these plots just discussed a fourth layer could be added, such as changing the size of the point by income in plots (a) and (b) and shape of the point by gender in plots (c) and (d). Another method to examine a multivariate relationship is to use paneling in two dimensions. Figure 4.20 demonstrates how we can examine the histogram of overall BSI (brief symptom inventory) score for each combination of employment status and highest educational level attained.

A **scatterplot matrix** is a common tool to examine the bivariate relationships between multiple continuous variables simultaneously. Figure 4.21 demonstrates a publication-ready version of a scatterplot matrix that has many features added, including the pairwise correlation (see Chapter



(a) *color*

(b) *shape*

(c) *size*

(d) *shading*

Figure 4.19: *Scatterplot of CESD as a function of age, with a third characteristic included using different methods*

7 for details), univariate histograms and lowess lines on the scatterplots. Each diagonal rectangle displays the histogram of a particular variable. This single plot lets us identify characteristics of the data such as (1) the distribution of BSI is skewed with a long tail to the right as demonstrated by the tall bar representing frequency of responses for low values of BSI, and a few very short bars for high values of BSI, (2) the parent bonding care sub-scale is also considered skewed since the bars are short for low values of the sub-scale and increase in height as the scale increases, (3) there is a moderate positive correlation ($r = 0.37$) between the age a youth starts smoking and drinking and (4) none of the other three variables seem to be correlated with BSI since the lowess lines through the scatterplots are all approximately horizontal.

We remind the reader that the example of a scatterplot matrix shown in Figure 4.21 was created with some customization applied. Each software program has different defaults such as background coloring and axis labels and tick marks that the user may want to modify. For example a researcher may want to change Figure 4.21 to show all axis labels on the left and bottom for familiar positioning.

Another way to visualize the relationship between many continuous variables is by examining the **correlation** between all pairs of variables. The correlation is a numeric summary statistic that quantify the direction and strength of a relationship between two continuous measures. The calculation and interpretation of the correlation can be found in Section 8.7. Figure 4.22 demonstrates one approach where the circles sizes represent both the direction and magnitude of the correlation. Here the shading gradient goes from white (+1) to black (-1), and is not actually recommended for

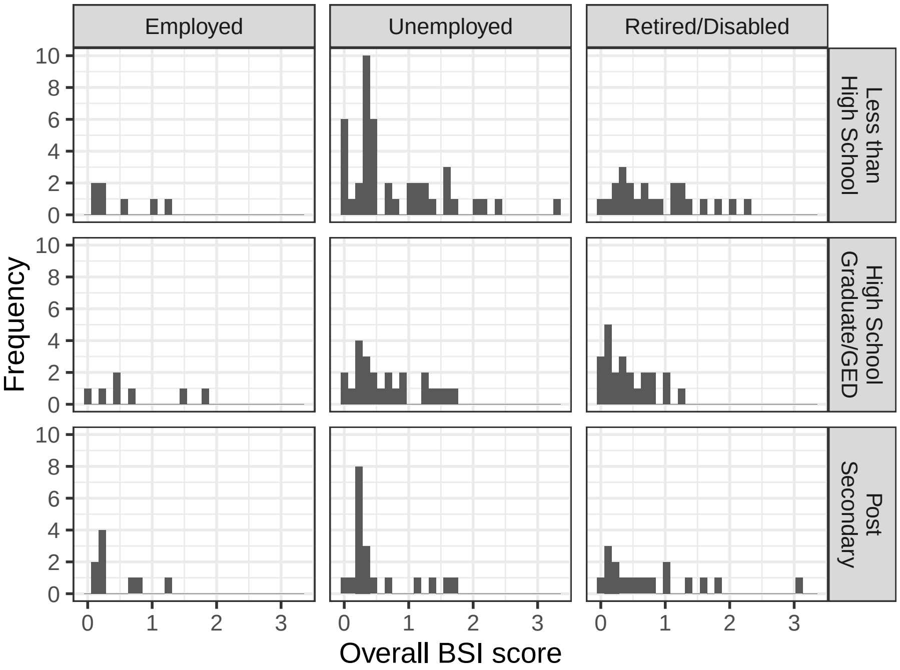


Figure 4.20: *A histogram of overall BSI score paneled on the combination of two other variables: employment status and highest educational level attained*

a diverging scale such as the correlation. For more information on choosing appropriate color gradients see Zeileis et al. (2009). This approach to visualizing the correlation matrix is useful in some multivariate analyses such as principle component analysis discussed in Chapter 14.

4.5 Discussion of computer programs

Each general statistical software package has commands or procedures to produce many, if not all, of the plots or visualizations we describe in this chapter. Table 4.7 shows which command can be used to produce a particular plot using the three major packages discussed in this book. The full R code for all tables and plots in this Chapter are available on the CRC Press and UCLA web sites (see Appendix A).

Additional notes for Table 4.7 and most other software command tables in the book:

- **R:** Entries that are in `monospace font` are functions within Base R. Entries in normal font are packages that contain functions (not specifically listed here) that are used to create the selected plot. All packages in R are user written and must be installed prior to use.
- **SAS:** All entries are individual procedures, called PROCs. Not all are part of BASE SAS. PROC GPLOT, GCHART, and GTL are part of SAS/GRAPH. PROC TEMPLATE is listed here as part of the Graph Template Language, which provides full customization of SAS Graphics.
- **SPSS:** With the exception of creating tables, all available graphics are best built using the Chart Builder. Table entries provide guidance for the reader to find the appropriate selection. The Chart Builder also has tools to easily change the color and shape of the point (or marker).
- **Stata:** Options within commands are written in *(italics)*. Entries marked with a dagger † are community-contributed commands.

Table 4.7: *Software commands for plotting*

	Visualizations	R[*]	SAS	SPSS	Stata
Univariate					
Categorical	Table	`table`	FREQ	FREQUENCIES	table, tabulate
	Bar Chart	`plot`, ggplot2	GCHART	Bar	graph bar, catplot†,tabplot†
	Cleveland Dot Plot	`dotchart`, ggplot2	SGPLOT, FREQ SGPLOT	Scatter/Dot -Summary Point Plot	graph dot
	Pie Chart	`pie`	GCHART	Pie	graph pie
Continuous	Stem-Leaf	`stem`	SGPLOT, UNIVARIATE	EXAMINE	stem
	Stripchart / Dotplot	`stripchart`, ggplot2	 SGPLOT	Scatter/Dot	dotplot, stripplot†
	Histograms	`hist`, ggplot2	SGPLOT, UNIVARIATE	Histogram	histogram
	Kernel Density	`plot`, ggplot2	SGPLOT, UNIVARIATE	Histogram	kdensity
	Boxplot	`boxplot`, ggplot2	BOXPLOT, SGPLOT	Boxplot	graph box,
	Violin Plots	ggplot2	TEMPLATE	-	vioplot†
Bivariate					
Cat v Cat	Two-way table	`table`	FREQ	CROSS TABS	table, tabulate
	Bar Chart	`plot`, ggplot2	GCHART	Bar	graph bar, catplot†,tabplot†
	Mosaic Plot	`mosaicplot`, vcd	TEMPLATE	-	spineplot†[a]
Cont v. cat	grouping	`plot`, ggplot2	BOXPLOT, SGPLOT	Histogram, Boxplot	histogram,graph box
Cont v Cont	Scatterplot	`plot`, ggplot2	GPLOT, SGPLOT	Scatter	scatter
	Line Plot	`plot`, ggplot2	GPLOT, SGPLOT	Line	line
Multivariate					
	paneling	ggplot2, lattice	SGPANEL	Groups tab	by
	colors, size	`plot`, ggplot2	GPLOT, GCHART	-	color, *(msize, weight)*
	scatterplot matrix	psych, lattice, `pairs`, car	SGSCATTER	Scatter/Dot -Scatterplot Matrix	graph matrix
	correlation matrix	corrplot			twoway scatter, corrtable†

[a] Technically not a mosaic plot (Hummel, 1996)

[*] `Monospace font` denotes the function name. Normal font denotes a user written package containing functions to perform the specified task

† Community contributed command

Italic text denotes options available within several commands

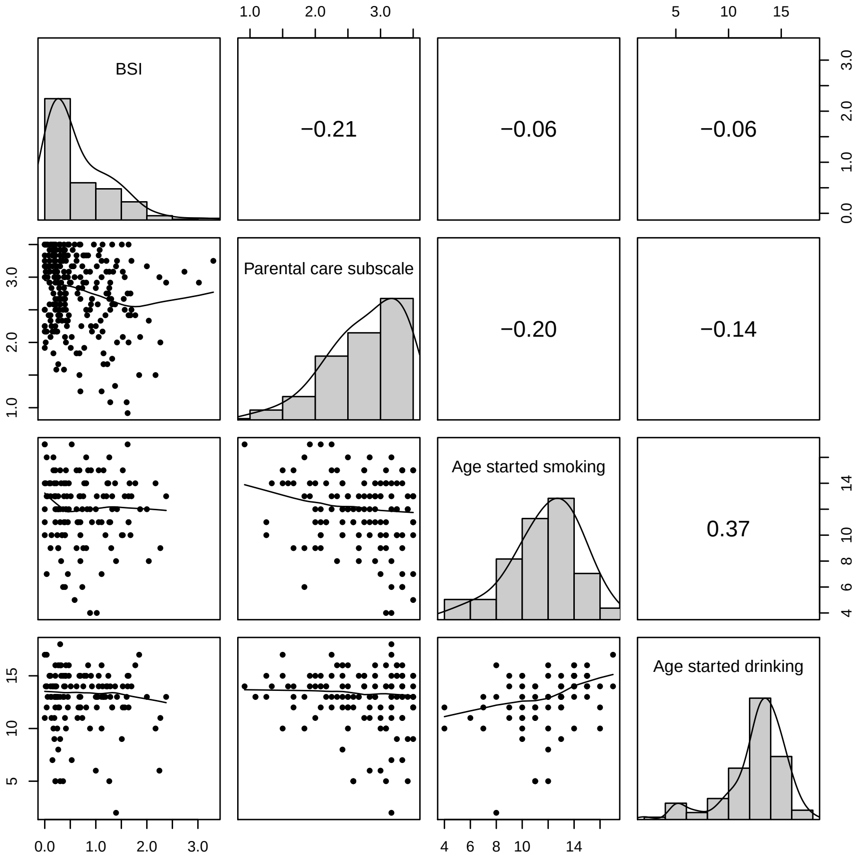


Figure 4.21: *Scatterplot matrix with histograms and density plots along the diagonal, and pairwise correlation values above the diagonal*

4.6 What to watch out for

- **Avoid complexity.** We advise against using too many enhancements on a single plot since doing so can confuse the reader instead of providing a better understanding. The purpose of most graphics is to understand distributional patterns, and to identify odd data points. Not all layers provide illumination. For example, in Figure 4.19c there is much over-plotting in the lower left so that it is difficult to see if there is a pattern emerging. In this case coloring the points for different income levels may be more helpful than changing the size, although barely.
- **Choose colors mindfully.** All plots in this textbook are either in black and white or shaded using a grayscale. This is a necessary adjustment for black and white printing, but also is a consideration for colorblind readers. We recommend using a colorblind-friendly color palette for publications involving color.
- **Do not add extra dimensions.** We do not demonstrate plots such as 3D pie charts, or 3D bar charts in this text. In these cases the third dimension does not provide true information, and is considered "chart junk" that can be very misleading. This is not a global recommendation; there are circumstances in which a third dimension does contain additional information that may be useful to include in the visualization.

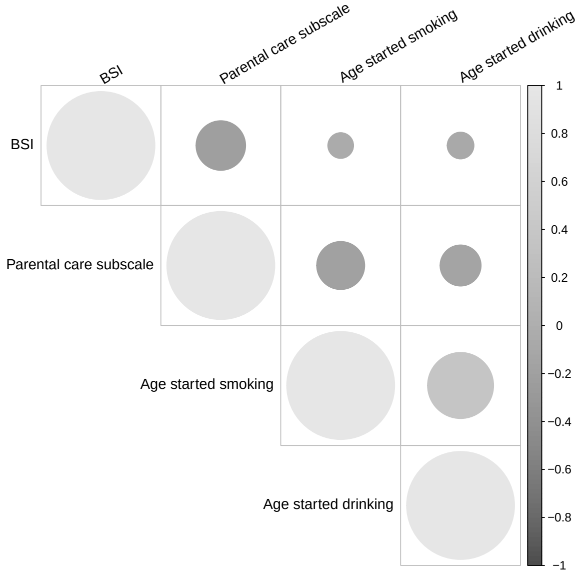


Figure 4.22: *Visual representation of a correlation matrix circle sizes and shading representing the direction and magnitude of the correlation between all pairs of variables.*

- **Be truthful with the scaling.** Be mindful of the scaling of the vertical axis. For example, Figure 4.2b plots a percentage on the vertical axis with a high value of near 50%. We scale the vertical axis to 100% here to put the difference in percentages in the context of the overall range. A 2% point difference between categories can appear huge if the vertical axis only has a total range of say 5%. Similarly Figure specifically has a zero mark for the frequency. Displaying a vertical axis that is too large or too small relative to the data is one of the most common ways in which graphics can be misleading.
- **Check publishing guidelines.** If the goal is to publish using graphics, then be sure to check the rules carefully. Some publications have rules regarding features such as whether there is a box around the plot versus only showing the horizontal and vertical axes.
- **Be consistent with selected themes.** For example, if the first plot has a clear background and a box outlining the edge, all subsequent plots should have that same theme. If a second categorical variable controls the color or shape of the points for one plot, then all subsequent plots that also use that same categorical variable should have the same color and/or shape scheme applied.
- **Do not over-interpret.** One should not judge statistical significance based on graphs unless they are specifically designed for such. Even if they are (seemingly) designed for it (e.g., graphing confidence intervals for group comparisons or 95% pointwise confidence intervals or confidence bands for graphs) they can produce different and potentially misleading results and interpretations as compared to appropriate hypothesis tests.
- **Plotting with missing data.** Each software program has slightly different default values for how missing data are handled in different plots. For example, the `table` function in R does not automatically include a column for missing data, but a bar chart created using ggplot2 will show a bar for the missing category. In all software packages; values that are missing will be omitted from continuous data plots such as histograms and scatterplots, with typically no mention of the sample size being used to create the plot shown.

4.7 Summary

Data visualization is a powerful tool that can be used to explore and understand your data. Data values that need to be recoded can easily be identified and trends in the data can be uncovered. Graphics can be used to confirm that the data meet certain criteria, or assumptions that are needed for further statistical analysis such as the specialized analysis methods presented in the remainder of the book. Data visualizations can also enhance understanding of statistical analysis results, but do not serve as a substitute. Even though the saying "a picture is worth a thousand words" may be true, and a graph can provide more information than a block of text or a table of numbers, when it comes to more than a few variables, the options for graphically representing the data are limited.

We have demonstrated a wide variety of visualizations in this chapter. Some plots require detailed written explanations and are more suitable for reports or publications that do not have length restrictions. Authors should attempt to strike a balance between complexity and interpretability of graphics, yet always aim to elucidate characteristics of data relevant to the question being asked or answered.

There are many other types of graphics that we do not discuss such as heatmaps, ridgelines, choropleth maps and word clouds. These are typically considered specialized graphics for specific analyses. We present some specialized plots in the appropriate chapters of this book but do not attempt to cover all possible ways to display information visually. We recommend looking at Edward Tufte's pioneering work for historical overviews and inspiration (Tufte, 2001, 2006), and Yau (2011) for how to tell stories real world data. Additional handbooks that include practical advice on how to choose and create effective visualizations include Munzner (2014); Cairo (2016); Kirk (2019) and Holtz and Conor (2018).

For the programming language details on how to make the graphics shown in this chapter and others, we refer readers to this book's supplemental webpage and reference books such as the *R Graphics Cookbook* by Chang (2013), *R Graphics* by Murrell (2011), *ggplot2: Elegant Graphics for Data Analysis* by Wickham (2016), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* by Wickham and Grolemund (2017), *Statistical Graphics in SAS* by Kuhfeld (2010), *A Visual Guide to Creating Graphs Interactively* by Matange and Bottitta (2016), *Handbook of Statistical Graphics using SAS* by Der and Everitt (2014), the *IBM SPSS Statistics 24 Brief Guide* (IBM, 2016), *Building SPSS Graphs to Understand Data* by Aldrich and Rodriguez (2012), *Speaking Stata Graphics* by Cox (2014), and *A Visual Guide to Stata Graphics* by Mitchell (2012).

4.8 Problems

Descriptions of data sets, how to obtain them, and the codebooks can be found in Section 1.2 and Appendix A.

4.1 From the lung function data set, determine how many families have one child, two children, and three children between the ages of 7 and 18.

4.2 For the depression data set, determine if any of the variables have observations that do not fall within the ranges given in Table 3.4, codebook for depression data.

4.3 For the lung function data set, create a new variable called AGEDIFF = (age of child 1) – (age of child 2) for families with at least two children. Produce a frequency count of this variable. Are there any negative values? Comment.

4.4 Construct histograms for mothers' and fathers' heights and weights from the lung function data set. Describe cases that you consider to be outliers.

4.5 For the lung cancer data set,

a) construct a histogram of the variable Days

b) for every other variable produce a frequency table of all possible values.

4.6 For the lung function data set, produce a two-way table of gender of child 1 versus gender of child 2 (for families with at least two children). Describe the distribution of genders for these families.

4.7 For the lung cancer data set,

a) produce a separate histogram of the variable Days for small and large tumor sizes (0 and 1 values of the variable Staget)

b) compute a two-way frequency table of the variable Staget versus the variable Death

c) comment on the results of (a) and (b).

4.8 Construct a scatterplot of income versus employment status from the depression data set. From the data in this table, decide if there are any adults whose incomes are unusual considering their employment status. Are there any adults in the data set whom you think are unusual?

4.9 Using the lung function data explore and describe the following relationships:

a) How does the residential area affect the lung function of the parents?

b) For the oldest child, plot the relationship between FEV1 and (i) age; (ii) height; (iii) weight using a scatterplot and lowess line.

4.10 Using the parental HIV data set create a few visualizations to explore the relationship between the variables of interest listed below. In a few sentences describe the information about the given relationship you learn from each graph, and what specific features of the graph led you to that conclusion.

a) The relationship between the age when a child starts smoking and when they start drinking.

b) Ethnicity, a choice of one neighborhood characteristic, and financial situation of the household.

c) Attendance of religious services and level of religiousness/spiritualism.

4.11 Using a scatterplot matrix, repeat the problem 4.8 for fathers' measurements instead of those of the oldest child. Did you find the same pattern of relationships between body measurements and FEV1 in fathers as you did for the oldest child?

4.12 Using the mice data, create a profile plot for the average weight of mice per group over time.

Chapter 5

Data screening and transformations

5.1 Transformations, assessing normality and independence

In Section 3.4 we discussed the use of transformations to create new variables. In this chapter we discuss transforming the data to obtain a distribution that is approximately normal. This is of particular interest for exploratory data analysis. For confirmatory data analysis (as described in Chapter 1) one should choose any appropriate transformations of variables prior to performing any analyses. Section 5.2 shows how transformations change the shape of distributions. Section 5.3 discusses several methods for deciding when a transformation should be made and how to find a suitable transformation. An iterative scheme is proposed that helps to zero in on a good transformation and statistical tests for normality are evaluated. Section 5.4 presents simple graphical methods for determining if the data are independent. Section 5.5 provides an overview of the methods used in the four statistical software packages for topics discussed in this chapter. In this chapter, we rely heavily on graphical methods: see Cook and Weisberg (1994) and Tufte (2001).

Each computer software package offers the users information to help decide if their data are normally distributed. The packages provide convenient methods for transforming the data to achieve approximate normality. They also include some output for checking the independence of the observations. Hence the assumption of independent, normally distributed data that is made in many statistical tests can be assessed, at least approximately. Note that it has been shown that inference can be robust in many research settings even with highly non-normal data (see Lumley et al., 2002). Additionally, many investigators may try to discard the most obvious outliers prior to assessing normality because such outliers can grossly distort the distribution, but discarding outliers is generally not advised unless it is concluded that an error has occurred in the measurement, recording or entry of these observations. Some researchers also consider removing inconsistent or extreme observations (see Osborne and Overbay, 2004), but such a decision depends heavily on the circumstances surrounding the research topic and should always be documented.

5.2 Common transformations

For exploratory analysis of data it may be useful to transform certain variables before performing the analyses. Examples are found in the next section and in Chapter 7. In this section we present some common transformations. If you are familiar with this subject, you may wish to skip to the next section.

To develop a feel for transformations, let us examine a plot of transformed values versus the original values of the variable. To begin with, a plot of values of a variable X against itself produces a 45° diagonal line going through the origin, as shown in Figure 5.1.

One of the most commonly performed transformations is taking the **logarithm** (log) to base 10. Recall that the logarithm is the number that satisfies the relationship $X = 10^Y$. That is, the logarithm of X is the power Y to which 10 must be raised in order to produce X. As shown in Figure 5.2 in plot **a**, the logarithm of 10 is 1 since $10 = 10^1$. Similarly, the logarithm of 1 is 0 since $1 = 10^0$, and the logarithm of 100 is 2 since $100 = 10^2$. Other values of logarithms can be obtained from tables of common logarithms, from a hand calculator with a log function, or from statistical packages by