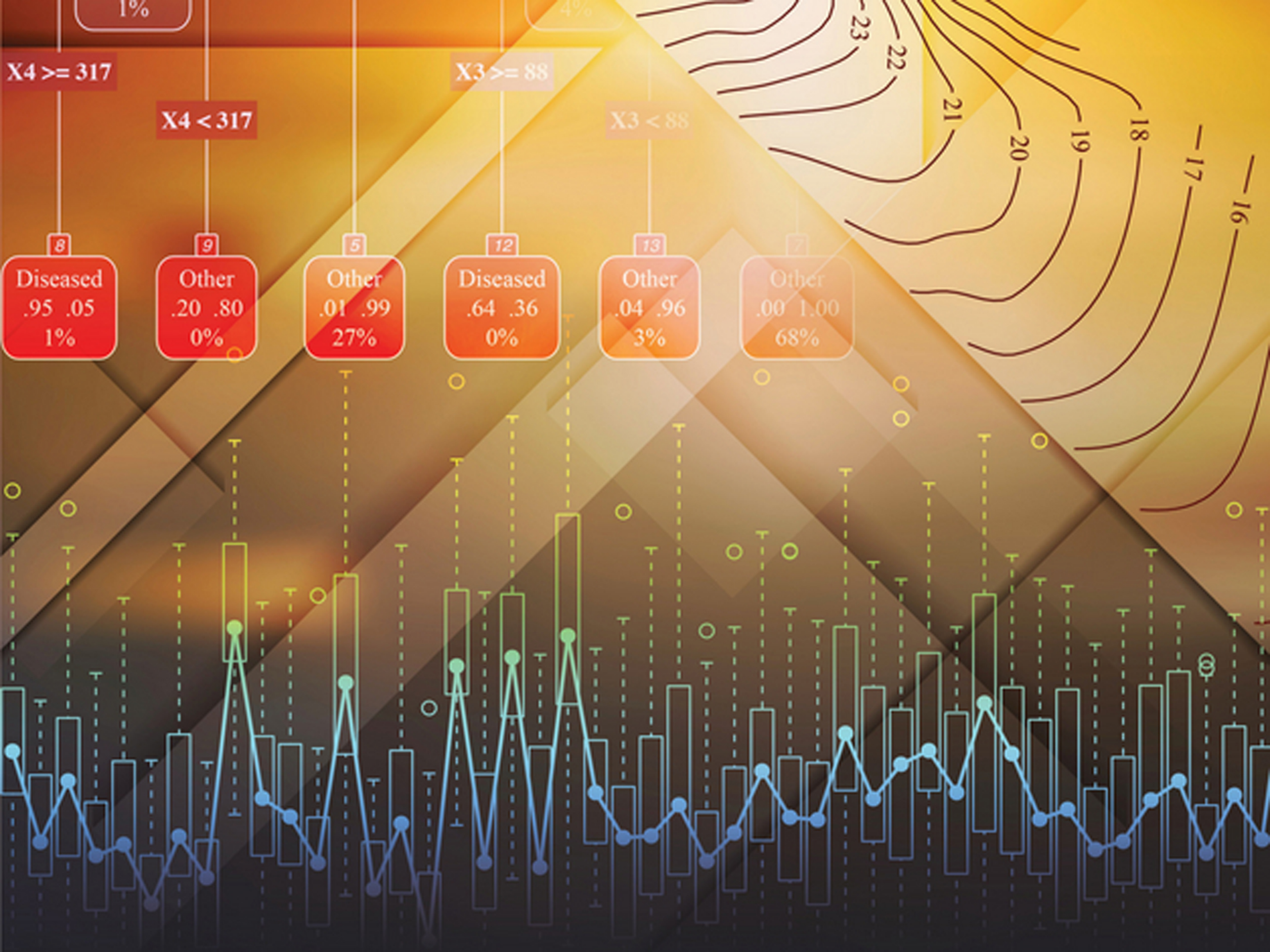

Statistical Data Analytics

Foundations for Data Mining, Informatics, and Knowledge Discovery

Walter W. Piegorsch

WILEY

Statistical Data Analytics

Statistical Data Analytics

Foundations for Data Mining, Informatics, and Knowledge Discovery

Walter W. Piegorsch

University of Arizona, USA

WILEY

This edition first published 2015

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Piegorsch, Walter W.
Statistical data analytics : foundations for data mining, informatics, and knowledge discovery / Walter W. Piegorsch.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-61965-0 (cloth : alk. paper) 1. Data mining–Mathematics. 2. Mathematical statistics. I. Title.
QA76.9.D343P535 2015
006.3′12—dc23
2015015327

A catalogue record for this book is available from the British Library.

Typeset in 10/12pt TimesLTStd by SPi Global, Chennai, India

1 2015

To Karen

Contents

Preface

Every data set tells a story. *Data analytics*, and in particular the statistical methods at their core, piece together that story's components, ostensibly to reveal the underlying message. This is the target paradigm of *knowledge discovery*: distill via statistical calculation and summarization the features in a data set/database that teach us something about the processes affecting our lives, the civilization which we inhabit, and the world around us. This text is designed as an introduction to the statistical practices that underlie modern data analytics.

Pedagogically, the presentation is separated into two broad themes: first, an introduction to the basic concepts of probability and statistics for novice users and second, a selection of focused methodological topics important in modern data analytics for those who have the basic concepts in hand. Most chapters begin with an overview of the theory and methods pertinent to that chapter's focal topic and then expand on that focus with illustrations and analyses of relevant data. To the fullest extent possible, data in the examples and exercises are taken from real applications and are not modified to simplify or "clean" the illustration. Indeed, they sometimes serve to highlight the "messy" aspects of modern, real-world data analytics. In most cases, sample sizes are on the order of 10^2–10^5, and numbers of variables do not usually exceed a dozen or so. Of course, far more massive data sets are used to achieve knowledge discovery in practice. The choice here to focus on this smaller range was made so that the examples and exercises remain manageable, illustrative, and didactically instructive. Topic selection is intended to be broad, especially among the exercises, allowing readers to gain a wider perspective on the use of the methodologies. Instructors may wish to use certain exercises as formal examples when their audience's interests coincide with the exercise topic(s).

Readers are assumed to be familiar with four semesters of college mathematics, through multivariable calculus and linear algebra. The latter is less crucial; readers with only an introductory understanding of matrix algebra can benefit from the refresher on vector and matrix relationships given in Appendix A. To review necessary background topics and to establish concepts and notation, Chapters 1–5 provide introductions to basic probability (Chapter 2), statistical description (Chapters 3 and 4), and statistical inference (Chapter 5). Readers familiar with these introductory topics may wish to move through the early chapters quickly, read only selected sections in detail (as necessary), and/or refer back to certain sections that are needed for better comprehension of later material. Throughout, sections that address more advanced material or that require greater familiarity with probability and/or calculus are highlighted with asterisks (*). These can be skipped or selectively perused on a first reading, and returned to as needed to fill in the larger picture.

The more advanced material begins in earnest in Chapter 6 with techniques for supervised learning, focusing on simple linear regression analysis. Chapters 7 and 8 follow with multiple linear regression and generalized linear regression models, respectively. Chapter 9 completes the tour of supervised methods with an overview of various methods for classification. The final two chapters give a complementary tour of methods for unsupervised learning, focusing on dimension reduction (Chapter 10) and clustering/association (Chapter 11).

Standard mathematical and statistical functions are used throughout. Unless indicated otherwise – usually by specifying a different base – log indicates the natural logarithm, so that $\log(x)$ is interpreted as $\log_e(x)$. All matrices, such as $\mathbf{X}$ or $\mathbf{M}$, are presented in bold uppercase. Vectors will usually display as bold lowercase, for example, $\mathbf{b}$, although some may appear as uppercase (typically, vectors of random variables). Most vectors are in column form, with the operator ^{T} used to denote transposition to row form. In selected instances, it will be convenient to deploy a vector directly in row form; if so, this is explicitly noted.

Much of modern data analytics requires appeal to the computer, and a variety of computer packages and programming languages are available to the user. Highlighted herein is the **R** statistical programming environment (R Core Team 2014). **R**'s growing ubiquity and statistical depth make it a natural choice. Appendix B provides a short introduction to **R** for beginners, although it is assumed that a majority of readers will already be familiar with at least basic **R** mechanics or can acquire such skills separately. Dedicated introductions to **R** with emphasis on statistics are available in, for example, Dalgaard (2008) and Verzani (2005), or online at the Comprehensive **R** Archive Network (CRAN): http://cran.r-project.org/. Also see Wilson (2012).

Examples and exercises throughout the text are used to explicate concepts, both theoretical and applied. All examples end with a ❑ symbol. Many present sample **R** code, which is usually intended to illustrate the methods and their implementation. Thus the code may not be most efficient for a given problem but should at least give the reader some inkling into the process. Most of the figures and graphics also come from **R**. In some cases, the **R** code used to create the graphic is also presented, although, for simplicity, this may only be "base" code without accentuations/options used to stylize the display.

Throughout the text, data are generally presented in reduced tabular form to show only a few representative observations. If public distribution is permitted, the complete data sets have been archived online at http://www.wiley.com/go/piegorsch/data_analytics or their online source is listed. A number of the larger data sets came from from the University of California–Irvine (UCI) Machine Learning Repository at http://archive.ics.uci.edu/ml (Frank and Asuncion, 2010); appreciative thanks are due to this project and their efforts to make large-scale data readily available.

Instructors may employ the material in a number of ways, and creative manipulation is encouraged. For an intermediate-level, one-semester course introducing the methods of data analytics, one might begin with Chapter 1, then deploy Chapters 2–5, and possibly Chapter 6 as needed for background. Begin in earnest with Chapters 6 or 7 and then proceed through Chapters 8–11 as desired. For a more complete, two-semester sequence, use Chapters 1–6 as a (post-calculus) introduction to probability and statistics for data analytics in the first semester. This then lays the foundations for a second, targeted-methods semester into the details of supervised and unsupervised learning via Chapters 7–11. Portions of any chapter (e.g., advanced subsections with asterisks) can be omitted to save time and/or allow for greater focus in other areas.

Experts in data analytics may canvass the material and ask, how do these topics differ from any basic selection of statistical methods? Arguably, they do not. Indeed, whole books can be (and have been) written on the single theme of essentially every chapter. The focus in this text, however, is to highlight methods that have formed at the core of data analytics and statistical learning as they evolved in the twenty-first century. Different readers may find certain sections and chapters to be of greater prominence than others, depending on their own scholarly interests and training. This eclectic format is unavoidable, even intentional, in a single volume such as this. Nonetheless, it is hoped that the selections as provided will lead to an effective, unified presentation.

Of course, many important topics have been omitted or noted only briefly, in order to make the final product manageable. Omissions include methods for missing data/imputation, spurious data detection, novelty detection, robust and ordinal regression, generalized additive models, multivariate regression, and ANOVA (analysis of variance, including multivariate analysis of variance, MANOVA), partial least squares, perceptrons, artificial neural networks and Bayesian belief networks, self-organizing maps, classification rule mining, and text mining, to name a few. Useful sources that consider some of these topics include (a) for missing data/imputation, Abrahantes et al. (2011); (b) for novelty detection, Pimentel et al. (2014); (c) for generalized additive models, Wood (2006); (d) for MANOVA, Huberty and Olejnik (2006); (e) for partial least squares, Esposito Vinzi and Russolillo (2013); (f) for neural networks, Stahl and Jordanov (2012); (g) for Bayesian belief networks, Phillips (2005); (h) for self-organizing maps, Wehrens and Buydens (2007); and (i) for text mining, Martinez (2010), and the references all therein. Many of these topics are also covered in a trio of dedicated texts on statistical learning – also referenced regularly throughout the following chapters – by Hastie et al. (2009), Clarke et al. (2009), and James et al. (2013). Interested readers are encouraged to peruse all these various sources, as appropriate.

By way of acknowledgments, sincere and heartfelt thanks are due numerous colleagues, including Alexandra Abate, Euan Adie, D. Dean Billheimer and the statisticians of the Arizona Statistical Consulting Laboratory (John Bear, Isaac Jenkins, and Shripad Sinari), Susan L. Cutter, David B. Hitchcock, Fernando D. Martinez, James Ranger-Moore, Martin Sill, Debra A. Stern, Hao Helen Zhang, and a series of anonymous reviewers. Their comments and assistance helped to make the presentation much more accessible. Of course, despite the fine efforts of all these individuals, some errors may have slipped into the text and these are wholly my own responsibility. I would appreciate hearing from readers who identify any inconsistencies that they may come across.

Most gracious thanks are also due the editorial team at John Wiley & Sons – Prachi Sinha Sahay, Kathryn Sharples, Heather Kay, and Richard Davies – and their LaTeX support staff led by Alistair Smith. Their patience and professionalism throughout the project's development were fundamental in helping it achieve fruition.

Walter W. Piegorsch
Tucson, Arizona
October 2014

Part I

BACKGROUND: INTRODUCTORY STATISTICAL ANALYTICS

1

Data analytics and data mining

1.1 Knowledge discovery: finding structure in data

The turn of the twenty-first century has been described as the beginning of the (or perhaps "an") Information Age, a moniker that is difficult to dismiss and likely to be understated. Throughout the period, contemporary science has evolved at a swift pace. Ever-faster scanning, sensing, recording, and computing technologies have developed which, in turn, generate data from ever-more complex phenomena. The result is a rapidly growing amount of "information." When viewed as quantitative collections, the term heard colloquially is "Big Data," suggesting a wealth of information – and sometimes disinformation – available for study and archiving. Where once computer processing and disk storage were relegated to the lowly kilobyte (1024 bytes) and megabyte (1024 KB) scales, we have moved past routine gigabyte- (1024 MB) and terabyte- (1024 GB) scale computing and now collect data on the petabyte (1024 TB) and even the exabyte (1024 PB) scales. Operations on the zettabyte scale (1024 EB) are growing, and yottabyte- (1024 ZB) scale computing looms on the horizon. Indeed, one imagines that the brontobyte (1024 YB) and perhaps geopbyte (1024 BB) scales are not far off (and may themselves be common by the time you read this).

Our modern society seems saturated by the "Big Data" produced from these technological advances. In many cases, the lot can appear disorganized and overwhelming – and sometimes it is! – engendering a sort of "quantitative paralysis" among decision makers and analysts. But we should look more closely: through clever study of the features and latent patterns in the underlying information, we can enhance decision- and policy-making in our rapidly changing society. The key is applying careful and proper analytics to the data.

At its simplest, and no matter the size, *data* are the raw material from which *information* is derived. This is only a first step, however: the information must itself be studied and its patterns analyzed further, leading to *knowledge discovery*. [An earlier term was *knowledge discovery in databases*, or "KDD" (Elder and Pregibon 1996), because the data often came from a

Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery,
First Edition. Walter W. Piegorsch.

www.wiley.com/go/piegorsch/data_analytics

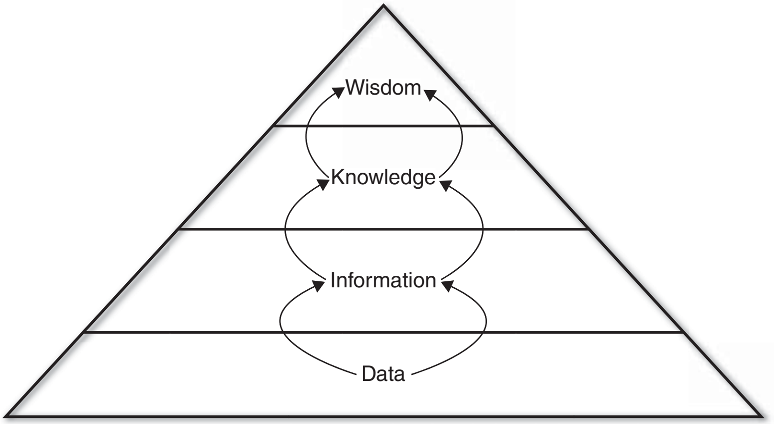

Figure 1.1 The DIKW pyramid.

database repository.] A capstone step in the process integrates and synthesizes the knowledge that has been gained on the phenomenon of interest, to produce true *wisdom* and advance the science. Thus from Data we derive Information, gaining Knowledge and producing Wisdom: D → I → K → W. The effort is sometimes described as a *DIKW pyramid* or *DIKW hierarchy* (Rowley 2007), because each step in the process represents a further refinement on the previous advance. (Some authors have derided the term, suggesting that it underemphasizes and misrepresents the complexities of synthesizing knowledge from information. Nonetheless, the DIKW pyramid still gives a useful framework for conceptualizing the knowledge- and wisdom-discovery process.) Figure 1.1 abstracts the concept.

The DIKW paradigm is by nature a multidisciplinary endeavor: computer scientists construct algorithms to manipulate and organize the data, aided by statisticians and mathematicians who instruct on development and application of quantitative methodology. Then, database experts collect and warehouse the data, software designers write programs that apply the analytic algorithms to the data, engineers build electronics and hardware to implement the programming, and subject-matter/domain experts – that is, biologists, chemists, economists, and social scientists – interpret the findings. To be successful, no one discipline or contributor can operate in a vacuum: each step in the process is enhanced by the interaction and interplay of all participants. (Indeed, the more each contributing discipline informs itself about and involves itself with the others, the more productive the DIKW effort becomes.) It is true interdisciplinarity at work, driving us to the targets, knowledge and wisdom, at the top of the pyramid.

At the base of the pyramid lies the foundation: the data. To advance successfully through each DIKW step, we must apply effective data collection, description, analysis, and interpretation. These all are the purview of statistical science, and it is the methods of modern statistical analysis that lie at the core of data analytics. Thus experience and familiarity with *statistical data analytics* has become a fundamental, necessary skill for any modern scientist dealing with Big Data. Since these methods are applied at the base of the pyramid – and often also throughout the advancing steps – this textbook views them as foundations for the DIKW process. When applied properly, and within the context of the larger interdisciplinary endeavor, features and structures in the data are revealed: for example, clinicians identify susceptible subpopulations in large databases of breast cancer patients, economists study credit card records for possible trends in purchasing behavior, sociologists track how networks develop among users of social media, and geographers catalog data on natural hazards and highlight localities with increased risk.

It is important to warn that domain-aided interpretation is a necessary component in this process: large data sets can contain structural features which when studied in greater depth represent nothing more than random noise. Teasing out real patterns from any apparent structure is often as much art as science. Nonetheless, when the analytics are successful, they facilitate our ultimate goal of knowledge discovery and advancement in wisdom.

The effort to bore through a large database studying possible patterns of response is often called *data mining*. The term conjures imagery of a miner digging through masses of rock in search of precious stones and is a surprisingly useful metaphor here. A more formal definition is “the process of seeking interesting or valuable information within large data sets” (Hand et al. 2000, p. 111). Larger still (although the two areas need not overlap) is the field of *informatics*, the study and development of quantitative tools for information processing. Many informatic subfields have emerged as data miners and analysts have specialized their focus. Examples include bioinformatics and medical informatics, ecoinformatics, geoinformatics, socioinformatics; the list grows daily! In all these areas, the data-analytic effort relies heavily on proper description, summarization, visualization, and, when necessary, inferential analysis of the collected data mass. The foundational statistical techniques for doing so are the basis of the material presented in this textbook. Some of the focus will be on issues associated with data mining, that is, how one explores collections of data statistically to identify important patterns and structure. Where pertinent, however, connections and extensions to larger applications in informatic science will also gain attention. The material is presented primarily at an introductory level, although the later chapters also give guidance on intermediate and (occasionally) advanced topics.

1.2 Data quality versus data quantity

An often-overlooked issue in data mining and data analytics is the need for sufficiently high quality in the data under study. Data miners regularly remind themselves of the GIGO principle: “if **G**arbage goes **I**n, **G**arbage come **O**ut” (Hand et al. 2001, Section 2.6). That is, the quality and value of any data mining or informatic analysis is contingent upon the quality of the underlying data. In and of itself, a large database is oftentimes an important resource; however, quantity of data does not always equate with quality of information. The data must themselves possess a level of quality commensurate with the questions they are asked to address.

This concern is worth emphasizing at an early stage in any data-analytic effort and should be kept in mind as the calculations and analyses unfold. Many informatic projects utilize data stores not under the control of the analyst or involve secondary analyses of existing databases. Thus it may not be possible to avoid or control data entry errors, coding inaccuracies, measurement mistakes, subject misidentifications, etc. If direct access and oversight is available, then some level of quality assurance/quality control (“QA/QC”) should be imposed on the data entry and curation process; see, for example, Fong (2001) or Pierchala and Surti (2009). Otherwise, potential data entry missteps or other errors in an existing database can sometimes be identified via statistical analysis, including single- or multi-dimensional graphical displays, data summarization techniques, or other forms of comparative statistical testing. (Many of these methods have more-general uses in statistical data analytics as well; these are described in the following chapters.) Of course, the analyst must also be wary of going too far: overcorrection of, say, missing data by imputing the missing values from the remainder of the database might just as quickly smooth away the very patterns the mining exercise is intended to detect.

Hand et al. (2000) distinguish between two general forms of data quality distortion: individual and collective. The first ("individual") occurs when the larger database is generally sound, but particular records in the database are affected by errors in collection, entry, or some other form of disruption. Classical examples include misplaced decimal points, transposed digits, measurement rounding errors, missing data records, and impossible combinations in classification fields (think: pregnant = "yes"/sex = "male"). These sorts of errors are often difficult to control, and the problem is common to almost any large collection of data: even the best data quality assurance program will on occasion let errors slip by. Data miners must be aware that sometimes, a feature or pattern uncovered in a large database could simply be the consequence of (a series of) individual-level data distortions. When examined in greater depth, these likely will be recognized as such and usually are afforded little value. Indeed, Hand et al. mention, only partly with tongue-in-cheek, that a large database found to be free of any errors may call into suspicion the quality of the database as a whole! More seriously though, they also note that certain patterns of distortion may in fact be of actual interest to the data miner; for example, large blocks of missing data can sometimes indicate a real, predictive classification feature in the population under study. Obviously, a kind of balancing act is required here: while outlying observations might be the purposeful target of an exercise in, say, credit-fraud detection, they more often hinder proper pattern detection in a typical data mining project (Hand et al. 2001, Section 2.7).

The second form of distortion ("collective") occurs when the larger collection suffers irregularities in the selection mechanisms under which the data were identified or sampled. Technically, data scientists define the *sampling frame* as the population of units from which a data set or database has been drawn and upon which measurements are taken/recorded. It is to this population that any inferences made from the data apply. For instance, suppose an analyst mines a database of patients suffering from a particular respiratory disease, such as asthma, in the warm and arid US Southwest. Any patterns of disease associated with, say, low-pressure weather systems gleaned from those records might not – indeed, likely will not – apply to asthma patients in more-humid, cooler north Britain/Scotland.

More generally, when data are inaccurately registered in a systematic manner, they contaminate the database and confuse the underlying sampling frame. A form of collective-scale data distortion ensues. To help to avoid collective sampling frame distortions, statistical practice encourages application of formal sampling strategies to the target population in order to construct the database. Preferred, at least at a basic level, is *simple random sampling*, where the units are sampled independently and with equally likely probabilities (see Section 3.1). By contrast, in selected instances, the database is large enough to contain the *entire* population of interest; for example, a grocery chain may collect shopping records of all its customers over a 6-month period. If so, the data now represent a full *census* of the population, and issues of sampling are less urgent. Complete enumerations of this sort are typically necessary if the informatic goal is one of fine-pattern detection across the target population.

More complex forms of probability-based sampling are also possible, although these exceed the scope here. For a deeper introduction to the theory and application of sampling methodology, see Thompson (2012) or Lohr (2010).

Of course, it is not always possible to control the sampling/selection process. In many cases, the data are recorded simply as the opportunity allows, at the convenience of the team building the database and with limited or no regard to sampling theory guidelines. This is called *convenience sampling* or *opportunity sampling*. Or, the selection process may by its very nature favor certain subjects; for instance, patients recruited for a study of genetic susceptibility to lung cancer may already be in the clinic for other disease-related reasons. (In the worst case, they all might be cigarette smokers under treatment for another, noncancerous disease such as emphysema, confounding study of the factors that lead to disease onset or progression. Upon reflection, it is perhaps obvious here that the subjects are being sampled preferentially; still, it is also surprising how often a selective process such as this goes unrecognized in practice.) The effect is known as *selection bias*, where the inclusion of a record in the database depends on what value(s) the variables take, or on some other, external, nonrandom feature (Wei and Cowan 2006). Selection bias can have a substantial confounding or contaminating effect on a large database.

Other forms of collection-level distortion include drift in the target population's attributes over time (e.g., oxygenation levels in an ecosystem's lakes may exhibit unrecognized changes due to increasing climate temperature) or overzealous data screening to expunge distortions that ends up excluding perfectly valid records. In the end, one can control for (some) data distortions via statistical adjustments in the data and/or in the analyses applied to them, but this is not always possible. At a minimum, the analyst must be aware of distorting influences on data quality in order to avoid falling victim to their ills. See Hand et al. (2000, Section 4) or Hand et al. (2001, Section 2.7) for more details and some instructive examples.

1.3 Statistical modeling versus statistical description

An important component in statistical analytics, and one that has exhibited the power of statistical science over the past century, is that of *statistical inference* (Casella and Berger 2002; Hogg and Tanis 2010). Statistical inference is the derivation of conclusions about a population from information in a random sample of that population. In many cases, formal statistical models are required to implement the inferential paradigm, using probability theory. By contrast, statistical *description* is the process of summarizing quantitative and qualitative features in a sample or population. The description process is often represented as simpler than the modeling/inferential process, but in fact, both require a level of skill and expertise beyond that of simple statistical arithmetic. A better distinction might be that inference is designed to make deductions about a feature of the population, while description is designed to bring features of a population to light.

Statistical description and statistical inference are typically applied in tandem, and the inferential process often contains descriptive aspects. They can also be employed separately, however, and it is not unusual in a data mining exercise to focus on only one of the two. For instance, an exploratory investigation of radio-transmitter data from tagged animals of a certain species may only involve simple description of their tracks and trajectories throughout a wildlife preserve. Alternatively, an inferential study on how two different species traverse the preserve might determine if a significant difference was evidenced in their trajectory patterns. In the former case, we call the effort one of *exploratory data analysis*, or "EDA," a statistical archetype popularized by Tukey (1977); more recently, see Gelman (2004) or Buja et al. (2009). The EDA approach shares similarities with many descriptive statistical methods employed in data analytics, and the two paradigms often overlap (Myatt 2007). As a result, the focus in this text will be on exploratory aspects of the data mining and knowledge discovery process, driven by statistical calculation. To provide a broader panorama, however, associated methods of statistical inference will also be considered. Chapter 2 begins with an introduction to basic probability models useful in statistical inference. Chapters 3 and 4 follow with an introduction to methods of statistical description, data manipulation, and data visualization. On the basis of these methods of probability and data description, Chapter 5 then formally introduces the inferential paradigm. Readers familiar with introductory concepts in the earlier chapters may wish to skip forward to Chapter 6 on regression techniques for supervised learning or on to further chapters where specific foundational statistical methods for data analytics and selected informatic applications are presented.

Exercises

1.1 Use an online search engine or any other means of textual search to give a list of at least three more specialized areas of “informatics”, beyond those mentioned (bioinformatics, ecoinformatics, etc.) in Section 1.1.

1.2 Give an application (from your own field of study, as appropriate) where data mining is used, and indicate instances of knowledge discovery generated from it.

1.3 Describe the nature of the database(s) from Exercise 1.2 on which the data mining was performed. What quantities were measured? What was the target population? What was/were the sampling frame/s?

1.4 Give an application (from your own field of study, as appropriate) where data distortion can occur for

(a) individual-level distortion.

(b) collective-level distortion.

1.5 As mentioned in Section 1.2, a grocery chain constructed a large database from shopping records of all its customers between January 1 and June 30 in a given year. The data only recorded each customer’s purchase(s) of (i) any cheese products at least once every month, (ii) any meat products at least once every month, and (iii) any seafood products at least once every month. It was not recorded whether the customers considered themselves vegetarians, however. Is this a form of individual-level data distortion or collective-level distortion? Justify your answer.

1.6 (Hand et al., 2000) A large database was constructed on male adult diastolic blood pressures (in mmHg). When graphed, the data showed that measurements ending in odd numbers were much more common at higher blood pressure readings. Upon deeper investigation, it was found that the pressures were taken with a digital instrument that could only display even values. When a male subject’s reading was exceptionally high, however, the technician repeated the measurement and recorded the average of the two readings. Thus although both original readings had to be even, the averaged reading could be odd. Is this a form of individual-level data distortion or collective-level distortion? Justify your answer.

1.7 To gauge students' opinions on proposed increases in statewide taxes, a polling firm sent operatives to every public college or university in their state. At each campus, the operatives stood outside the Student Union or main cafeteria just before lunch. For 30 minutes, they asked any student entering the building if he or she supported or opposed the tax increases. They also recorded the student's age, sex, and class standing (freshman, sophomore, etc.). Describe in what way(s) this can be viewed as a form of convenience sampling. Can you imagine aspects that could be changed to make it more representative and less opportunistic?

1.8 A financial firm builds a large database of its customers to study their credit card usage. In a given month, customers who had submitted at least their minimum monthly payment but less than the total amount due on that month's statement were included. By how much, if at all, the monthly payment exceeded the minimum payment level was recorded. These values were then mined for patterns using the customers' ages, lengths of patronage, etc. Is there any selection bias evident in this approach? Why or why not?

1.9 As mentioned in Section 1.2, a physician collected a large database of records on asthma patients in the US Southwest. He determined whether temporal patterns occurred in the patients' asthma onset when low-pressure weather fronts passed through the region. Is this a question of statistical description or statistical inference? Justify your answer.

1.10 Return to the asthma study in Exercise 1.9. The physician there also mined the database for associative patterns of patient proximity to construction sites where large amounts of airborne particulates were generated. Is this a question of statistical description or statistical inference? Again, justify your answer.

1.11 A geographer constructs a database on the county-by-county occurrence of natural disasters in the US Southeast over a 40-year period, along with corresponding county-level information on concurrent property damage (in $). She then uses the database to determine statistically if a difference exists in property losses due to a particular form of disaster (floods) among counties in two adjoining US states. Is this a question of statistical description or statistical inference? Why or why not?

2

Basic probability and statistical distributions

The elements of probability theory serve as a cornerstone to most, if not all, statistical operations, be they descriptive or inferential. In this chapter, a brief introduction is given to these elements, with focus on the concepts that underlie the foundations of statistical informatics. Readers familiar with basic probability theory may wish to skip forward to Section 2.3 on special statistical distributions or farther on to Chapter 3 and its introduction to basic principles of data manipulation.

2.1 Concepts in probability

Data are generated when a random process produces a quantifiable or categorical outcome. We collect all possible outcomes from a particular random process together into a set, $\mathcal{S}$, called the *sample space* or *support space*. Any subcollection of possible outcomes, including a single outcome, is called an *event*, $\mathcal{E}$. Notice that an event is technically a subset of the sample space $\mathcal{S}$. Standard set notation for this is $\mathcal{E} \subset \mathcal{S}$.

Probabilities of observing events are defined in terms of their long-term frequencies of occurrence, that is, how frequent the events (or combinations of events) occur relative to all other elements of the sample space. Thus if we generate a random outcome in a repeated manner and count the number of occurrences of an event $\mathcal{E}$, then the ratio of this count to the total number of times the outcome could occur is the probability of the event of interest. This is the *relative frequency interpretation* of probability. The shorthand for P[Observe event $\mathcal{E}$] is $P[\mathcal{E}]$ for any $\mathcal{E} \subset \mathcal{S}$. To illustrate, consider the following simple, if well recognized, example.

Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery,
First Edition. Walter W. Piegorsch.

www.wiley.com/go/piegorsch/data_analytics

Example 2.1.1 **Six-sided die roll.** Roll a fair, six-sided die and observe the number of 'pips' seen on that roll. The sample space is the set of all possible outcomes from one roll of that die: $\mathcal{S} = \{1, 2, \dots, 6\}$. Any individual event is a single number, say, $\mathcal{E} = \{6\} = \{$a roll showing 6 pips$\}$. Clearly, the single event $\mathcal{E} = \{6\}$ is contained within the larger sample space $\mathcal{S}$.

If the die is fair, then each individual event is equally likely. As there are six possible events in $\mathcal{S}$, to find $P[\mathcal{E}]$, divide 1 (for the single occurrence of $\mathcal{E}$) by 6 (for the six possible outcomes): $P[\mathcal{E}] = \frac{1}{6}$, that is, in one out of every six tosses, we expect to observe a $\{6\}$. □

2.1.1 Probability rules

A variety of fundamental axioms are applied in the interpretation of a probability $P[\mathcal{E}]$. The most well known are

(1a) $0 \le P[\mathcal{E}] \le 1$, and

(1b) $P[\mathcal{S}] = 1$.

In addition, a number of basic rules apply for combinations of two events, $\mathcal{E}_1$ and $\mathcal{E}_2$. These are

(2a) *Addition Rule*. $P[\mathcal{E}_1 \text{ or } \mathcal{E}_2] = P[\mathcal{E}_1] + P[\mathcal{E}_2] - P[\mathcal{E}_1 \text{ and } \mathcal{E}_2]$.

(2b) *Conditionality Rule*. $P[\mathcal{E}_1 \text{ given } \mathcal{E}_2] = P[\mathcal{E}_1 \text{ and } \mathcal{E}_2]/P[\mathcal{E}_2]$ for any event $\mathcal{E}_2$ such that $P[\mathcal{E}_2] > 0$. For notation, conditional probabilities are written with the symbol '|', for example, $P[\mathcal{E}_1|\mathcal{E}_2] = P[\mathcal{E}_1 \text{ given } \mathcal{E}_2]$.

(2c) *Multiplication Rule*. $P[\mathcal{E}_1 \text{ and } \mathcal{E}_2] = P[\mathcal{E}_1|\mathcal{E}_2]\ P[\mathcal{E}_2]$.

Special cases of these rules occur when the events in question relate in a certain way. For example, two events $\mathcal{E}_1$ and $\mathcal{E}_2$ that can never occur simultaneously are called *disjoint* (or equivalently, *mutually exclusive*). In this case, $P[\mathcal{E}_1 \text{ and } \mathcal{E}_2] = 0$. Notice that if two events are disjoint, the Addition Rule in (2a) simplifies to $P[\mathcal{E}_1 \text{ or } \mathcal{E}_2] = P[\mathcal{E}_1] + P[\mathcal{E}_2]$. Two disjoint events, $\mathcal{E}_1$ and $\mathcal{E}_2$, are *complementary* if the joint event $\{\mathcal{E}_1 \text{ or } \mathcal{E}_2\}$ makes up the entire sample space $\mathcal{S}$. Notice that this implies $P[\mathcal{E}_1 \text{ or } \mathcal{E}_2] = 1$. If two events, $\mathcal{E}_1$ and $\mathcal{E}_2$, are complementary so are their probabilities. This is known as the *Complement Rule*:

(2d) **Complement Rule**: If, for two disjoint events $\mathcal{E}_1$ and $\mathcal{E}_2$, the joint event $\{\mathcal{E}_1 \text{ and } \mathcal{E}_2\}$ equals the entire sample space $\mathcal{S}$, then $P[\mathcal{E}_1] = 1 - P[\mathcal{E}_2]$ and $P[\mathcal{E}_2] = 1 - P[\mathcal{E}_1]$.

Example 2.1.2 **Six-sided die roll (Example 2.1.1, continued).** Return to the roll of a fair, six-sided die. As seen in Example 2.1.1, the sample space is $\mathcal{S} = \{1, 2, \dots, 6\}$. As the die is only rolled once, no two singleton events can occur together, so, for example, observing a $\{6\}$ and observing a $\{4\}$ are disjoint events. Thus from the Addition Rule (2a) with disjoint events, $P[4 \text{ or } 6] = P[4] + P[6] = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

More involved constructions are also possible. For instance, from the Complement Rule (2d), $P[\text{not observing a } 6] = P[1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5] = 1 - P[6] = 1 - \frac{1}{6} = \frac{5}{6}$. □

The case where disjoint events completely enumerate the sample space $\mathcal{S}$ has a special name: it is called a *partition*. One need not be restricted to only two events, however. If a set of $h \ge 2$ events, $\{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_h\}$, exists such that (i) all the h events are disjoint from each other – technically, if every pair of events $\mathcal{E}_i$ and $\mathcal{E}_j$, $i \neq j$, is disjoint – and (ii) the collective event $\{\mathcal{E}_1$ and $\mathcal{E}_2$ and $\cdots$ and $\mathcal{E}_h\}$ equals $\mathcal{S}$, we say the set forms a partition of $\mathcal{S}$. (Mathematically, the partition can even consist of a countably infinite set of pairwise-disjoint events $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ if they satisfy these conditions.) This leads to another important rule from probability theory:

(2e) *The Law of Total Probability*. For any event $\mathcal{B} \subset \mathcal{S}$ and any partition, $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_h\}$, of $\mathcal{S}$,

$$P[\mathcal{B}] = \sum_{i=1}^{h} P[\mathcal{B}|\mathcal{E}_i]P[\mathcal{E}_i].$$

The Law of Total Probability in (2e) is an important building block in another famous result from probability theory, known as *Bayes' rule*. It describes how the probability of an event $P[\mathcal{E}]$ can be 'updated' using external information. The result is credited to eighteenth century Presbyterian minister Sir Thomas Bayes (Bayes 1763), although, see Stigler (1983) regarding the particulars behind that assignment. In its simplest form, Bayes' rule also shows how conditional probabilities can be reversed: by recognizing that $P[\mathcal{B}$ and $\mathcal{E}] = P[\mathcal{E}$ and $\mathcal{B}]$ and manipulating the Multiplication Rule (2c) with this fact in mind, one can show (Exercise 2.5) that

$$P[\mathcal{E}|\mathcal{B}] = P[\mathcal{B}|\mathcal{E}]\frac{P[\mathcal{E}]}{P[\mathcal{B}]}. \tag{2.1}$$

More generally, the result can be applied to full partitions of the sample space:

(2f) *Bayes' Rule*. For any event $\mathcal{B} \subset \mathcal{S}$ and any partition, $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_h\}$, of $\mathcal{S}$,

$$P[\mathcal{E}_i|\mathcal{B}] = \frac{P[\mathcal{B}|\mathcal{E}_i]P[\mathcal{B}]}{\sum_{i=1}^{h} P[\mathcal{B}|\mathcal{E}_i]}$$

for every $i = 1, \ldots, h$.

A different relationship occurs between two events if they do not impact each other in any way. Suppose the knowledge that one event $\mathcal{E}_1$ occurs has absolutely no impact on the probability that a second event $\mathcal{E}_2$ occurs and that the reverse is also true. Two such events are called *independent*. In effect, independent events modify the Conditionality Rule (2b) into $P[\mathcal{E}_1|\mathcal{E}_2] = P[\mathcal{E}_1]$ and $P[\mathcal{E}_2|\mathcal{E}_1] = P[\mathcal{E}_2]$. More importantly, for two independent events, the Multiplication Rule (2c) simplifies to $P[\mathcal{E}_1$ and $\mathcal{E}_2] = P[\mathcal{E}_1]P[\mathcal{E}_2]$.

2.1.2 Random variables and probability functions

Suppose a random outcome can be quantified formally, either because (i) it is an actual measurement or count or (ii) it is a qualitative outcome that has been unambiguously coded into a numeric value. Such a quantified random outcome is called a *random variable*. Standard notation for random variables is uppercase Roman letters, such as X or Y. To distinguish between a conceptual random variable and one that has already been realized in practice, the realized value is denoted by a lowercase Roman character: x or y. The basic probability rules for events as discussed in Section 2.1.1 can then be expanded to describe random variables.