

Making Everything Easier!™

Statistics for Big Data

FOR
DUMMIES®
A Wiley Brand

Learn to:

- Collect, clean, and interpret data
- Effectively communicate data analysis
- Make good predictions

Alan Anderson, PhD

*Finance, economics, statistics,
and math instructor*

with David Semmelroth

*Author of Data Driven Marketing
For Dummies*



Statistics For Big Data

FOR
DUMMIES[®]
A Wiley Brand

**by Alan Anderson, PhD
with David Semmelroth**

FOR
DUMMIES[®]
A Wiley Brand

Statistics For Big Data For Dummies®

Published by: **John Wiley & Sons, Inc.**, 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Copyright © 2015 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc., and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., 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. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES REPRESENTATIVES OR WRITTEN SALES MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR YOUR SITUATION. YOU SHOULD CONSULT WITH A PROFESSIONAL WHERE APPROPRIATE. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit www.wiley.com/techsupport.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2015943222

ISBN 978-1-118-94001-3 (pbk); ISBN 978-1-118-94002-0 (ePub); ISBN 978-1-118-94003-7 (ePDF)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Introduction	1
Part I: Introducing Big Data Statistics	7
Chapter 1: What Is Big Data and What Do You Do With It?	9
Chapter 2: Characteristics of Big Data: The Three Vs	19
Chapter 3: Using Big Data: The Hot Applications	27
Chapter 4: Understanding Probabilities	41
Chapter 5: Basic Statistical Ideas.....	57
Part II: Preparing and Cleaning Data	81
Chapter 6: Dirty Work: Preparing Your Data for Analysis	83
Chapter 7: Figuring the Format: Important Computer File Formats	99
Chapter 8: Checking Assumptions: Testing for Normality	107
Chapter 9: Dealing with Missing or Incomplete Data	119
Chapter 10: Sending Out a Posse: Searching for Outliers.....	129
Part III: Exploratory Data Analysis (EDA)	141
Chapter 11: An Overview of Exploratory Data Analysis (EDA).....	143
Chapter 12: A Plot to Get Graphical: Graphical Techniques	155
Chapter 13: You're the Only Variable for Me: Univariate Statistical Techniques	173
Chapter 14: To All the Variables We've Encountered: Multivariate Statistical Techniques	191
Chapter 15: Regression Analysis.....	215
Chapter 16: When You've Got the Time: Time Series Analysis.....	243
Part IV: Big Data Applications	269
Chapter 17: Using Your Crystal Ball: Forecasting with Big Data	271
Chapter 18: Crunching Numbers: Performing Statistical Analysis on Your Computer.....	297
Chapter 19: Seeking Free Sources of Financial Data.....	319
Part V: The Part of Tens	331
Chapter 20: Ten (or So) Best Practices in Data Preparation.....	333
Chapter 21: Ten (or So) Questions Answered by Exploratory Data Analysis (EDA).....	339
Index	349

Table of Contents

<i>Introduction</i>	1
About This Book	2
Foolish Assumptions	3
Icons Used in This Book	4
Beyond the Book	4
Where to Go From Here	5
 <i>Part I: Introducing Big Data Statistics</i>	7
 Chapter 1: What Is Big Data and What Do You Do With It?	9
Characteristics of Big Data	9
Exploratory Data Analysis (EDA)	10
Graphical EDA techniques	10
Quantitative EDA techniques	11
Statistical Analysis of Big Data	11
Probability distributions	12
Regression analysis	13
Time series analysis	14
Forecasting techniques	14
 Chapter 2: Characteristics of Big Data: The Three Vs	19
Characteristics of Big Data	19
Volume	20
Velocity	21
Variety	22
Traditional Database Management Systems (DBMS)	22
Relational model databases	22
Hierarchical model databases	24
Network model databases	25
Alternatives to traditional database systems	26
 Chapter 3: Using Big Data: The Hot Applications.	27
Big Data and Weather Forecasting	28
Big Data and Healthcare Services	30
Big Data and Insurance	31
Big Data and Finance	33
Big Data and Electric Utilities	34
Big Data and Higher Education	35

Big Data and Retailers.....	36
Nordstrom	36
Walmart.....	37
Amazon.com.....	37
Big Data and Search Engines.....	38
Big Data and Social Media	39
Chapter 4: Understanding Probabilities.....	41
The Core Structure: Probability Spaces	41
Discrete Probability Distributions	43
Counting outcomes.....	44
When only two things can happen:	
The binomial distribution	45
Continuous Probability Distributions.....	49
The normal distribution.....	49
Introducing Multivariate Probability Distributions	53
Joint probabilities.....	54
Unconditional probabilities.....	54
Conditional probabilities	55
Chapter 5: Basic Statistical Ideas.....	57
Some Preliminaries Regarding Data.....	57
Nominal data	58
Ordinal data.....	58
Summary Statistical Measures.....	58
Measures of central tendency.....	59
Measures of dispersion.....	63
Overview of Hypothesis Testing.....	66
The null hypothesis	66
The alternative hypothesis.....	67
The level of significance.....	67
The test statistic	68
The critical value (s)	68
To reject or not to reject, that is the question.....	69
Measures of association.....	70
Higher-Order Measures	74
Skewness.....	75
Kurtosis.....	77
Part II: Preparing and Cleaning Data.....	81
Chapter 6: Dirty Work: Preparing Your Data for Analysis.....	83
Passing the Eye Test: Does Your Data Look Correct?	84
Checking your sources.....	84
Verifying formats	85
Typecasting your data	86

Being Careful with Dates.....	87
Dealing with datetime formats.....	87
Taking geography into account	88
How your software thinks about dates.....	89
Does the Data Make Sense?.....	90
Checking discrete data.....	90
Checking continuous data	91
Frequently Encountered Data Headaches.....	93
Missing values	93
Duplicate records	95
Other Common Data Transformations	95
Percentiles	96
Standard scores	96
Dummy variables	97
Chapter 7: Figuring the Format: Important Computer File Formats . . .	99
Spreadsheet Formats	99
Comma-separated variables (.csv).....	100
Text.....	101
Microsoft Excel.....	102
Web formats	104
Database Formats.....	105
Microsoft Access (.accdb).....	105
MySQL (.frm)	106
Chapter 8: Checking Assumptions: Testing for Normality	107
Goodness of fit test.....	107
The chi-square distribution.....	108
The null and alternative hypotheses.....	109
The level of significance.....	109
Computing the test statistic	109
The critical value	113
The decision	114
Jarque-Bera test.....	115
Skewness.....	115
Kurtosis.....	115
Excess kurtosis.....	116
The null and alternative hypotheses.....	116
Computing the test statistic	116
The critical value	117
Chapter 9: Dealing with Missing or Incomplete Data	119
Missing Data: What's the Problem?.....	119
Nonresponse	120
Missingness	120

Techniques for Dealing with Missing Data	122
Deletion techniques	122
Imputation techniques	124
Expectation-maximization (EM)	127
Chapter 10: Sending Out a Posse: Searching for Outliers	129
Testing for Outliers	130
Graphical tests of outliers	131
Hypothesis tests for outliers	135
Robust Statistics	138
Dealing with Outliers	139
 <i>Part III: Exploratory Data Analysis (EDA)</i>	 141
Chapter 11: An Overview of Exploratory Data Analysis (EDA).	143
Graphical EDA Techniques	144
Box plots	145
Histograms	145
Scatter plots	146
Normal probability plots	148
EDA Techniques for Testing Assumptions	148
Run sequence plot	148
Lag plot	149
Histogram	150
Normal probability plot	151
Quantitative EDA Techniques	152
Interval estimation	152
Hypothesis testing	153
Chapter 12: A Plot to Get Graphical: Graphical Techniques	155
Stem-and-Leaf Plots	155
Scatter Plots	157
Box Plots	161
Histograms	163
Quantile-Quantile (QQ) Plots	165
Autocorrelation Plots	168
Chapter 13: You're the Only Variable for Me: Univariate Statistical Techniques	173
Counting Events Over a Time Interval: The Poisson Distribution	174
Continuous Probability Distributions	176
The Student's t-distribution	177
The lognormal distribution	181
The chi-square distribution	184
The F-distribution	187

Chapter 14: To All the Variables We've Encountered: Multivariate Statistical Techniques.191

Testing Hypotheses about Two Population Means	192
The null hypothesis for two population means	192
Alternative hypotheses for two population means	193
Level of significance	193
Test statistics and critical values for testing hypotheses about two population means.....	193
Independent populations.....	194
The decision	196
The case of dependent populations	196
Using Analysis of Variance (ANOVA) to Test Hypotheses about Population Means	199
The F-Distribution.....	204
Finding the critical values using the F-table.....	205
Making a decision	206
F-Test for the Equality of Two Population Variances	206
Null hypothesis	207
Alternative hypothesis	207
Level of significance	207
Test statistic	207
Critical values	208
Decision rule.....	208
Correlation	209
Pearson's product-moment correlation coefficient	210
Spearman's rank correlation coefficient	212

Chapter 15: Regression Analysis.215

The Fundamental Assumption: Variables Have a Linear Relationship.....	216
Defining the Population Regression Equation	216
Estimating the Population Regression Equation	217
Testing the Estimated Regression Equation	222
The coefficient of determination (R^2).....	222
Computing the coefficient of determination	224
The t-test.....	224
Using Statistical Software	228
Excel	228
Using the p-value.....	229
Assumptions of Simple Linear Regression	230
Violations of the assumptions.....	230
Multiple Regression Analysis	231
Predicting the value of Y.....	233
Testing the results of the multiple regression equation.....	234
Multicollinearity	241

Chapter 16: When You've Got the Time: Time Series Analysis 243

Key Properties of a Time Series.....	243
Trend component	244
Seasonal component	244
Cyclical component.....	244
Irregular component.....	245
Forecasting with Decomposition Methods	245
Multiplicative decomposition	245
Additive decomposition.....	247
Smoothing Techniques	247
Moving averages	248
Centered moving averages with an odd period size	250
Centered moving averages with an even period size.....	250
Exponential smoothing	252
Seasonal Components.....	255
Modeling a Time Series with Regression Analysis	257
Identifying the trend.....	258
Estimating the trend.....	260
Forecasting with time series regression	262
Estimating a quadratic trend.....	263
Comparing Different Models: MAD and MSE.....	267
Mean absolute deviation (MAD)	267
Mean square error (MSE)	268

Part IV: Big Data Applications 269

Chapter 17: Using Your Crystal Ball: Forecasting with Big Data 271

ARIMA Modeling	271
Testing for stationarity	272
Adjustments for nonstationarity.....	275
Steps used in ARIMA modeling	276
Moving average (MA) processes	279
Autoregressive (AR) processes	281
Autoregressive moving average (ARMA) processes.....	283
Autoregressive integrated moving average (ARIMA) processes	285
Simulation Techniques	286
Historical simulation	287
Monte Carlo simulation.....	290

Chapter 18: Crunching Numbers: Performing Statistical Analysis on Your Computer297

Excelling at Excel	298
Key Excel statistical functions	298
Updated statistical functions	304
Analysis ToolPak.....	307
Programming with Visual Basic for Applications (VBA)	310
R, Matey!	314

Chapter 19: Seeking Free Sources of Financial Data319

Yahoo! Finance.....	320
The ticker symbol	320
Downloading historical stock prices	321
Finding stock option prices	322
Analyzing options strategies	323
Finding key statistics for a corporation	325
Other information on Yahoo! Finance.....	325
Federal Reserve Economic Data (FRED).....	326
Finding macroeconomic data on FRED	327
Finding financial data on FRED.....	327
Board of Governors of the Federal Reserve System	328
U.S. Department of the Treasury	329
Other Useful Financial Websites.....	330

Part V: The Part of Tens 331

Chapter 20: Ten (or So) Best Practices in Data Preparation333

Check Data Formats	333
Verify Data Types	334
Graph Your Data	334
Verify Data Accuracy	334
Identify Outliers	335
Deal with Missing Values.....	335
Check Your Assumptions about How the Data Is Distributed.....	336
Back Up and Document Everything You Do.....	337

Chapter 21: Ten (or So) Questions Answered by Exploratory Data Analysis (EDA)339

What Are the Key Properties of a Dataset?	339
What's the Center of the Data?	340
How Much Spread Is There in the Data?	341
Is the Data Skewed?	343

What Distribution Does the Data Follow? 344

Are the Elements in the Dataset Uncorrelated? 345

Does the Center of the Dataset Change Over Time? 346

Does the Spread of the Dataset Change Over Time? 347

Are There Outliers in the Data? 347

Does the Data Conform to Our Assumptions? 348

***Index* 349**

Introduction

Welcome to *Statistics For Big Data For Dummies*! Every day, what has come to be known as *big data* is making its influence felt in our lives. Some of the most useful innovations of the past 20 years have been made possible by the advent of massive data-gathering capabilities combined with rapidly improving computer technology.

For example, of course, we have become accustomed to finding almost any information we need through the Internet. You can locate nearly anything under the sun immediately by using a search engine such as Google or DuckDuckGo. Finding information this way has become so commonplace that Google has slowly become a verb, as in “I don’t know where to find that restaurant — I’ll just Google it.” Just think how much more efficient our lives have become as a result of search engines. But how does Google work? Google couldn’t exist without the ability to process massive quantities of information at an extremely rapid speed, and its software has to be extremely efficient.

Another area that has changed our lives forever is e-commerce, of which the classic example is Amazon.com. People can buy virtually every product they use in their daily lives online (and have it delivered promptly, too). Often online prices are lower than in traditional “brick-and-mortar” stores, and the range of choices is wider. Online shopping also lets people find the best available items at the lowest possible prices.

Another huge advantage to online shopping is the ability of the sellers to provide reviews of products and recommendations for future purchases. Reviews from other shoppers can give extremely important information that isn’t available from a simple product description provided by manufacturers. And recommendations for future purchases are a great way for consumers to find new products that they might not otherwise have known about. Recommendations are enabled by one application of big data — the use of highly sophisticated programs that analyze shopping data and identify items that tend to be purchased by the same consumers.

Although online shopping is now second nature for many consumers, the reality is that e-commerce has only come into its own in the last 15–20 years, largely thanks to the rise of big data. A website such as Amazon.com must process quantities of information that would have been unthinkable just a few years ago, and that processing must be done quickly

and efficiently. Thanks to rapidly improving technology, many traditional retailers now also offer the option of making purchases online; failure to do so would put a retailer at a huge competitive disadvantage.

In addition to search engines and e-commerce, big data is making a major impact in a surprising number of other areas that affect our daily lives:

- ✓ Social media
- ✓ Online auction sites
- ✓ Insurance
- ✓ Healthcare
- ✓ Energy
- ✓ Political polling
- ✓ Weather forecasting
- ✓ Education
- ✓ Travel
- ✓ Finance

About This Book

This book is intended as an overview of the field of big data, with a focus on the statistical methods used. It also provides a look at several key applications of big data. Big data is a broad topic; it includes quantitative subjects such as math, statistics, computer science, and data science. Big data also covers many applications, such as weather forecasting, financial modeling, political polling methods, and so forth.

Our intentions for this book specifically include the following:

- ✓ Provide an overview of the field of big data.
- ✓ Introduce many useful applications of big data.
- ✓ Show how data may be organized and checked for bad or missing information.
- ✓ Show how to handle outliers in a dataset.
- ✓ Explain how to identify assumptions that are made when analyzing data.
- ✓ Provide a detailed explanation of how data may be analyzed with graphical techniques.

- ✓ Cover several key *univariate* (involving only one variable) statistical techniques for analyzing data.
- ✓ Explain widely used *multivariate* (involving more than one variable) statistical techniques.
- ✓ Provide an overview of modeling techniques such as regression analysis.
- ✓ Explain the techniques that are commonly used to analyze time series data.
- ✓ Cover techniques used to forecast the future values of a dataset.
- ✓ Provide a brief overview of software packages and how they can be used to analyze statistical data.

Because this is a *For Dummies* book, the chapters are written so you can pick and choose whichever topics that interest you the most and dive right in. There's no need to read the chapters in sequential order, although you certainly could. We do suggest, though, that you make sure you're comfortable with the ideas developed in Chapters 4 and 5 before proceeding to the later chapters in the book. Each chapter also contains several tips, reminders, and other tidbits, and in several cases there are links to websites you can use to further pursue the subject. There's also an online Cheat Sheet that includes a summary of key equations for ease of reference.

As mentioned, this is a big topic and a fairly new field. Space constraints make possible only an introduction to the statistical concepts that underlie big data. But we hope it is enough to get you started in the right direction.

Foolish Assumptions

We make some assumptions about you, the reader. Hopefully, one of the following descriptions fits you:

- ✓ You've heard about big data and would like to learn more about it.
- ✓ You'd like to use big data in an application but don't have sufficient background in statistical modeling.
- ✓ You don't know how to implement statistical models in a software package.

Possibly all of these are true. This book should give you a good starting point for advancing your interest in this field. Clearly, you are already motivated.

This book does not assume any particularly advanced knowledge of mathematics and statistics. The ideas are developed from fairly mundane mathematical operations. But it may, in many places, require you to take a deep breath and not get intimidated by the formulas.

Icons Used in This Book

Throughout the book, we include several icons designed to point out specific kinds of information. Keep an eye out for them:



A Tip points out especially helpful or practical information about a topic. It may be hard-won advice on the best way to do something or a useful insight that may not have been obvious at first glance.



A Warning is used when information must be treated carefully. These icons point out potential problems or trouble you may encounter. They also highlight mistaken assumptions that could lead to difficulties.



Technical Stuff points out stuff that may be interesting if you're really curious about something, but which is not essential. You can safely skip these if you're in a hurry or just looking for the basics.



Remember is used to indicate stuff that may have been previously encountered in the book or that you will do well to stash somewhere in your memory for future benefit.

Beyond the Book

Besides the pages or pixels you're presently perusing, this book comes with even more goodies online. You can check out the Cheat Sheet at www.dummies.com/cheatsheet/statisticsforbigdata.

We've also written some additional material that wouldn't quite fit in the book. If this book were a DVD, these would be on the Bonus Content disc. This handful of extra articles on various mini-topics related to big data is available at www.dummies.com/extras/statisticsforbigdata.

Where to Go From Here

You can approach this book from several different angles. You can, of course, start with Chapter 1 and read straight through to the end. But you may not have time for that, or maybe you are already familiar with some of the basics. We suggest checking out the table of contents to see a map of what's covered in the book and then flipping to any particular chapter that catches your eye. Or if you've got a specific big data issue or topic you're burning to know more about, try looking it up in the index.

Once you're done with the book, you can further your big data adventure (where else?) on the Internet. Instructional videos are available on websites such as YouTube. Online courses, many of them free, are also becoming available. Some are produced by private companies such as Coursera; others are offered by major universities such as Yale and M.I.T. Of course, many new books are being written in the field of big data due to its increasing importance.

If you're even more ambitious, you will find specialized courses at the college undergraduate and graduate levels in subject areas such as statistics, computer science, information technology, and so forth. In order to satisfy the expected future demand for big data specialists, several schools are now offering a concentration or a full degree in Data Science.

The resources are there; you should be able to take yourself as far as you want to go in the field of big data. Good luck!

Part I

Introducing Big Data Statistics

getting started
with

Big Data Statistics



Visit www.dummies.com for Great Dummies content online.

In this part . . .

- ✓ Introducing big data and stuff it's used for
- ✓ Exploring the three Vs of big data
- ✓ Checking out the hot big data applications
- ✓ Discovering probabilities and other basic statistical idea

Chapter 1

What Is Big Data and What Do You Do with It?

In This Chapter

- ▶ Understanding what big data is all about
 - ▶ Seeing how data may be analyzed using Exploratory Data Analysis (EDA)
 - ▶ Gaining insight into some of the key statistical techniques used to analyze big data
-

Big data refers to sets of data that are far too massive to be handled with traditional hardware. Big data is also problematic for software such as database systems, statistical packages, and so forth. In recent years, data-gathering capabilities have experienced explosive growth, so that storing and analyzing the resulting data has become progressively more challenging.

Many fields have been affected by the increasing availability of data, including finance, marketing, and e-commerce. Big data has also revolutionized more traditional fields such as law and medicine. Of course, big data is gathered on a massive scale by search engines such as Google and social media sites such as Facebook. These developments have led to the evolution of an entirely new profession: the *data scientist*, someone who can combine the fields of statistics, math, computer science, and engineering with knowledge of a specific application.

This chapter introduces several key concepts that are discussed throughout the book. These include the characteristics of big data, applications of big data, key statistical tools for analyzing big data, and forecasting techniques.

Characteristics of Big Data

The three factors that distinguish big data from other types of data are *volume*, *velocity*, and *variety*.

Clearly, with big data, the *volume* is massive. In fact, new terminology must be used to describe the size of these datasets. For example, one *petabyte* of data consists of 1.0×10^{15} bytes of data. That's 1,000 *trillion* bytes!



A *byte* is a single unit of storage in a computer's memory. A byte is used to represent a single number, character, or symbol. A byte consists of eight *bits*, each consisting of either a 0 or a 1.

Velocity refers to the speed at which data is gathered. Big datasets consist of data that's continuously gathered at very high speeds. For example, it has been estimated that Twitter users generate more than a quarter of a million tweets *every minute*. This requires a massive amount of storage space as well as real-time processing of the data.

Variety refers to the fact that the contents of a big dataset may consist of a number of different formats, including spreadsheets, videos, music clips, email messages, and so on. Storing a huge quantity of these incompatible types is one of the major challenges of big data.

Chapter 2 covers these characteristics in more detail.

Exploratory Data Analysis (EDA)

Before you apply statistical techniques to a dataset, it's important to examine the data to understand its basic properties. You can use a series of techniques that are collectively known as *Exploratory Data Analysis* (EDA) to analyze a dataset. EDA helps ensure that you choose the correct statistical techniques to analyze and forecast the data. The two basic types of EDA techniques are *graphical* techniques and *quantitative* techniques.

Graphical EDA techniques

Graphical EDA techniques show the key properties of a dataset in a convenient format. It's often easier to understand the properties of a variable and the relationships between variables by looking at graphs rather than looking at the raw data. You can use several graphical techniques, depending on the type of data being analyzed. Chapters 11 and 12 explain how to create and use the following:

- ✓ Box plots
- ✓ Histograms
- ✓ Normal probability plots
- ✓ Scatter plots

Quantitative EDA techniques

Quantitative EDA techniques provide a more rigorous method of determining the key properties of a dataset. Two of the most important of these techniques are

- ✓ Interval estimation (discussed in Chapter 11).
- ✓ Hypothesis testing (introduced in Chapter 5).

Interval estimates are used to create a *range* of values within which a variable is likely to fall. *Hypothesis* testing is used to test various propositions about a dataset, such as

- ✓ The mean value of the dataset.
- ✓ The standard deviation of the dataset.
- ✓ The probability distribution the dataset follows.

Hypothesis testing is a core technique in statistics and is used throughout the chapters in Part III of this book.

Statistical Analysis of Big Data

Gathering and storing massive quantities of data is a major challenge, but ultimately the biggest and most important challenge of big data is putting it to good use.

For example, a massive quantity of data can be helpful to a company's marketing research department only if it can identify the key drivers of the demand for the company's products. Political polling firms have access to massive amounts of demographic data about voters; this information must be analyzed intensively to find the key factors that can lead to a successful political campaign. A hedge fund can develop trading strategies from massive quantities of financial data by finding obscure patterns in the data that can be turned into profitable strategies.

Many statistical techniques can be used to analyze data to find useful patterns:

- ✓ Probability distributions are introduced in Chapter 4 and explored at greater length in Chapter 13.
- ✓ Regression analysis is the main topic of Chapter 15.
- ✓ Time series analysis is the primary focus of Chapter 16.
- ✓ Forecasting techniques are discussed in Chapter 17.

Probability distributions

You use a *probability distribution* to compute the probabilities associated with the elements of a dataset. The following distributions are described and applied in this book:

- ✓ **Binomial distribution:** You would use the binomial distribution to analyze variables that can assume only one of two values. For example, you could determine the probability that a given percentage of members at a sports club are left-handed. See Chapter 4 for details.
- ✓ **Poisson distribution:** You would use the Poisson distribution to describe the likelihood of a given number of events occurring over an interval of time. For example, it could be used to describe the probability of a specified number of hits on a website over the coming hour. See Chapter 13 for details.
- ✓ **Normal distribution:** The normal distribution is the most widely used probability distribution in most disciplines, including economics, finance, marketing, biology, psychology, and many others. One of the characteristic features of the normal distribution is *symmetry* — the probability of a variable being a given distance below the mean of the distribution equals the probability of it being the same distance above the mean. For example, if the mean height of all men in the United States is 70 inches, and heights are normally distributed, a randomly chosen man is equally likely to be between 68 and 70 inches tall as he is to be between 70 and 72 inches tall. See Chapter 4 and the chapters in Parts III and IV for details.

The normal distribution works well with many applications. For example, it's often used in the field of finance to describe the returns to financial assets. Due to its ease of interpretation and implementation, the normal distribution is sometimes used even when the assumption of normality is only approximately correct.

- ✓ **The Student's t-distribution:** The Student's t-distribution is similar to the normal distribution, but with the Student's t-distribution, extremely small or extremely large values are much more likely to occur. This distribution is often used in situations where a variable exhibits too much variation to be consistent with the normal distribution. This is true when the properties of small samples are being analyzed. With small samples, the variation among samples is likely to be quite considerable, so the normal distribution shouldn't be used to describe their properties. See Chapter 13 for details.

Note: The Student's t-distribution was developed by W.S. Gosset while employed at the Guinness brewing company. He was attempting to describe the properties of small sample means.

- ✓ **The chi-square distribution:** The chi-square distribution is appropriate for several types of applications. For example, you can use it to determine whether a population follows a particular probability distribution. You can also use it to test whether the variance of a population equals a specified value, and to test for the independence of two datasets. See Chapter 13 for details.
- ✓ **The F-distribution:** The F-distribution is derived from the chi-square distribution. You use it to test whether the variances of two populations equal each other. The F-distribution is also useful in applications such as regression analysis (covered next). See Chapter 14 for details.

Regression analysis

Regression analysis is used to estimate the strength and direction of the relationship between variables that are *linearly* related to each other. Chapter 15 discusses this topic at length.



Two variables X and Y are said to be *linearly* related if the relationship between them can be written in the form

$$Y = mX + b$$

where

m is the *slope*, or the change in Y due to a given change in X

b is the *intercept*, or the value of Y when $X = 0$

As an example of regression analysis, suppose a corporation wants to determine whether its advertising expenditures are actually increasing profits, and if so, by how much. The corporation gathers data on advertising and profits for the past 20 years and uses this data to estimate the following equation:

$$Y = 50 + 0.25X$$

where

Y represents the annual profits of the corporation (in millions of dollars).

X represents the annual advertising expenditures of the corporation (in millions of dollars).

In this equation, the slope equals 0.25, and the intercept equals 50. Because the slope of the regression line is 0.25, this indicates that on average, for every \$1 million increase in advertising expenditures, profits rise by \$.25 million, or \$250,000. Because the intercept is 50, this indicates that with no advertising, profits would still be \$50 million.

This equation, therefore, can be used to forecast future profits based on planned advertising expenditures. For example, if the corporation plans on spending \$10 million on advertising next year, its expected profits will be as follows:

$$Y = 50 + 0.25X$$

$$Y = 50 + 0.25(10) = 50 + 2.5 = 52.5$$

Hence, with an advertising budget of \$10 million next year, profits are expected to be \$52.5 million.

Time series analysis

A *time series* is a set of observations of a single variable collected over time. This topic is talked about at length in Chapter 16. The following are examples of time series:

- ✓ The daily price of Apple stock over the past ten years.
- ✓ The value of the Dow Jones Industrial Average at the end of each year for the past 20 years.
- ✓ The daily price of gold over the past six months.

With time series analysis, you can use the statistical properties of a time series to predict the future values of a variable. There are many types of models that may be developed to explain and predict the behavior of a time series.

One place where time series analysis is used frequently is on Wall Street. Some analysts attempt to forecast the future value of an asset price, such as a stock, based entirely on the history of that stock's price. This is known as *technical analysis*. Technical analysts do not attempt to use other variables to forecast a stock's price — the only information they use is the stock's own history.



Technical analysis can work only if there are inefficiencies in the market. Otherwise, all information about a stock's history should already be reflected in its price, making technical trading strategies unprofitable.

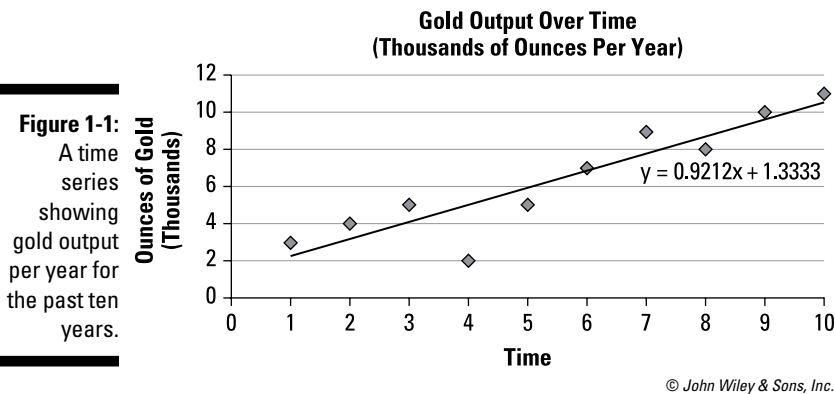
Forecasting techniques

Many different techniques have been designed to forecast the future value of a variable. Two of these are time series regression models (Chapter 16) and simulation models (Chapter 17).

Time series regression models

A *time series regression model* is used to estimate the trend followed by a variable over time, using regression techniques. A *trend line* shows the direction in which a variable is moving as time elapses.

As an example, Figure 1-1 shows a time series that represents the annual output of a gold mine (measured in thousands of ounces per year) since the mine opened ten years ago.



The equation of the trend line is estimated to be

$$Y = 0.9212X + 1.3333$$

where

X is the year.

Y is the annual production of gold (measured in thousands of ounces).

This trend line is estimated using regression analysis. The trend line shows that on average, the output of the mine grows by 0.9212 thousand (921.2 ounces) each year.

You could use this trend line to predict the output next year (the 11th year of operation) by substituting 11 for X , as follows:

$$Y = 0.9212X + 1.3333$$

$$Y = 0.9212(11) + 1.3333 = 11.4665$$

Based on the trend line equation, the mine would be expected to produce 11,466.5 ounces of gold next year.

Simulation models

You can use *simulation* models to forecast a time series. Simulation models are extremely flexible but can be extremely time-consuming to implement. Their accuracy also depends on assumptions being made about the time series data's statistical properties.

Two standard approaches to forecasting financial time series with simulation models are historical simulation and Monte Carlo simulation.

Historical simulation

Historical simulation is a technique used to generate a probability distribution for a variable as it evolves over time, based on its past values. If the properties of the variable being simulated remain stable over time, this technique can be highly accurate. One drawback to this approach is that in order to get an accurate prediction, you need to have a lot of data. It also depends on the assumption that a variable's past behavior will continue into the future.

As an example, Figure 1-2 shows a histogram that represents the returns to a stock over the past 100 days.

This histogram shows the probability distribution of returns on the stock based on the past 100 trading days. The graph shows that the most frequent return over the past 100 days was a loss of 2 percent, the second most frequent was a loss of 3 percent, and so on. You can use the information contained within this graph to create a probability distribution for the most likely return on this stock over the coming trading day.

