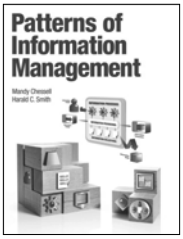
Getting Started with Data Science

Making Sense of Data with Analytics

Murtaza Haider

Related Books of Interest



Patterns of Information Management

By Mandy Chessell and Harald C. Smith

ISBN: 978-0-13-315550-1

Use Best Practice Patterns to Understand and Architect Manageable, Efficient Information Supply Chains That Help You Leverage All Your Data and Knowledge

Building on the analogy of a supply chain, Mandy Chessell and Harald Smith explain how information can be transformed, enriched, reconciled, redistributed, and utilized in even the most complex environments. Through a realistic, end-to-end case study, they help you blend overlapping information management, SOA, and BPM technologies that are often viewed as competitive.

Using this book's patterns, you can integrate all levels of your architecture—from holistic, enterprise, system-level views down to low-level design elements. You can fully address key non-functional requirements such as the amount, quality, and pace of incoming data. Above all, you can create an IT landscape that is coherent, interconnected, efficient, effective, and manageable.

The Business of IT

How to Improve Service and Lower Costs

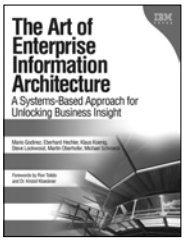
By Robert Ryan and Tim Raducha-Grace

ISBN: 978-0-13-700061-6

Drive More Business Value from IT…and Bridge the Gap Between IT and Business Leadership

IT organizations have achieved outstanding technological maturity, but many have been slower to adopt world-class business practices. This book provides IT and business executives with methods to achieve greater business discipline throughout IT, collaborate more effectively, sharpen focus on the customer, and drive greater value from IT investment. Drawing on their experience consulting with leading IT organizations, Robert Ryan and Tim Raducha-Grace help IT leaders make sense of alternative ways to improve IT service and lower cost, including ITIL, IT financial management, balanced scorecards, and business cases. You'll learn how to choose the best approaches to improve IT business practices for your environment and use these practices to improve service quality, reduce costs, and drive top-line revenue growth.

Related Books of Interest



The Art of Enterprise Information Architecture

A Systems-Based Approach for Unlocking Business Insight

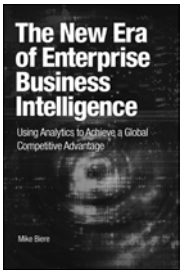
By Mario Godinez, Eberhard Hechler, Klaus Koenig, Steve Lockwood, Martin Oberhofer, and Michael Schroeck

ISBN: 978-0-13-703571-7

Architecture for the Intelligent Enterprise: Powerful New Ways to Maximize the Real-Time Value of Information

Tomorrow's winning "Intelligent Enterprises" will bring together far more diverse sources of data, analyze it in more powerful ways, and deliver immediate insight to decision-makers throughout the organization. Today, however, most companies fail to apply the information they already have, while struggling with the complexity and costs of their existing information environments.

In this book, a team of IBM's leading information management experts guide you on a journey that will take you from where you are today toward becoming an "Intelligent Enterprise."



The New Era of Enterprise Business Intelligence:

Using Analytics to Achieve a Global Competitive Advantage

By Mike Biere

ISBN: 978-0-13-707542-3

A Complete Blueprint for Maximizing the Value of Business Intelligence in the Enterprise

The typical enterprise recognizes the immense potential of business intelligence (BI) and its impact upon many facets within the organization—but it's not easy to transform BI's potential into real business value. Top BI expert Mike Biere presents a complete blueprint for creating winning BI strategies and infrastructure and systematically maximizing the value of information throughout the enterprise.

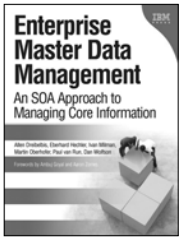
This product-independent guide brings together start-to-finish guidance and practical checklists for every senior IT executive, planner, strategist, implementer, and the actual business users themselves.

Listen to the author's podcast at:
ibmpressbooks.com/podcasts

IBM Press™

Visit ibmpressbooks.com
for all product information

Related Books of Interest



Enterprise Master Data Management

An SOA Approach to Managing Core Information

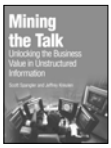
By Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul Van Run, and Dan Wolfson

ISBN: 978-0-13-236625-0

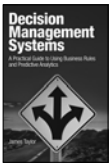
The Only Complete Technical Primer for MDM Planners, Architects, and Implementers

Enterprise Master Data Management provides an authoritative, vendor-independent MDM technical reference for practitioners: architects, technical analysts, consultants, solution designers, and senior IT decision makers. Written by the IBM® data management innovators who are pioneering MDM, this book systematically introduces MDM's key concepts and technical themes, explains its business case, and illuminates how it interrelates with and enables SOA.

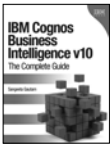
Drawing on their experience with cutting-edge projects, the authors introduce MDM patterns, blueprints, solutions, and best practices published nowhere else—everything you need to establish a consistent, manageable set of master data, and use it for competitive advantage.



Mining the Talk
Unlocking the Business Value in Unstructured Information
Spangler, Kreulen
ISBN: 978-0-13-233953-7

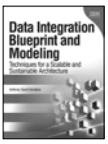


Decision Management Systems
A Practical Guide to Using Business Rules and Predictive Analytics
Taylor
ISBN: 978-0-13-288438-9



IBM Cognos Business Intelligence v10
The Complete Guide
Gautam
ISBN: 978-0-13-272472-2

IBM Cognos 10 Report Studio
Practical Examples
Draskovic, Johnson
ISBN: 978-0-13-265675-7



Data Integration Blueprint and Modeling
Techniques for a Scalable and Sustainable Architecture
Giordano
ISBN: 978-0-13-708493-7

This page intentionally left blank

Praise for *Getting Started with Data Science*

"A coauthor and I once wrote that data scientists held 'the sexiest job of the 21st century.' This was not because of their inherent sex appeal, but because of their scarcity and value to organizations. This book may reduce the scarcity of data scientists, but it will certainly increase their value. It teaches many things, but most importantly it teaches how to tell a story with data."

—Thomas H. Davenport, Distinguished Professor, Babson College;
Research Fellow, MIT; author of *Competing on Analytics* and *Big Data @ Work*

"We have produced more data in the last two years than all of human history combined. Whether you are in business, government, academia, or journalism, the future belongs to those who can analyze these data intelligently. This book is a superb introduction to data analytics, a must-read for anyone contemplating how to integrate big data into their everyday decision making."

—Professor Atif Mian, Theodore A. Wells '29 Professor of Economics and
Public Affairs, Princeton University;
Director of the Julis-Rabinowitz Center for Public Policy and Finance
at the Woodrow Wilson School; author of the best-selling book *The House of Debt*

"The power of data, evidence, and analytics in improving decision-making for individuals, businesses, and governments is well known and well documented. However, there is a huge gap in the availability of material for those who should use data, evidence, and analytics but do not know how. This fascinating book plugs this gap, and I highly recommend it to those who know this field and those who want to learn."

—Munir A. Sheikh, Ph.D., Former Chief Statistician of Canada;
Distinguished Fellow and Adjunct Professor at Queen's University

"*Getting Started with Data Science* (*GSDS*) is unlike any other book on data science you might have come across. While most books on the subject treat data science as a collection of techniques that lead to a string of insights, Murtaza shows how the application of data science leads to uncovering of coherent stories about reality. *GSDC* is a hands-on book that makes data science come alive."

—Chuck Chakrapani, Ph.D., President, Leger Analytics

"This book addresses the key challenge facing data science today, that of bridging the gap between analytics and business value. Too many writers dive immediately into the details of specific statistical methods or technologies, without focusing on this bigger picture. In contrast, Haider identifies the central role of narrative in delivering real value from big data.

"The successful data scientist has the ability to translate between business goals and statistical approaches, identify appropriate deliverables, and communicate them in a compelling and comprehensible way that drives meaningful action. To paraphrase Tukey, 'Far better an approximate answer to the right question, than an exact answer to a wrong one.' Haider's book never loses sight of this central tenet and uses many real-world examples to guide the reader through the broad range of skills, techniques, and tools needed to succeed in practical data-science.

"Highly recommended to anyone looking to get started or broaden their skillset in this fast-growing field."

—Dr. Patrick Surry, Chief Data Scientist, www.Hopper.com

Getting Started with Data Science

Making Sense of Data with Analytics

Murtaza Haider

IBM Press: Pearson plc
Boston • Columbus • Indianapolis • New York • San Francisco
Amsterdam • Cape Town • Dubai • London • Madrid • Milan • Munich
Paris • Montreal • Toronto • Delhi • Mexico City • Sao Paulo • Sidney
Hong Kong • Seoul • Singapore • Taipei • Tokyo

ibmpressbooks.com

IBM Press Program Managers: Steven M. Stansel, Natalie Troia
Cover design: IBM Corporation

Associate Publisher: Dave Dusthimer
Marketing Manager: Stephane Nakib
Executive Editor: Mary Beth Ray
Publicist: Heather Fox
Editorial Assistant: Vanessa Evans
Development Editor: Box Twelve Communications
Managing Editor: Kristy Hart
Cover Designer: Alan Clements
Senior Project Editor: Lori Lyons
Copy Editor: Paula Lowell
Senior Indexer: Cheryl Lenser
Senior Compositor: Gloria Schurick
Proofreader: Kathy Ruiz
Manufacturing Buyer: Dan Uhrig

Published by Pearson plc

Publishing as IBM Press

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact international@pearsoned.com.

Library of Congress Control Number: 2015947691

ISBN-13: 978-0-13-399102-4
ISBN-10: 0-13-399102-4

Text printed in the United States on recycled paper at R.R. Donnelley in Crawfordsville, Indiana.

First printing: December 2015

This book is dedicated to my parents, Lily and Ajaz

Contents-at-a-Glance

Contents

Preface

It was arguably the largest trading floor in Canada. Yet when it came to data and analytics, parts of the trading floor were still stuck in the digital Stone Age.

It was the mid-nineties, and I was a new immigrant in Toronto trying to find my place in a new town. I had to my credit a civil engineering degree and two years of experience as a newspaper reporter. The Fixed Income Derivatives desk at BMO Nesbitt Burns, a Toronto-based investment brokerage, was searching for a data analyst. I was good in Excel and proficient in engineering-grade Calculus. I applied for the job and landed it.

I was intimidated, to say the least, on the very first day I walked onto the sprawling trading floor. Right in the heart of downtown Toronto, a vast floor with thousands of computer screens, specialized telephone panels, huge fax machines, and men dressed in expensive suits with a telephone set in each hand, were busy buying and selling stocks, bonds, derivatives, and other financial products.

I walked over to the Derivatives desk where a small group of youngish traders greeted me with the least bit of hospitality and lots of disinterest. I was supposed to work for them.

My job was to determine the commission owed to Nesbitt Burns by other brokerages for the trades executed by the Derivatives traders. Before Futures, Options, and other securities are traded, the brokerages representing buyers and sellers draw up contracts to set commissions for the trade. Thus, for every executed trade, the agents representing buyers and sellers receive their agreed upon commissions. The Derivatives desk had fallen behind in invoicing other brokerages for their commission. My job was to analyze the trading data and determine how much other brokerages owed to Nesbitt Burns.

I was introduced to the person responsible for the task. She had been struggling with the volume of trades, so I was hired to assist her. Because she was not trained in statistics or analytics, she did what a layperson would have done with the task.

She received daily trades data in a text file that she re-keyed in Excel. This was the most time-consuming and redundant step in the process. The data identified other brokerages, the number of trades executed, and the respective commission for each block of trades executed. After she had entered the data from the printouts of the digital text files, she would then sort the data by brokerages and use Excel's **`Subtotal`** command to determine the daily commissions owed

by other brokerages. She would tabulate the entire month in roughly 30 different Excel files, and then add the owed commissions using her large accounting calculator. Finally, she would type up the invoice for each brokerage in Microsoft® Word.

I knew right away that with her methodology, we would never be able to catch up. I also knew that this could be done much faster. Therefore, I got on with the task.

I first determined that the text files containing raw data were in fixed width format that allowed me to import them directly into Excel using the `Import` Function. Once I perfected the automated importing, I asked my superiors to include a month's worth of trading data in each text file. They were able to oblige.

This improved the most time-consuming part of the process. Right there, I had eliminated the need to re-key the data. As soon as I received the text file carrying a month's worth of trade data, I would import it into Microsoft Excel.

With data imported in a spreadsheet, I was ready to analyze it. I did not sort data and then obtain subtotals for each brokerage. Instead, I used Pivot Tables, which were part of the Excel's Data Analysis toolpack. Within seconds, I had a table ready with a brokerage-by-brokerage breakdown of amounts owed. The task that took weeks in the past took me less than 10 minutes to accomplish.

I was elated. Using analytics, I had solved a problem that troubled workers at one of the most sophisticated workplaces in Canada. At the same time, I realized the other big problem I was about to create for myself. If I were to advise my employers of the breakthrough, they might have no need for me anymore. Using analytics, I had effectively made my job redundant.

I came up with a plan and shared it with one of the traders, David Barkway—who, unfortunately, died later on September 11, 2001, at the World Trade Center.[1] He was there to meet with bond traders at Cantor Fitzgerald. David was the scholarly looking, well-mannered person on the Derivatives Desk who had helped me with the project. I shared with him my breakthrough. He advised me to renegotiate my contract to include the additional responsibilities to recover the amounts. I took his advice and was fortunate to have a renegotiated contract.

Within weeks, I had switched to Microsoft Access and automated the entire task, including data import, analysis, and invoice generation. I would then fax invoices to my counterparts at other brokerages. I successfully recovered substantial amounts for Nesbitt Burns. A year later, I quit and started graduate studies at the University of Toronto.

Why I Wrote This Book

The day I left Nesbitt Burns, I knew that one day I would write a book on analytics. I did not realize then it would take me 19 years to write the book. Since Nesbitt, I have completed consulting assignments for numerous public and private sector agencies. In every consulting assignment, I walked away with the same realization that a little bit of improvement in data savviness would significantly improve the profitability of these firms.

1. Haider, M. (2011, September 11). "They killed my husband." Retrieved September 13, 2015, from http://www.dawn.com/2011/09/11/they-killed-my-husband-for-no-good-reason/

In 2005, I helped establish the Geographic Information Systems (GIS) Laboratory for the Federal Bureau of Statistics in Pakistan. The United Nations Population Fund supported the project. Imagine a national statistical agency not having the digital maps of its census geography. As a result, the agency could not perform any spatial analysis on the Census data that cost hundreds of millions of dollars to collect. With analytics, the Bureau could help improve government planning.

A few years later, I analyzed hundreds of thousands of customer complaints filed with Canada's largest public transit agency, the Toronto Transit Commission (TTC). Using analytics, I discovered that the Commission received much more customer complaints on days with inclement weather than it did otherwise. This allowed TTC to forewarn its employees and plan for customer care on days with heavy rains, snowfalls, or high-speed winds.

Similarly, I analyzed the failure rates for water pumps for a firm who had just acquired the business. The client was concerned about the higher than expected failure rates of the pumps. My analysis revealed that failure rates were higher than what was disclosed at the time the business was acquired. It allowed the firm to renegotiate the terms with the former seller to recuperate the existing and expected future losses.

Repeatedly I reached the same conclusion: If these businesses had invested in analytics, they would have determined the answers to problems they encountered. As a result, they would have adapted their business strategies in light of the insights drawn from analytics.

Over the past 15 years, we have seen tremendous advancements in data collection, data storage, and analytic capabilities. Businesses and governments now routinely analyze large amounts of data to improve evidence-based decision-making. Examples include designing more effective medicines or developing better economic policies. If analytics and data science were important 20 years ago, they are now the distinguishing factor for successful firms who increasingly compete on analytics.[2]

I am neither the first nor the only person to realize the importance of analytics. Chelsea Clinton, the daughter of former President Bill Clinton and Secretary Hillary Clinton, the 2016 presidential candidate, is also acutely aware of the importance of analytics for the success of businesses, firms, and not-for-profit organizations. At present, Chelsea Clinton is the vice chair of the Clinton Foundation, a U.S.-based NGO focused on improving global health.

She once revealed that the most important thing she learned in her graduate studies was statistics.[3] Ms. Clinton obtained a master's degree in public health from Columbia Mailman School of Public Health. Of all the training she received in the graduate program at the University, Ms. Clinton identified statistics (analytics) as the one that had the most utility for her. She credited statistical analysis software (Stata) for helping her to "absorb information more quickly and mentally sift through and catalog it."

2. Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning* (1 edition). Harvard Business Review Press.

3. Feldscher, K. (2015, April 14). Chelsea Clinton, TOMS founder Blake Mycoskie share insights on global health leadership | News | Harvard T.H. Chan School of Public Health. Retrieved September 13, 2015, from http://www.hsph.harvard.edu/news/features/chelsea-clinton-toms-founder-blake-mycoskie-share-insights-on-global-health-leadership/.

You may note that Ms. Clinton is neither a statistician nor is she aspiring to be a data scientist. Still, she believes that as a global leader in public health, she is able to perform better because of her proficiency in analytics and statistical analysis. Put simply, data science and analytics make Chelsea Clinton efficient and successful at her job where she is tasked to improve global health.

Who Should Read This Book?

While the world is awash with large volumes of data, inexpensive computing power, and vast amounts of digital storage, the skilled workforce capable of analyzing data and interpreting it is in short supply. A 2011 McKinsey Global Institute report suggests that "the United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."[4]

Getting Started with Data Science (GSDS)is a purpose-written book targeted at those professionals who are tasked with analytics, but do not have the comfort level needed to be proficient in data-driven analytics. GSDS appeals to those students who are frustrated with the impractical nature of the prescribed textbooks and are looking for an affordable text to serve as a long-term reference. GSDS embraces the 24/7 streaming of data and is structured for those users who have access to data and software of their choice, but do not know what methods to use, how to interpret the results, and most importantly, how to communicate findings as reports and presentations in print or online.

GSDS is a resource for millions employed in knowledge-driven industries where workers are increasingly expected to facilitate smart decision-making using up-to-date information that often takes the form of continuously updating data.

At the same time, the learning-by-doing approach in the book is equally suited for independent study by senior undergraduate and graduate students who are expected to conduct independent research for their coursework or dissertations.

About the Book

Getting Started with Data Science (*GSDS*) is an applied text on analytics written for professionals like Chelsea Clinton who either perform or manage analytics for small and large corporations. The text is equally appealing to those who would like to develop skills in analytics to pursue a career as a data analyst (statistician), which Google's chief economist, Hal Varian, calls the new sexiest job.[5]

4. http://www.mckinsey.com/features/big_data

5. Lohr, S. (2009, August 5). For Today's Graduate, Just One Word: Statistics. *The New York Times*. Retrieved from http://www.nytimes.com/2009/08/06/technology/06stats.html.

GSDS is uniquely poised to meet the needs for hands-on training in analytics. The existing texts have missed the largest segment of the analytics market by focusing on the extremes in the industry. Some books are too high-level as they extoll the virtues of adopting analytics with no hands-on training to experience the art and craft of data analytics. On the other hand, are the textbooks in statistics or econometrics written for senior undergraduate or graduate students? These textbooks require extensive basic knowledge and understanding of the subject matter. Furthermore, the textbooks are written for the academic audience, designed to fit a four-month semester. This structured approach may serve the academic needs of students, but it fails to meet the immediate needs of working professionals who have to learn and deliver results while being on the job full-time.

The Book's Three Key Ingredients: Narrative, Graphs, and Tables

Most books on statistics and analytics are often written by academics for students and, at times, display a disconnect with the needs of the industry. Unlike academic research, industry research delivers reports that often have only three key ingredients: namely summary tabulations, insightful graphics, and the narrative. A review of the reports produced by the industry leaders, such as PricewaterhouseCoopers, Deloitte, and large commercial banks, revealed that most used simple analytics—i.e., summary tabulations and insightful graphics to present data-driven findings. Industry reports seldom highlighted advanced statistical models or other similar techniques. Instead, they focused on creative prose that told stories from data.

GSDS appreciates the fact that most working analysts will not be required to generate reports with advanced statistical methods, but instead will be expected to summarize data in tables and charts (graphics) and wrap these up in convincing narratives. Thus, *GSDS* extensively uses graphs and tables to summarize findings, which then help weave a compelling and intriguing narrative.

The Story Telling Differentiator

This book is as much about storytelling as it is about analytics. I believe that a data scientist is a person who uses data and analytics to find solutions to problems, and then uses the findings to tell the most convincing and compelling story. Most books on analytics are tool or method focused. They are either written to demonstrate the analytics features of one or more software, or are written to highlight the capabilities of discipline-specific methods, such as data mining, statistics, and econometrics. Seldom a book attempts to teach the reader the art and craft of turning data into insightful narratives.

I believe that unless a data scientist is willing to tell the story, she will remain in a back office job where others will use her analytics and findings to build the narrative and receive praise, and in time, promotions. Storytelling is, in fact, the final and most important stage of analytics. Therefore, successful communication of findings to stakeholders is as important as conducting robust analytics.

Understanding Analytics in a 24/7 World

GSDS is written for the world awash with data where the focus is on how to turn data into insights. The chapter on data focuses on data availability in the public and private sectors. With FRED, Quandl, development banks, municipalities, and governments embracing the open data paradigm, opportunities to perform innovative analytics on current and often real-time data are plenty. The book expects readers to be keen to work with real-world data.

The book is organized to meet the needs of applied analysts. Instead of subjecting them to hundreds of initial pages on irrelevant background material, such as scarcely used statistical distributions, *GSDS* exposes the readers to tools they need to turn data into insights.

Furthermore, each chapter focuses on one or more research questions and incrementally builds upon the concepts, repeating worked-out examples to illustrate key concepts. For instance, the chapter on modeling binary choices introduces binary logit and probit models. However, the chapter repeatedly demonstrates the most commonly used tools in analytics: summary tables and graphics.

The book demonstrates all concepts for the most commonly used analytics software, namely R, Stata, SPSS®, and SAS. While each chapter demonstrates one, or at times two, software, the book's accompanying website (www.ibmpressbooks.com/title/9780133991024) carries additional, detailed documentation on how to perform the same analysis in other software. This ensures that the reader has the choice to work with her preferred analytics platform.

The book relies on publically available market research survey data from several sources, including PEW Research Center. The PEW global data sets offer survey data on 24 countries, including the U.S. (http://www.pewglobal.org/). The data sets offer information on a variety of topics, ranging from the use of social media and the Internet to opinions about terrorism. I use Pew data and other similar global data sets, which would be appealing to readers in the U.S. and other countries.

A Quick Walkthrough of the Book

Chapter 1, "The Bazaar of Storytellers," establishes the basic definitions of what data science is and who is a data scientist. It may surprise you to know that individuals and corporations are battling over who is a data scientist. My definition is rather simple and straightforward. If you analyze data to find solutions to problems and are able to tell a compelling story from your findings, you are a data scientist. I also introduce in Chapter 1 the contents of the book in some detail. I conclude the first chapter by answering questions about data science as a career.

Chapter 2, "Data in the 24/7 Connected World," serves as an introduction to the brave new world of data. It is astonishing to realize that until recently we complained about the lack of sufficient data. Now we face data deluge. The Internet of (every) things and wearable and ubiquitous computing have turned human beings into data generation machines. The second chapter thus provides a bird's eye view of the data world. I identify sources of propriety and open data. I also offer an introduction to big data, an innovation that has taken the business world by storm.

I, however, warn about the typical pitfall of focusing only on the size and not the other equally relevant attributes of the data we collectively generate today. The coverage of big data is rather brief in this chapter. However, I point the reader to a 2015 paper I co-wrote on big data in the *International Journal of Information Management*, which provides a critical review of the hype around big data.[6] Lastly, I introduce readers to survey-based data.

Chapter 3, "The Deliverable," focuses on the differentiator: storytelling. I do not consider data science to be just about big or small data, algorithms, coding, or engineering. The strength of data science lies in the power of the narrative. I first explain the systematic process one may adopt for analytics. Later, I present several examples of writings (copies of my syndicated published blogs) to illustrate how you can tell stories with data and analytics.

Chapter 4, "Serving Tables," incrementally evolves the discussion on generating summary tables from very basic tables to rather advanced, multidimensional tables. I highlight the need for effective communication with tabular summaries by urging data scientists to avoid ambiguous labeling and improper formatting.

Chapter 5, "Graphic Details," introduces readers to a systematic way of generating illustrative graphics. Graphics are uniquely suited to present the interplay between several variables to help communicate the complex interdependencies that often characterize the socioeconomic questions posed to data scientists. Using real-world data sets, including the data on *Titanic*'s survivors, I illustrate how one can generate self-explaining and visually pleasing graphics. I also demonstrate how graphics help tease out complex, and at times latent, relationships between various phenomena.

Chapter 6, "Hypothetically Speaking," formally introduces readers to hypothesis testing using the standard statistical methods, such as `t-tests` and correlation analysis. Often we are interested in determining whether the differences we observe in data are real or merely a matter of chance. The chapter first introduces the very fundamental building blocks of statistical thought, probability distributions, and then gradually builds on the discussion by demonstrating hypothesis testing.

All else being equal is the focus of Chapter 7, "Why Tall Parents Don't Have Even Taller Children." Using regression analysis, I demonstrate that only after controlling for other relevant factors, we are able to isolate the influence of the variable of interest on the phenomenon we analyze. Chapter 7 begins by remembering Sir Frances Galton's pioneering work that introduced regression models as a tool for scientific inquiry. I use examples from housing markets, consumer spending on food and alcohol, and the relationship between teaching evaluations and instructors' looks to illustrate regression models.

Chapter 8, "To Be or Not to Be," introduces the tools to analyze binary or dichotomous variables. I first demonstrate that ordinary least squares (OLS) regression models are not ideally suited to analyze binary dependent variables. I then introduce two alternatives: logit and probit

6. Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

models. I also demonstrate the use of grouped logit models using sample Census data on public transit ridership in and around New York City. I explain interpreting logit and probit models using marginal effects and graphics.

Chapter 9, "Categorically Speaking About Categorical Data," begins with a brief refresher on binary logit models. I then introduce multinomial variables, which are categorical variables with more than two categories. The bulk of the discussion focusses on how to analyze multinomial variables. A key feature of Chapter 9 is the use of multiple data sets to estimate a variety of models. Finally, I illustrate how to estimate conditional logit models that use the attributes of choice and the characteristics of the decision-maker as explanatory variables.

Chapter 10, "Spatial Data Analytics," addresses the oft-neglected aspect of data science: spatial analytics. The emergence of Geographic Information Systems (GIS) has enabled spatial analytics to explore the spatial interdependencies in data. The marriage between the GIS software and demographic/spatial data has improved to market research and data science practice. I illustrate how sophisticated algorithms and tools packaged in affordable or freeware GIS software can add spatial analytical capabilities to a data scientist's portfolio.

Chapter 11, "Doing Serious Time with Time Series," presents a detailed discussion of the theory and application of time series analysis. The chapter first introduces the types of time series data and its distinctive features, which are the trend, seasonality, autocorrelation, lead, and lags. The chapter shows how one can adopt the OLS regression model to work with time series data. I then introduce the advanced time series modeling techniques and their applications to forecasting housing markets in Canada and the U.S.

The final Chapter 12, "Data Mining for Gold," serves as an introduction to the rich and rapidly evolving field of data mining. The topics covered are a very small subset of techniques and models being used for data mining purposes. The intent is to give readers a taste of some commonly used data mining techniques. I believe that over the next few years, data-mining algorithms are likely to experience a revolution. The availability of large data sets, inexpensive data storage capabilities, and advances in computing platforms are all set to change the way we go about data analysis and data mining.

Acknowledgments

With aging parents in Pakistan and a young family in North America, I, like millions of other immigrants, struggle to balance parts of my life that are spread across oceans. Immigrants might be CEOs, academics, doctors, and engineers; to a varying degree, they all struggle to take care of those who once took care of them. My circumstances are therefore not unique.

This book is a few months behind schedule. I have been flying across the Atlantic to assist my siblings and their spouses who have been the primary caregivers for my parents. This has also allowed me the opportunity to be near my mother, who has had the most profound impact on my aspirations and dreams. She died in August 2015 after battling illnesses with courage and dignity. Exactly a month later, my father also passed away.

Raising a child is a challenge irrespective of place and circumstances. Hilary Clinton reminds us that "it takes a village." But raising a physically disabled child in an impoverished country with poor health care poses extraordinary challenges. My mother, though, was relentless to beat the odds to ensure that I grew up a confident and productive member of society, irrespective of my abilities and disabilities.

I was born in 1969. A birth trauma severely affected my right arm and hand. My parents tried for years to reverse the damage, but in vain. Back then, the medical technology and expertise in Pakistan was inadequate. As a result, I grew up with only one functional arm.

My mother decided that she would not let this setback change her plans for me. She had the good fortune of being an educated woman. She was a professor of literature, a writer, and a poet. She, like my father, was also a fearless individual. As an academic, she ensured that I inherited her love for Persian and Urdu poetry.

My mother was born in the British India. A British nurse, who attended her birth, called her Lily. That became her nickname, which soon became her primary identity.

Lily grew up in one of the most conservative parts of this world, and yet she tirelessly pursued higher education and dedicated her life to spreading it. She was among the pioneers who founded several institutions of higher learning to educate young women in Pakistan's most conservative Frontier province.

This book and I owe a mountain of debt to her. She instilled in me a sense of duty to others. She wanted me to dream, write, and lead. I did that and more because she enabled me to look beyond my disabilities.

My father, Ajaz, is the other person from whom I have not just inherited genes, but also great values. Like my mother, he was also a professor of literature. I inherited his love for books and the respect for the published word. My father was an incorruptible man who cherished how little he needed rather than what material wealth he possessed. He ensured that my siblings and I inherit his sense of fairness and social justice.

It was a privilege just to know them, and my good fortune to have them as my loving parents.

Along the way, hundreds, if not thousands, of individuals, including teachers, neighbors, colleagues, mentors, and friends, have made remarkable contributions to my intellectual growth. I am indebted to all and would like to acknowledge those who influenced this book.

I came formally to statistics rather late in my academic life. My formal introduction to statistical analysis happened at the graduate school when I took a course in engineering probability and statistics with Professor Ezra Hauer, an expert in traffic safety at the University of Toronto.[7]

It did not take Professor Hauer long to realize that I was struggling in his course. He brought a set of VHS tapes on introductory statistics for me. He ensured that I watched the videos and other reading material he selected for me. I most certainly would have failed his course if it were not for his dedicated efforts to help me learn statistics.

Several other academics at the University of Toronto were also instrumental in developing my interest in analytics. Professor Eric Miller introduced me to discrete choice models. Professors Ferko Csillag (late) and Carl Amrhein introduced me to spatial analytics and Geographic Information Systems (GIS). Professor Richard DiFrancesco helped me with input-output models.

I worked on the book during a sabbatical leave from Ryerson University. I am grateful to Ryerson University for providing me with the opportunity to concentrate on the book. Several individuals at Ryerson University have influenced my thoughts and approach to analytics. Professor Maurice Yeates, the former Dean of graduate studies, and Professor Ken Jones, former Dean of the Ted Rogers School of Management, have mentored me over the past 15 years in spatial statistical analysis. I served briefly as their graduate research assistant during doctoral studies. Later, as their colleague at Ryerson University, I continued to benefit from their insights.

I had the privilege of working along Dr. Chuck Chakrapani at Ryerson University. Chuck was a senior research fellow at Ryerson. Currently, he is the President of Leger Analytics, a leading market research firm in Canada. Chuck is a prolific writer and the author of several bestselling texts on market research. I benefited from dozens of conversations with him on statistical methods.

I am also grateful to Kevin Manuel, who is a data librarian at Ryerson University. Kevin has been a strong supporter of my analytics efforts at Ryerson University.

I have been fortunate to receive generous support from the analytics industry. Dr. Howard Slavin and Jim Lam of Caliper Corporation ensured that I always had enough licenses for GIS software to learn and teach spatial analytics. Later, colleagues at Pitney Bowes, makers of MapInfo and other customer analytics software, donated significant data and software to support my

7. Huer, E. (2015). *The Art of Regression Modeling in Road Safety.* Springer.

research and teaching. Dr. Mark Smith, who earlier managed Pitney Bowes' Portrait Miner, Dr. Robert Cordery, Dr. John Merola, Ali Tahmasebi, and Hal Hopson are some of the amazing colleagues at Pitney Bowes who have supported and influenced my work on spatial analytics.

I am also grateful to Professor Luc Anselin and his team for creating several tools to undertake spatial analytics. I have illustrated applications of Geoda for spatial analytics in Chapter 10. I have greatly benefitted from Professor Anselin's training workshops in spatial analytics, which he conducted at the University of Michigan in Ann Arbor.

I am indebted to numerous authors whose works have inspired me over the years. In econometrics, I am grateful to Professors Damodar Gujarati, Jeffery Wooldridge, and William Greene for their excellent texts that I have relied on over the past 20 years.

I have used R extensively in the book. R is a collective effort of thousands of volunteers. I owe them tons of gratitude. Some I would like to mention. I am grateful to Professor John Fox of McMaster University for creating R Commander, an intuitive GUI for R. The amazing contribution to graphics in R by Deepyan Sarkar (Lattice package) and Hadley Wickham (ggplot2, dplyr, and several other packages) is also recognized. Other authors of texts on R and statistical modeling, who are too numerous to list, are collectively acknowledged for helping me understand analytics.

I had the pleasure of working with several editors for this book. Professor Amar Anwar at the Cape Breton University and Professor Sumeet Gulati at the University of British Columbia (UBC) have been great friends. Along with Alastair Fraser of the UBC, Sumeet and Amar went over the draft and suggested many improvements. I am grateful for their diligence and support.

I am also grateful to Rita Swaya, Liam Donaldson, and Ioana Moca for their help with editing.

Several colleagues at IBM® have been instrumental in getting this book published. Raul Chong and Leon Katsnelson of IBM Canada were the first two IBM colleagues who championed the idea for this book. They introduced me to Susan Visser and Steven Stansel at IBM Publishing, who then introduced me to colleagues at Pearson North America. I am grateful to colleagues at IBM and Pearson who shared my enthusiasm for this book and favorably reviewed the book proposal.

Mary Beth Ray, Executive Editor at Pearson, guided me through the complex task of authoring a book. She has been a great support during the writing and editing phases. Chris Cleveland and Lori Lyons at Pearson North America also provided abundant help with technical matters and editing.

Jeff Riley, Paula A. Lowell, Lori Lyons, and Kathy Ruiz have combed through this book. Once they go through a chapter, I receive it back with thousands of big and small edits. I have learned much about editing and publishing by merely responding to their edits. Their attention to detail and commitment to quality has helped improve this book.

I would like to express my gratitude to Antoine Haroun, CIO at Mohawk College, who has been a steadfast friend over the past two decades and a strong supporter of this book. He kept cheering me on to ensure that I complete the manuscript. Also, I am profoundly grateful to my brothers and their families for attending to our parents, providing me the opportunity to work on this book.

Over the past year, my family has made numerous sacrifices to ensure that this book becomes a reality. My wife, Sophia, has sacrificed the most to help me get this book completed. She brings the sharp eye and the no-nonsense approach of a banker to my ideas. To see an example, you might want to skip to the section on the "Department of Obvious Conclusions" in Chapter 7. Sophia remains the foremost editor and critic of my writings and the original source of numerous ideas I have articulated. Our sons, Mikael (7) and Massem (5), saw me type through many afternoons and evenings. I even smuggled my work on family vacations. I am not proud of it.

Without my family's support and love, I would not have been able to follow my passions in life: analytics and writing.

About the Author

Murtaza Haider, Ph.D., is an Associate Professor at the Ted Rogers School of Management, Ryerson University, and the Director of a consulting firm Regionomics Inc. He is also a visiting research fellow at the Munk School of Global Affairs at the University of Toronto (2014-15). In addition, he is a senior research affiliate with the Canadian Network for Research on Terrorism, Security, and Society, and an adjunct professor of engineering at McGill University.

Haider specializes in applying analytics and statistical methods to find solutions for socio-economic challenges. His research interests include analytics; data science; housing market dynamics; infrastructure, transportation, and urban planning; and human development in North America and South Asia. He is an avid blogger/data journalist and writes weekly for the *Dawn* newspaper and occasionally for the *Huffington Post*.

Haider holds a Masters in transport engineering and planning and a Ph.D. in Urban Systems Analysis from the University of Toronto.

This page intentionally left blank

This page intentionally left blank

CHAPTER 6

Hypothetically Speaking

The Great Recession in 2008 wiped clean the savings portfolios of hundreds of millions in North America and Europe. Before the recession, people like columnist Margaret Wente, who were fast approaching retirement, had a 10-year plan. But then a "black swan pooped all over it."[1]

Nassim Nicholas Taleb, a New York-based professor of finance and a former finance practitioner, used the black swan analogy in his book of the same title to explain how past events are limited in forecasting the future.[2] He mentions the surprise of the European settlers when they first arrived in Western Australia and spotted a black swan. Until then, Europeans believed all swans to be white. However, a single sighting of a black swan changed that conclusion forever. This is why Professor Taleb uses the black swan metaphor to highlight the importance of extremely rare events, which the data about the past might not be able to forecast. The same phenomenon is referred to as the "fat tails" of probability distributions that explain the odds of the occurrence of rare events.

The fat tails of probability distributions resulted in trillions of dollars of financial losses during the 2007–08 financial crisis. In a world where almost nothing looks and feels normal, the empirical analysis is rooted in a statistical model commonly referred as the Normal distribution. Because of the ease it affords, the Normal distribution and its variants serve as the backbone of statistical analysis in engineering, economics, finance, medicine, and social sciences. Most readers of this book are likely to be familiar with the bell-shaped symmetrical curve that stars in every text on statistics and econometrics.

Simply stated, the Normal distribution assigns a probability to a particular outcome from a range of possible outcomes. For instance, when meteorologists advise of a 30% chance of rainfall, they are relying on a statistical model to forecast the likelihood of rain based on past data. Such models are usually good in forecasting the likelihood of events that have occurred more frequently in the past. For instance, the models usually perform well in forecasting the likelihood of a small change in a stock market's value. However, the models fail miserably in forecasting

large swings in the stock market value. The rarer an event, the poorer will be the model's ability to forecast its likelihood to occur. This phenomenon is referred to as the fat tail where the model assigns a very low, sometimes infinitely low, possibility of an event to occur. However, in the real world, such extreme events occur more frequently.

In *Financial Risk Forecasting*, Professor Jon Danielsson illustrates fat tails using the Great Recession as an example.[3] He focuses on the S&P 500 index returns, which are often assumed to follow Normal distribution. He illustrates that during 1929 and 2009, the biggest one-day drop in S&P 500 returns was 23% in 1987. If returns were indeed Normally distributed, the probability of such an extreme crash would be 2.23×10^{-97} Professor Danielsson explains that the model predicts such an extreme crash to occur only once in 10^{95} years. Here lies the problem. The earth is believed to be roughly 10^7 years old and the universe is believed to be 10^{13} years old. Professor Danielsson explains that if we were to believe in Normal distribution, the 1987 single day-crash of 23% would happen in "once out of every 12 universes."

The reality is that the extreme fluctuations in stock markets occur much more frequently than what the models based on Normal distribution suggest. Still, it continues to serve as the backbone of empirical work in finance and other disciplines.

With these caveats, I introduce the concepts discussed in this chapter. Most empirical research is concerned with comparisons of outcomes for different circumstances or groups. For instance, we are interested in determining whether male instructors receive higher teaching evaluations than female instructors. Such analysis falls under the umbrella of hypothesis testing, which happens to be the focus of this chapter.

I begin by introducing the very basic concepts of random numbers and probability distributions. I use the example of rolling two dice to introduce the fundamental concepts of probability distribution functions. I then proceed to a formal introduction of Normal and t-distributions, which are commonly used for statistical models. Finally, I explore hypothesis testing for the comparison of means and correlations.

I use data for high-performing basketball players to compare their career performances using statistical models. I also use the teaching evaluations data, which I have introduced in earlier chapters, to compare means for two or more groups.

Random Numbers and Probability Distributions

Hypothesis testing has a lot to do with probability distributions. Two such distributions, known as the normal or Gaussian distribution and the t-distribution, are frequently used. I restrict the discussion about probability distributions to these frequently used distributions.

Probability is a measure between zero and one of the likelihood that an event might occur. An event could be the likelihood of a stock market falling below or rising above a certain threshold. You are familiar with the weather forecast that often describes the likelihood of rainfall in

terms of probability or chance. Thus, you often hear the meteorologists explain that the likelihood of rainfall is, for instance, 45%. Thus, 0.45 is the probability that the event, rainfall, might occur.

The subjective definitions of probability might be expressed as the probability of one's favorite team winning the World Series or the probability that a particular stock market will fall by 10% in a given time period. Probability is also described as outcomes of experiments. For instance, if one were to flip a fair coin, the outcome of head or tail can be explained in probabilities. Similarly, the quality control division of a manufacturing firm often defines the probability of a defective product as the number of defective units produced for a certain predetermined production level.

I explain here some basic rules about probability calculations. The probability associated with any outcome or event must fall in the zero and one (0–1) interval. The probability of all possible outcomes must equate to one.

Tied with probability is the concept of randomness. A random variable is a numerical description of the outcome of an experiment. Random variables could be discrete or continuous. Discrete random variables assume a finite or infinite countable number of outcomes. For instance, the imperfections on a car that passes through the assembly line or the number of incorrectly filled orders at a grocery store are examples of random variables.

A continuous random variable could assume any real value possible within some limited range. For instance, if a factory produces and ships cereal in boxes, the average weight of a cereal box will be a continuous random variable. In finance, the daily return of a stock is a continuous variable.

Let us build on the definition of probability and random variables to describe probability distributions. A probability distribution is essentially a theoretical model depicting the possible values a random variable may assume along with the probability of occurrence. We can define probability distributions for both discrete and continuous random variables.

Consider the stock for Apple computers for the period covering January 2011 and December 2013. During this time, the average daily returns equaled 0.000706 with a standard deviation of 0.01774. I have plotted a histogram of daily return for the Apple stock and I have overlaid a normal distribution curve on top of the histogram shown in Figure 6.1. The bars in the histogram depict the actual distribution of the data and the theoretical probability distribution is depicted by the curve. I can see from the figure that the daily returns equaled zero or close to zero more frequently than depicted by the Normal distribution. Also, note that some negative daily returns were unusually large in magnitude, which are reflected by the values to the very left of the histogram beyond –0.05. I can conclude that while the Normal distribution curve assigns very low probability to very large negative values, the actual data set suggests that despite the low theoretical probability, such values have realized more frequently.

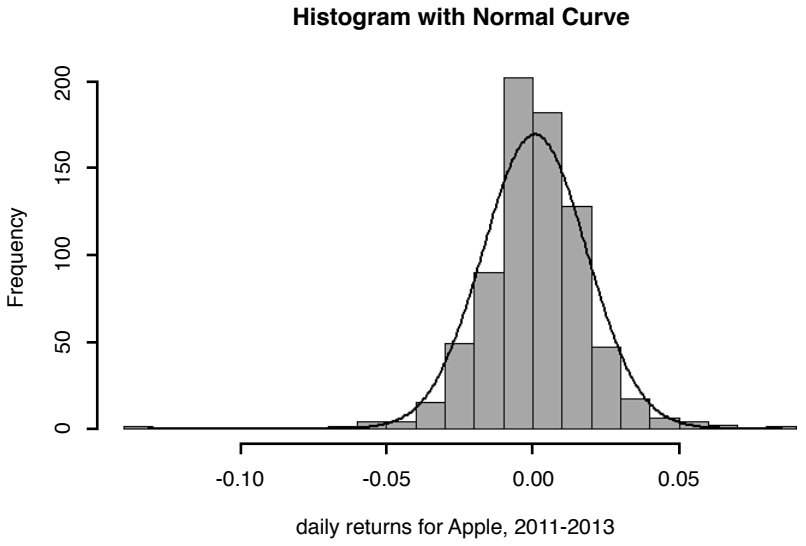


Figure 6.1 Histogram of daily Apple returns with a normal curve

Casino Royale: Roll the Dice

I illustrate the probability function of a discrete variable using the example of rolling two fair dice. A die has six faces, so rolling two dice can assume one of the 36 discrete outcomes, because each die can assume one of the six outcomes in a roll. Hence rolling two dice together will return one out of 36 outcomes. Also, note that if one (1) comes up on each die, the outcome will be 1 + 1 = 2, and the probability associated with this outcome is one out of thirty-six (1/36) because no other combination of the two dice will return two (2). On the other hand, I can obtain three (3) with the roll of two dice by having either of the two dice assume one and the other assuming two and vice versa. Thus, the probability of an outcome of three with a roll of two dice is 2 out of 36 (2/36).

The 36 possible outcomes obtained from rolling two dice are illustrated in Figure 6.2.

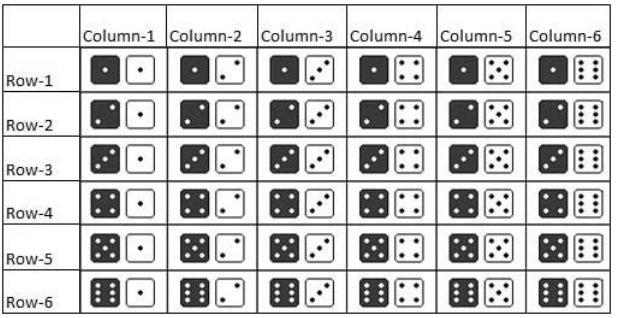


Figure 6.2 All possible outcomes of rolling two dice

Source: http://www.edcollins.com/backgammon/diceprob.htm

Based on the possible outcomes of rolling two dice, I can generate the probability density function. I present the outcomes and the respective probabilities in the Probability column in Table 6.1.

Table 6.1 Probability Calculations for Rolling Two Dice

Sum of Two Dice, x	f(x)	Probability	F(x)	Prob <=x	Prob > x
2	1/36	0.028	1/36	0.028	0.972
3	2/36	0.056	3/36	0.083	0.917
4	3/36	0.083	6/36	0.167	0.833
5	4/36	0.111	10/36	0.278	0.722
6	5/36	0.139	15/36	0.417	0.583
7	6/36	0.167	21/36	0.583	0.417
8	5/36	0.139	26/36	0.722	0.278
9	4/36	0.111	30/36	0.833	0.167
10	3/36	0.083	33/36	0.917	0.083
11	2/36	0.056	35/36	0.972	0.028
12	1/36	0.028	1	1	0

The cumulative probability function F(x) specifies the probability that the random variable will be less than or equal to some value x. For the two dice example, the probability of obtaining five with a roll of two dice is 4/36. Similarly, the cumulative probability of 10/36 is the probability that the random variable will be either five or less (Table 6.1).

Figure 6.3 offers a vivid depiction of probability density functions. Remember that the probability to obtain a certain value for rolling two dice is the ratio of the number of ways that particular value can be obtained and the total number of possible outcomes for rolling two dice (36). The highest probable outcome of rolling two dice is 7, which is $\frac{6}{36} = .167$

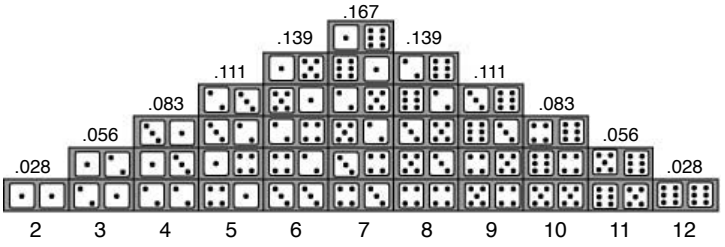


Figure 6.3 Histogram of the outcomes of rolling two dice

Source: http://hyperphysics.phy-astr.gsu.edu/hbase/math/dice.html

I have plotted the probability density function and the cumulative distribution function of the discrete random variable representing the roll of two dice in Figure 6.4 and Figure 6.5. Notice again that the probability of obtaining seven as the sum of rolling two dice is the highest and the probability of obtaining 2 or 12 are the lowest. Probability density function is a continuous function that describes the probability of outcomes for the random variable X. A histogram of a random variable approximates the shape of the underlying density function.

Figure 6.5 depicts an important concept that I will rely on this chapter. The figure shows the probability of finding a particular value or less from rolling two dice, also known as the cumulative distribution function. For instance, the probability of obtaining four (4) or less from rolling two dice is 0.167. Stated otherwise, the probability of obtaining a value greater than four (4) from rolling two dice is 0.833.

Recall Figure 6.1, which depicted the histogram of the daily returns for Apple stock. The shape of the histogram approximated the shape of underlying density function.

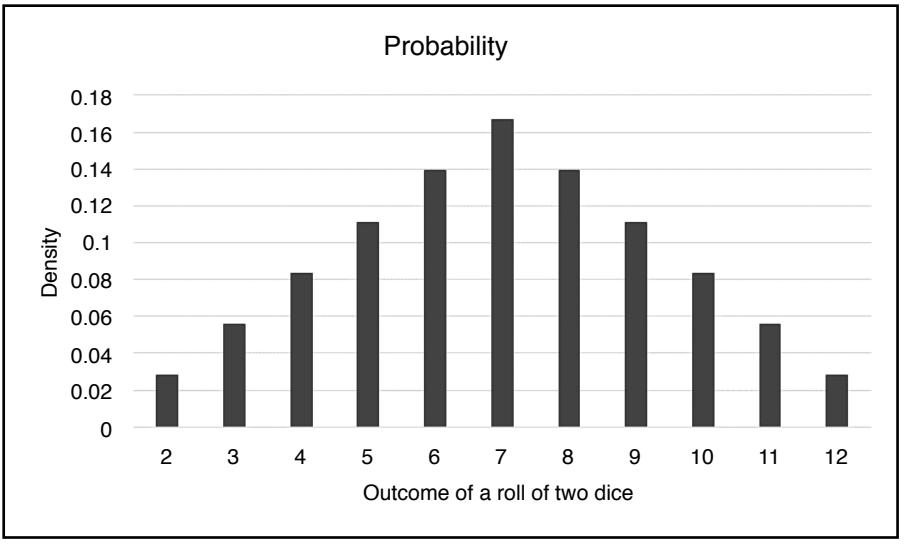


Figure 6.4 Histogram of the outcomes for rolling two dice

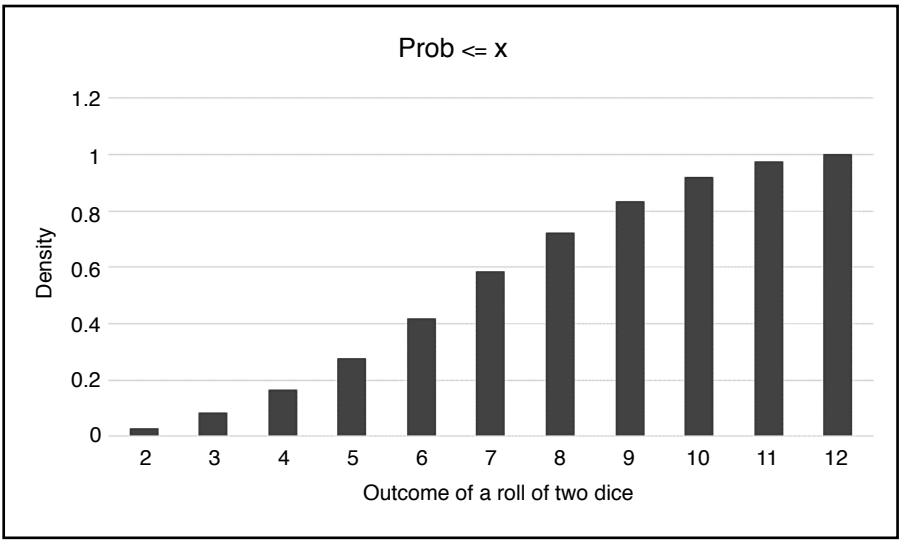


Figure 6.5 Probability distribution function for rolling two dice

Normal Distribution

The Normal distribution is one of the most commonly referred to distributions in statistical analysis and even in everyday conversations. A large body of academic and scholarly work rests on the fundamental assumption that the underlying data follows a Normal distribution that generates a bell-shaped curve. Mathematically, Normal distribution is expressed as shown in Equation 6.1:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2 * pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \textbf{Equation 6.1}$$

Where x is a random variable, μ is the mean and σ is the standard deviation. The standard normal curve refers to 0 mean and constant variance; that is, σ = 1 and is represented mathematically as shown in Equation 6.2:

$$f(x, 0, 1) = \frac{1}{\sqrt{2 * pi}} e^{-\frac{(x)^2}{2}} \quad \textbf{Equation 6.2}$$

I generate a regular sequence of numbers (x) between –4 and 4. I can transform x as per Equation 6.2 into y. The plot in Figure 6.6 presents the standard normal curve or the probability density function, which plots the random variable x on the x-axis and density (y) on the y-axis.

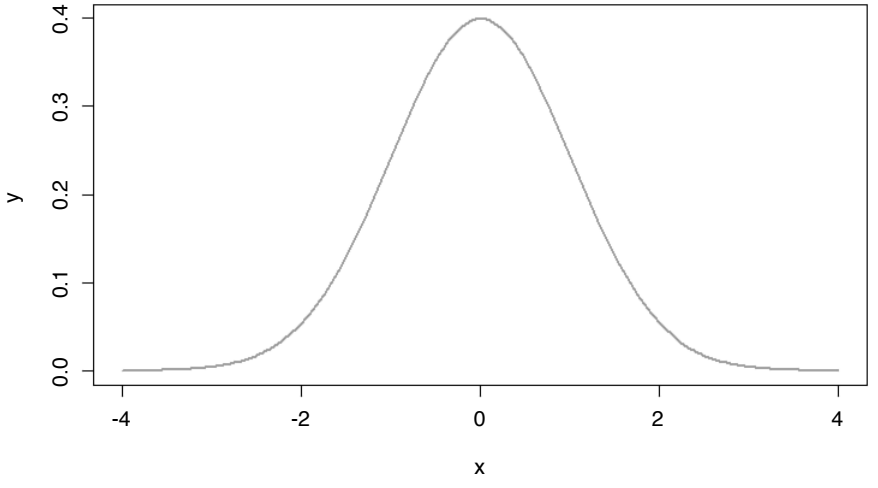


Figure 6.6 The bell-shaped Normal distribution curve

The Student Who Taught Everyone Else

The other commonly used distribution is the Student's t-distribution, which was specified by William Sealy Gosset. He published a paper in Biometrika in 1908 under the pseudonym, Student. Gosset worked for the Guinness Brewery in Durbin, Ireland, where he worked with small samples of barley.

Mr. Gosset is the unsung hero of statistics. He published his work under a pseudonym because of the restrictions from his employer. Apart from his published work, his other contributions to statistical analysis are equally significant. The *Cult of Statistical Significance*, a must read for anyone interested in data science, chronicles Mr. Gosset's work and how other influential statisticians of the time, namely Ronald Fisher and Egon Pearson, by way of their academic bona fides ended up being more influential than the equally deserving Mr. Gosset.

The t-distribution refers to a family of distributions that deal with the mean of a normally distributed population with small sample sizes and unknown population standard deviation. The Normal distribution describes the mean for a population, whereas the t-distribution describes the mean of samples drawn from the population. The t-distribution for each sample could be different and the t-distribution resembles the normal distribution for large sample sizes.

In Figure 6.7, I plot t-distributions for various sample sizes, also known as the degrees of freedom, along with the normal distribution. Note that the t-distribution with a sample size of 30 resembles the normal distribution the most.

Over the years, 30 has emerged as the preferred threshold for a large enough sample that may prompt one to revert to the Normal distribution. Many researchers, though, question the suitability of 30 as the threshold. In the world of big data, 30 obviously seems awfully small.

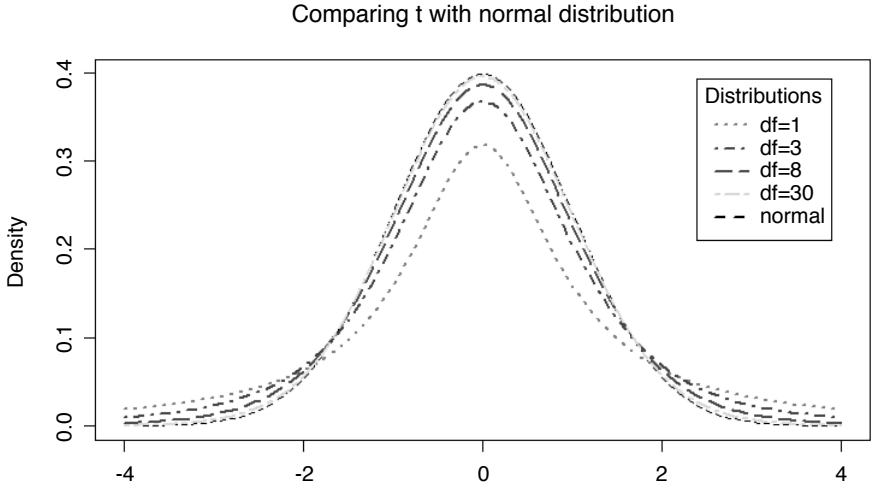


Figure 6.7 Probability distribution curves for normal and t-distributions for different sample sizes

Statistical Distributions in Action

I now illustrate some applied concepts related to the Normal distribution. Assuming that the data are "normally" distributed, I can in fact determine the likelihood of an event. For instance, consider the Teaching Ratings data, which I have discussed in detail in Chapters 4 and 5. The data set contains details on course evaluations for 463 courses and the attributes of the instructors. Professor Hamermesh and his co-author wanted to determine whether the subjective measure of an instructor's appearance influenced his or her teaching evaluation score.

The descriptive statistics for some variables in the data set are presented in Table 6.2. I also report the R code in this chapter used to generate the output. The following code launches two R packages: `xtable` and `psych`. It commits the data file to R's dedicated reference memory and then runs summary statistics, which are formatted and produced using the RMarkdown extensions.

```
### download data from the course's website
library(xtable)
library(psych)
attach(TeachingRatings)
tab <- xtable(describe(cbind(eval, age, beauty, students, allstudents),
                       skew=F, ranges=F), digits=3)
rownames(tab)<- c("teaching evaluation score", "instructor's age",
             "beauty score", "students responding to survey","students
             registered in course")
print(tab, type="html")
```

Table 6.2 Summary Statistics for Teaching Evaluation Data

	vars	n	mean	sd	se
teaching evaluation score	1	463	3.998	0.554	0.026
instructor's age	2	463	48.365	9.803	0.456
beauty score	3	463	0.000	0.789	0.037
students responding to survey	4	463	36.624	45.018	2.092
students registered in course	5	463	55.177	75.073	3.489

Table 6.2 demonstrates that the average course evaluation (non-weighted mean) was 3.998 and the standard deviation (SD) was 0.554. I also report descriptive statistics for other variables in the table and plot the histogram for the teaching evaluation score to visualize the distribution (see Figure 6.8). I see that the teaching evaluation scores peak around 4.0 with a relatively larger spread on the right, suggesting more frequent occurrence of higher than average teaching evaluation scores than lower teaching evaluation scores.

```
histogram(TeachingRatings$eval, nint=15, aspect=2,
          xlab="teaching evaluation score", col=c("dark grey"))
```

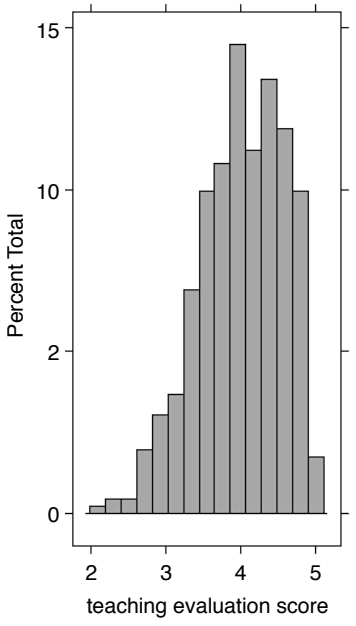


Figure 6.8 Histogram of teaching evaluation scores

I can plot a Normal density function based on descriptive statistics for the evaluation data. If I were to assume that the teaching evaluation scores were normally distributed, I only need the mean and standard deviation to plot the Normal density curve. The resulting plot is presented in Figure 6.9.

Unlike the histogram presented in Figure 6.8, the theoretical distribution in Figure 6.9 is more symmetrical, suggesting that the theoretical distributions are neater and symmetrical, whereas the real-world data is "messier."

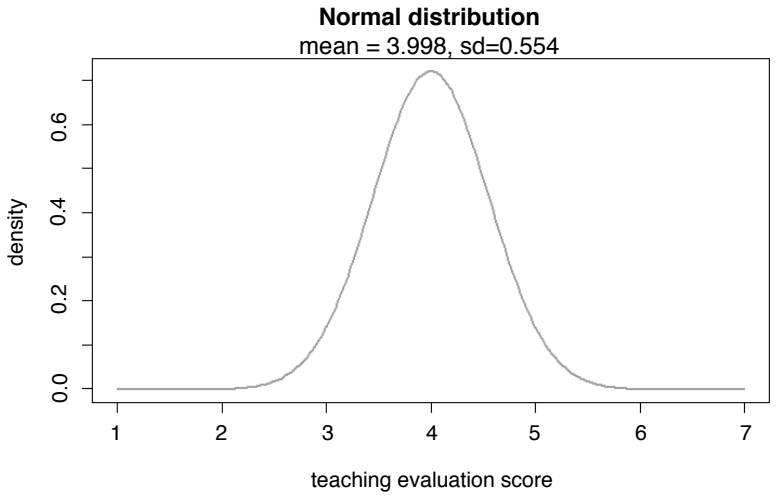


Figure 6.9 Normal distribution curve with mean=3.998, and sd=0.554

Z-Transformation

A related and important concept is the z-transformation of a variable. One can transform a variable such that its transformed version returns a mean of 0 and a standard deviation of 1. I use the formula in Equation 6.3 for z-transformation:

$$z = \frac{x - \mu}{\sigma}$$ **Equation 6.3**

The preceding equation showcases x as the raw data, μ as the mean and σ as the standard deviation. For example, if an instructor received a course evaluation of 4.5, the z-transformation can be calculated as follows:

$$z = \frac{4.5 - 3.998}{.554} = 0.906$$

I can create a new variable by standardizing the variable `eval` and plot a histogram of the standardized variable. The mean of the standardized variable is almost 0 and the standard deviation equals 1 (see Figure 6.10).

```
z.eval<-as.matrix((TeachingRatings$eval-3.998)/.554)
histogram(z.eval, nint=15, aspect=2,
   xlab=" normalized teaching evaluation score", col=c("dark grey"))
```

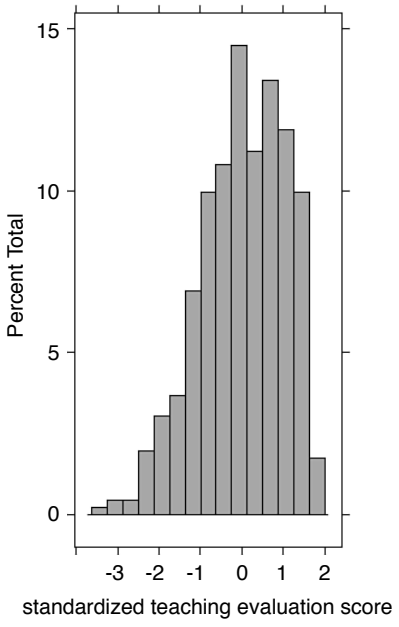


Figure 6.10 Histogram of standardized teaching evaluation score

The z-transformed data is useful in determining the probability of an event being larger or smaller than a certain threshold. For instance, assuming that the teaching evaluations are normally distributed, I can determine the probability of an instructor receiving a teaching evaluation greater or lower than a particular value. I explain this concept in the following section.

Probability of Getting a High or Low Course Evaluation

Let us assume that the variable `eval` is Normally distributed. I can determine the probability of obtaining the evaluation higher or lower than a certain threshold. For instance, let us determine the probability of an instructor receiving a course evaluation of higher than 4.5 when the mean evaluation is 3.998 and the standard deviation (SD) is 0.554. All statistical software, including spreadsheets such as Microsoft Excel, provide built-in formulae to compute these probabilities. See the following R code.

```
pnorm(c(4.5), mean=3.998, sd=0.554, lower.tail=FALSE)
```

R readily computes 0.1824, or simply 18.24%. This suggests that the probability of obtaining a course evaluation of higher than 4.5 is 18.24%.

Another way of conceptualizing the probability of obtaining a teaching evaluation of higher than 4.5 is to see it illustrated in a plot (see Figure 6.11). Notice that the area under the curve to the right of the value 4.5 is shaded gray, which represents the probability of receiving a teaching evaluation score of higher than 4.5.

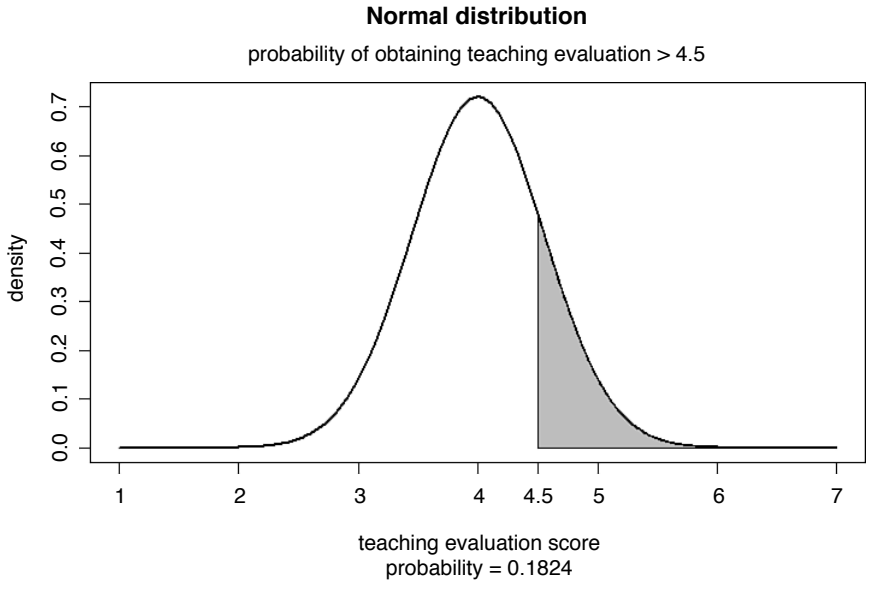


Figure 6.11 Probability of obtaining a teaching evaluation score of greater than 4.5

The gray-shaded part of the area represents 18.24% of the area under the curve. The area under the normal distribution curve is assumed as one, or in percentage terms, 100%. This is analogous to a histogram and represents the collective probability of all possible values attained by a variable. Thus, the probability of obtaining a course evaluation of greater than or equal to 4.5 (the area shaded in gray) is 0.1824. The probability of obtaining a teaching evaluation of less than or equal to 4.5 will be 1 – 0.1824 = 0.8176 or 81.76%, which is shaded gray in Figure 6.12.

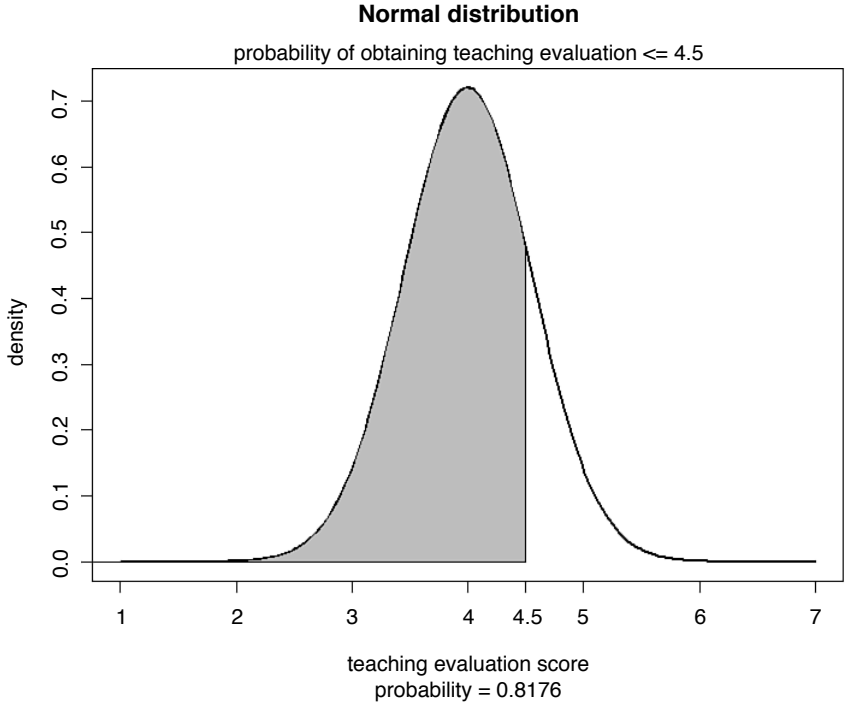


Figure 6.12 Probability of obtaining a teaching evaluation score of less than 4.5

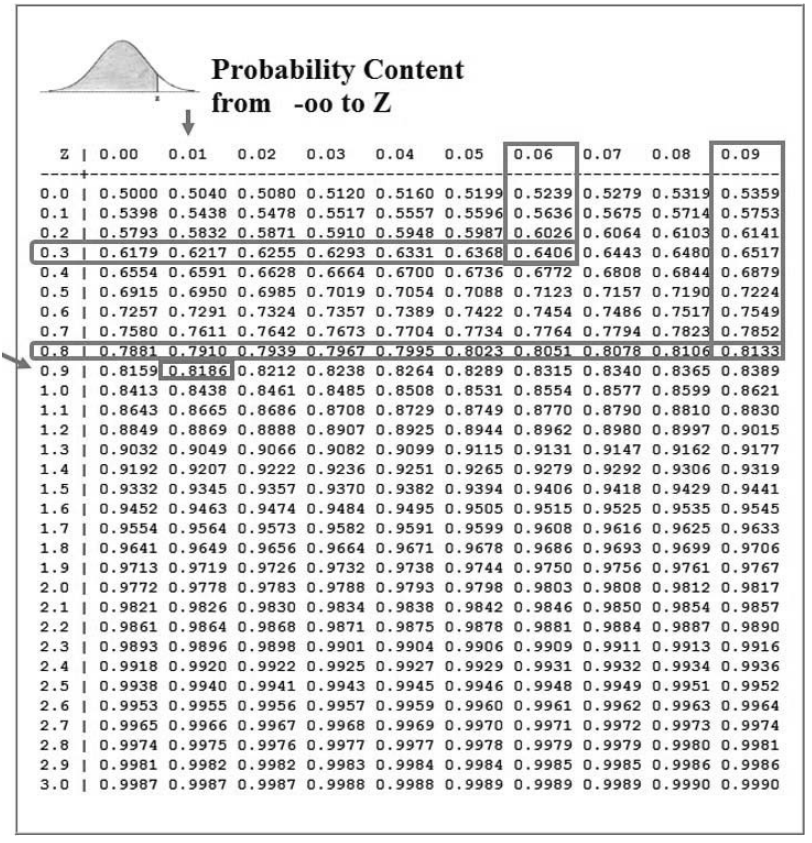
Probabilities with Standard Normal Table

Now let us repeat the same calculations using a calculator and the probability tables. I first standardize the raw data to determine the probability of a teaching evaluation score of higher than 4.5. I have demonstrated earlier the calculations to standardize 4.5, which equals 0.906.

The next step is to determine the probability value (`p-value`) from the probability table in Figure 6.13. Notice that the `p-values` listed in the table are for probability of less than or equal to a certain value, which is referred to as the left tail of the distribution. I need to subtract the `p-value` from 1 to obtain the probability value for greater than the selected value. The table expresses z-scores up to two decimal points.

From the calculations, you see that z approximates to 0.91, for a teaching evaluation score of 4.5. I search for the `p-value` corresponding to Z = 0.91 in Figure 6.13. I locate 0.9 in the first column and then locate 0.01 in the first row. The `p-value` reported in the cell at the intersection

of the aforementioned row and column is 0.8186 or 81.86% (also highlighted in the table with a box). Notice that this is the probability of getting a course evaluation of 4.5 or less. Notice also that this value is almost the same as reported in the value listed in the figure generated by R. Slight differences are due to rounding.



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure 6.13 Normal distribution table

Source: http://www.math.unb.ca/~knight/utility/NormTble.htm

To obtain the probability of receiving a course evaluation of higher than 4.5, I simply subtract 0.8186 from 1; I have 1 – 0.8186 = 0.1814 or 18.14%.

Let us now try to determine the probability of receiving a course evaluation between 3.5 and 4.2. I first need to standardize both scores. Here are the calculations.

Remember that:

$$z = \frac{x - \mu}{\sigma}$$

Thus,

$$z = \frac{4.2 - 3.998}{.554} = 0.36$$

$$z = \frac{3.5 - 3.998}{.554} = -0.89$$

From the Standard Normal Table you need to search for two values: one for Z = 0.36 and the other for Z = –0.89. The difference between the corresponding `p-values` will give the probability for course evaluations falling between 3.5 and 4.2. The calculations are straightforward for 4.2. The standardized value (`z-score`) is 0.36, which is highlighted in the table where the corresponding row and column intersect to return a `p-value` of 0.64 or 64%. This implies that the probability of receiving a course evaluation of 4.2 or less is 64%.

The z-transformation for 3.5 returns a negative z-score of –0.899. I again use Figure 6.13 to first locate 0.8 in the first column and then 0.09 in the first row and search for the corresponding `p-value` that is located at the intersection of the two. The resulting value is 0.8133. However, this is the `p-value` that corresponds to a z-score of +0.899. The `p-value` corresponding to a z-score of –0.899 is `1-p-value`, which happens to be 1–0.8133 = 0.18 or 18%, which suggests that the probability of obtaining a course evaluation of 3.5 or less is 18%. The results are presented in Table 6.3.

Table 6.3 Standardizing Teaching Evaluation Scores

Raw Data	Z-transformed	`P-value` ≤ Z
4.2	0.362	0.64 or 64%
3.5	–0.899	1–0.8133 = 0.186 or 18.6%

I still have not found the answer to the question regarding the probability of obtaining a course evaluation of greater than 3.5 and less than or equal to 4.2. To illustrate the concept consider Figure 6.14. The shaded area represents the probability of obtaining a teaching evaluation between 3.5 and 4.2. From Table 6.3, I see that the probability of a teaching evaluation of 4.2 or less is 0.64 and for a teaching evaluation of 3.5 or lower is 0.186, so the difference between the two will have our answer.

Mathematically:

$$0.64 - 0.18 = 0.46$$

Thus, 46% is the probability of obtaining a course evaluation of greater than 3.5 and 4.2 or lower. It is marked by the unshaded area in Figure 6.14.

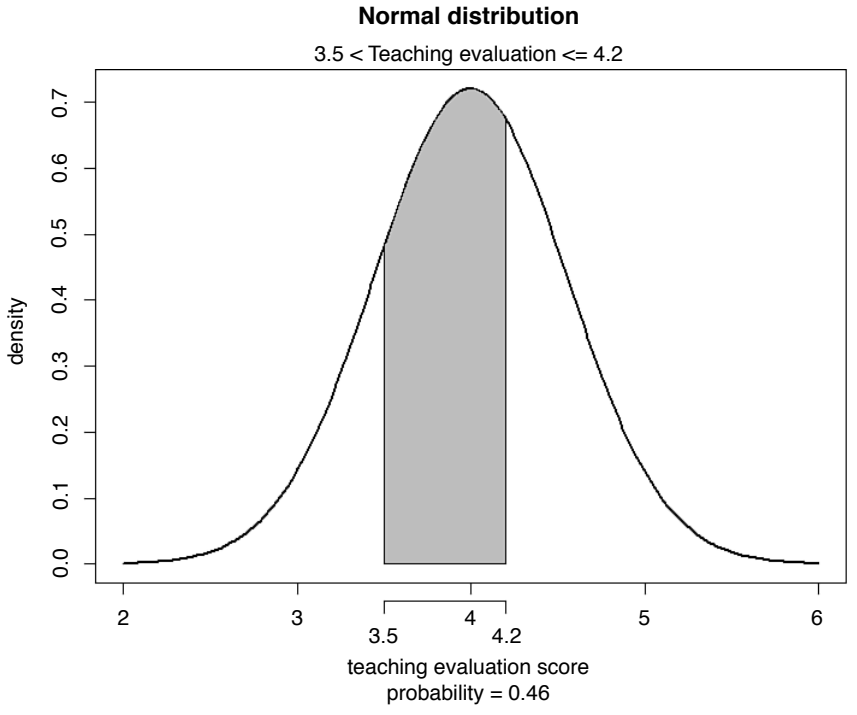


Figure 6.14 Probability of obtaining teaching evaluation score of greater than 3.5 and 4.2 or less

The probability can be readily obtained in statistical software, such as R and Stata.

R code:

```
1-(pnorm(c(4.2, 3.5), mean = 3.998, sd = .554, lower.tail=F))
or
pnorm(.362)-pnorm(-.899)
```

Stata code:

```
di 1-(1-normal(.362) + normal(-.899))
```

Hypothetically Yours

Most empirical analysis involves comparison of two or more statistics. Data scientists and analysts are often asked to compare costs, revenues, and other similar metrics related to socioeconomic outcomes. Often, this is accomplished by setting up and testing hypotheses. The purpose of the analysis is to determine whether the difference in values between two or more entities or outcomes is a result of chance or whether fundamental and statistically significant differences are at play. I explain this in the following section.

Consistently Better or Happenstance

Nate Silver, the former blogger for the *New York Times* and founder of fivethirtyeight.com has been at the forefront of popularizing data-driven analytics. His blogs are followed by hundreds of thousands readers. He is also credited with popularizing the use of data and analytics in sports. The same trend was highlighted in the movie *Moneyball*, in which a baseball coach with a data-savvy assistant puts together a team of otherwise regular players who were more likely to win as a team. The coach and his assistant, instead of relying on the traditional criterion, based their decisions on data. They were, therefore, able to put together a winning team.

Let me illustrate hypothesis testing using basketball as an example. Michael Jordan is one of the greatest basketball players. He was a consistently high scorer throughout his career. In fact, he averaged 30.12 points per game in his career, which is the highest for any basketball player in the NBA.[4] He spent most of his professional career playing for the Chicago Bulls. Jordan was inducted into the Hall of Fame in 2009. In his first professional season in 1984–85, Jordan scored on average 28.2 points per game. He recorded his highest seasonal average of 37.1 points per game in the 1986–87 season. His lowest seasonal average of 20 points per game was observed toward the end of his career in the 2002–03 season.

Another basketball giant is Wilt Chamberlain, who is one of basketball's first superstars and is known for his skill, conspicuous consumption, and womanizing. With 30.07 points on average per game, Chamberlain is a close second to Michael Jordan for the highest average points scored per game. Whereas Michael Jordan's debut was in 1984, Chamberlain began his professional career with the NBA in 1959 and was inducted into the Hall of Fame in 1979.

Just like Michael Jordan, who scored the highest average points per game in his third professional season, Chamberlain also scored an average of 50.4 points per game in 1961–6, his third season. Again, just like Michael Jordan, Chamberlain's average dropped at the tail end of his career when he scored 13.2 points per game on average in 1972–73.

Jordan's 30.12 points per game on average and Chamberlain's 30.06 points per game on average are very close. Notice that I am referring here to average points per game weighted by the number of games played in each season. A simple average computed from the average of points per game per season will return slightly different results.

While both Jordan and Chamberlain report very similar career averages, there are, however, significant differences in the consistency of their performance throughout their careers.

Instead of the averages, if we compare standard deviations, we see that Jordan with a standard deviation of 4.76 points is much more consistent in his performance than Chamberlain was with a standard deviation of 10.59 points per game. If we were to assume that these numbers are normally distributed, I can plot the Normal curves for their performance. Note in Figure 6.15 that Michael Jordan's performance returns a sharper curve (colored in black), whereas Chamberlain's curve (colored in gray) is flatter and spread wide. We can see that Jordan's score is mostly in the 20 to 40 points per game range, whereas Chamberlain's performance is spread over a much wider interval.

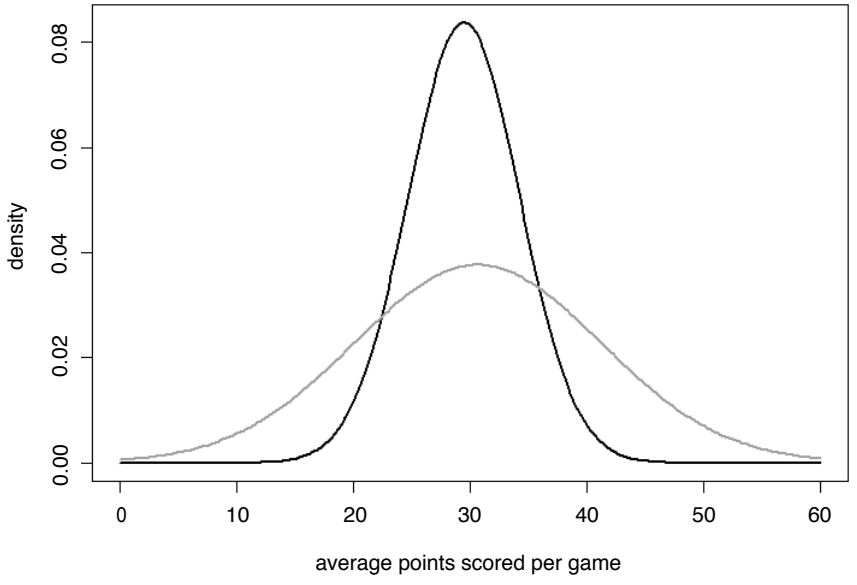


Figure 6.15 Normal distribution curves for Michael Jordan and Wilt Chamberlain

Mean and Not So Mean Differences

I use statistics to compare the consistency of scoring between two basketball giants. The comparison of means (averages) comes in three flavors. First, you can assume that the mean points per game scored by both Jordan and Chamberlain are the same. That is, the difference between the mean scores of the two basketball legends is zero. This becomes our null hypothesis. Let μ_j represent the average points per game scored by Jordan and μ_c represent the average points per game scored by Chamberlain. My null hypothesis, denoted by H_0, is expressed as follows:

H_0:$\mu_j = \mu_c$

The alternative hypothesis, denoted as H_a, is as follows:

H_a:$\mu_j \neq \mu_c$; their average scores are different.

Now let us work with a different null hypothesis and assume that Michael Jordan, on average, scored higher than Wilt Chamberlain did. Mathematically:

H_0:$\mu_j > \mu_c$

The alternative hypothesis will state that Jordan's average is lower than that of Chamberlain's, H_a:$\mu_j < \mu_c$. Finally, I can restate our null hypothesis to assume that Michael Jordan, on average, scored lower than Wilt Chamberlain did. Mathematically:

H_0:$\mu_j < \mu_c$

The alternative hypothesis in the third case will be as follows:

H_a:$\mu_j > \mu_c$; Jordan's average is higher than that of Chamberlain's.

I can test the hypothesis using a `t-test`, which I will explain later in the chapter.

Another less common test is known as the `z-test`, which is based on the normal distribution. Suppose a basketball team is interested in acquiring a new player who has scored on average 14 points per game. The existing team's average score has been 12.5 points per game with a standard deviation of 2.8 points per game. The team's manager wants to know whether the new player is indeed a better performer than the existing team. The manager can use the `z-test` to find the answer to this riddle. I explain the `z-test` in the following section.

Handling Rejections

After I state the null and alternative hypotheses, I conduct the `z-` or the `t-test` to compare the difference in means. I calculate a value for the test and compare it against the respective critical value. If the calculated value is greater than the critical value, I can reject the null hypothesis. Otherwise, I fail to reject the null hypothesis.

Another way to make the call on the hypothesis tests is to see whether the calculated value falls in the rejection region of the probability distribution function. The fundamental principal here is to determine, given the distribution, how likely it is to get a value as extreme as the one we observed. More often than not, we use the 95% threshold. We would like to determine whether the likelihood of obtaining the observed value for the test is less than 5% for a 95% confidence level. If the calculated value for the `z-` or the `t-test` falls in the region that covers the 5% of the distribution, which we know as the rejection region, we reject the null hypothesis. I illustrate the regions for normal and t-distributions in the following sections.

Normal Distribution

Recall from the last section that the alternative hypothesis comes in three flavors: The difference in means is not equal to 0, the difference is greater than 0, and the difference is less than 0. We have three rejection regions to deal with the three scenarios.

Let us begin with the scenario where the alternative hypothesis is that the mean difference is not equal to 0. We are not certain whether it is greater or less than zero. We call this the *two-tailed test*. We will define a rejection region in both tails (left and right) of the normal distribution. Remember, we only consider 5% of the area under the normal curve to define the rejection region. For a two-tailed test, we divide 5% into two halves and define rejection regions covering 2.5% under the curve in each tail, which together sum up to 5%. See the two-tailed test illustrated in Figure 6.16.

Recall that the area under the normal density plot is 1. The gray-shaded area in each tail identifies the rejection region. Taken together, the gray area in the left (2.5% of the area) and in the right (2.5% of the area) constitute the 5% rejection region. If the absolute value of the `z-test` is greater than the absolute value of 1.96, we can safely reject the null hypothesis that the difference in means is zero and conclude that the two average values are significantly different.

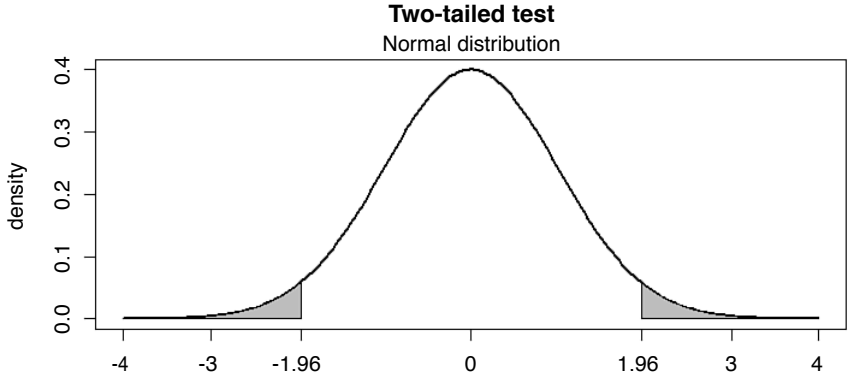


Figure 6.16 Two-tailed test using Normal distribution

Let us now consider a scenario where we believe that the difference in means is less than zero. In the Michael Jordan–Wilt Chamberlain example, we are testing the following alternative hypothesis:

$H_a: \mu_j < \mu_c$; Jordan's average is lower than that of Chamberlain's.

In this particular case, I will only define the rejection region in the left tail (the gray-shaded area) of the distribution. If the calculated `z-test` value is less than –1.64, for example, –1.8, we will know that it falls in the rejection region (see Figure 6.17) and we will reject the null hypothesis that the difference in means is greater than zero. The test is also called one-tailed test.

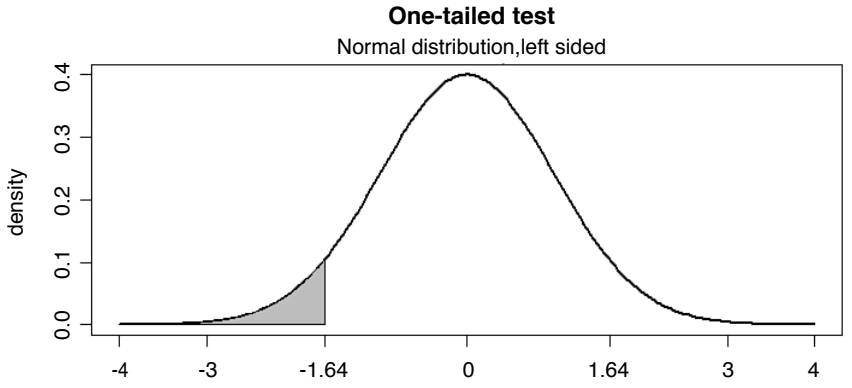


Figure 6.17 One-tailed test (left-tail)

Along the same lines, let us now consider a scenario where we believe that the difference in means is greater than zero. In the Michael Jordan–Wilt Chamberlain example, we are testing the following alternative hypothesis:

H_a:$\mu_j > \mu_c$; Jordan's average is higher than that of Chamberlain's.

In this particular case, I will only define the rejection region (the gray-shaded area) in the right tail of the distribution. If the calculated `z-test` value is greater than 1.64, for example, 1.8, we will know that it falls in the rejection region (see Figure 6.18) and we will reject the null hypothesis that the difference in means is less than zero.

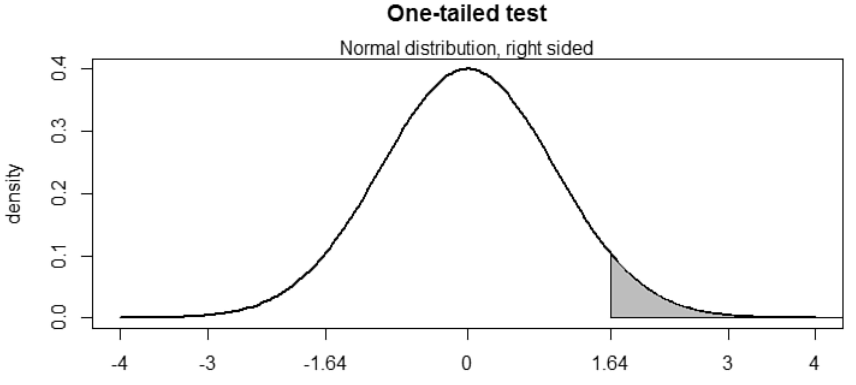


Figure 6.18 One-tailed test (right-tail)

t-distribution

Unlike the `z-test`, most comparison of means tests are performed using the t-distribution, which is preferred because it is more sensitive to small sample sizes. For large sample sizes, the t-distribution approximates the normal distribution. Those interested in this relationship should explore the *central limit theorem*. For the readers of this text, suffice it to say that for large sample sizes, the distribution looks and feels like the normal distribution, a phenomenon I have already illustrated in Figure 6.7.

I have chosen not to illustrate the rejection regions for the t-distribution, because they look the same as the ones I have illustrated for the normal distribution for sample sizes of say 200 or greater. Instead, I define the critical t-values.

As the number of observations increases, the critical t-values approach the ones we obtain from the `z-test`. For a left tail (mean is less than zero) test with 30 degrees of freedom, the critical value for a `t-test` at 5% level is –1.69. However, for a large sample with 400 degrees of freedom, the critical value is –1.648, which comes close to –1.64 for the normal distribution. Thus, one can see from Figure 6.19 that the critical values for t-distribution approaches the ones for normal distribution for larger samples. I, of course, would like to avoid the controversy for now on what constitutes as a large sample.

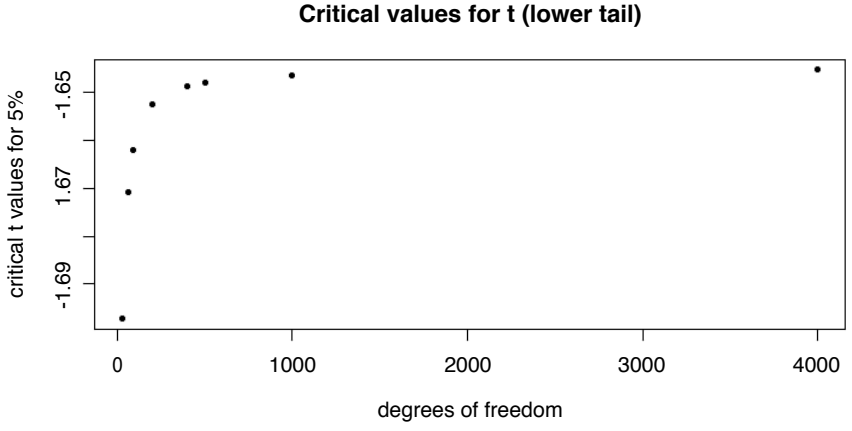


Figure 6.19 Critical t-values for left-tail test for various sample sizes

General Rules Regarding Significance

Another way of testing the hypothesis is to use the probability values associated with the `z-` or the `t-test`. Consider the general rules listed in Table 6.4 for testing statistical significance.

Table 6.4 Rules of Thumb for Hypothesis Testing

Type of Test	z or t Statistics*	Expected `p-value`	Decision
Two-tailed test	The absolute value of the calculated z or t statistics is greater than 1.96	Less than 0.05	Reject the null hypothesis
One-tailed test	The absolute value of the calculated z or t statistics is greater than 1.64	Less than 0.05	Reject the null hypothesis

* With large samples only for t statistics

Note that for `t-tests`, the 1.96 threshold works with large samples. For smaller samples, the critical t-value will be larger and depend upon the degrees of freedom (the sample size).

The Mean and Kind Differences

We are often concerned with comparing two or more outcomes. For instance, we might be interested in comparing sales from one franchise location with the rest. Statistically, we might have four conditions when we are concerned with comparing the difference in means between groups. These are

1. Comparing the sample mean to a population mean when the population standard deviation is known
2. Comparing the sample mean to a population mean when the population standard deviation is not known
3. Comparing the means of two independent samples with unequal variances
4. Comparing the means of two independent samples with equal variances

I discuss each of the four scenarios with examples in the following sections.

Comparing a Sample Mean When the Population SD Is Known

I continue to work with the basketball example. Numerous basketball legends are known for their high-scoring performance. The two in the lead are Wilt Chamberlain and Michael Jordan. However, there appears to be a huge difference between the two leading contenders and others. LeBron James, who is third in the NBA rankings for career average points per game at 27.5, is very much behind Jordan and Chamberlain, and is marginally better than Elgin Baylor at 27.36 (see Figure 6.20).

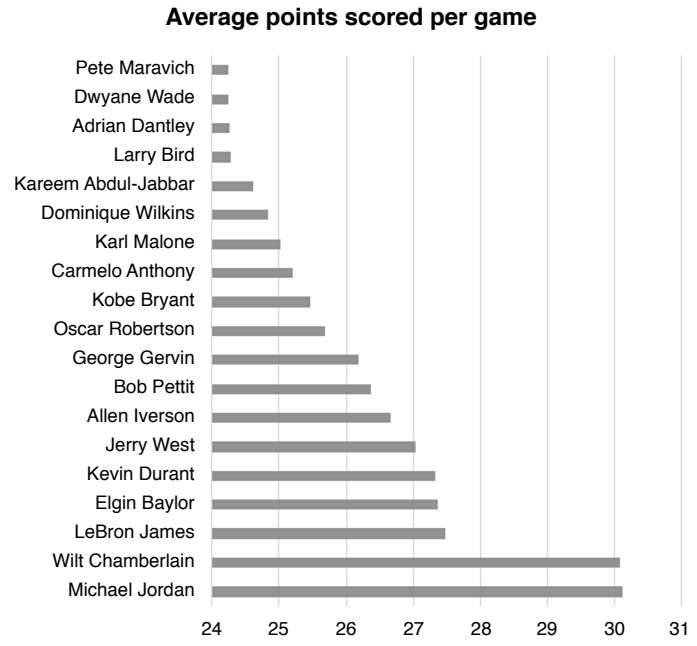


Figure 6.20 Average points scored per game for the leading NBA players

Source: http://www.basketball-reference.com/leaders/pts_per_g_career.html

Many believe that the hefty salaries of celebrity athletes reflect their exceptional performance. In basketball, scoring high points is one of the criteria, among others, that determines a player's worth. There is some truth to it. During 2013–14, Kobe Bryant of the LA Lakers earned more than $30 million. However, he was not the highest scorer in the team (see Figure 6.21). The basketball example provides us the backdrop to conduct the comparison of means test.

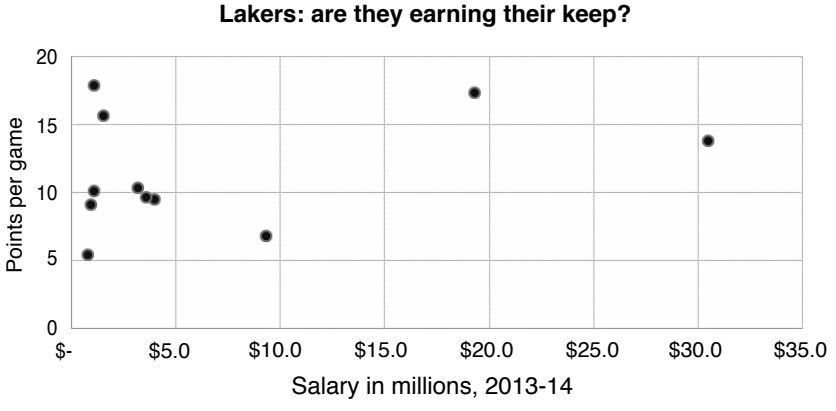


Figure 6.21 Lakers: Are they earning their keep?

The Basketball Tryouts

Let us assume that a professional basketball team wants to compare its performance with that of players in a regional league. The pros are known to have a historic mean of 12 points per game with a standard deviation of 5.5. A group of 36 regional players recorded on average 10.7 points per game. The pro coach would like to know whether his professional team scores on average are different from that of the regional players. I will use the two-tailed test from a normal distribution to determine whether the difference is statistically different. I start by calculating the z-value as shown in Equation 6.4:

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Equation 6.4

The specs are as follows:

Average points per player for the regional players: 10.7 ($\overline{x}$)

Std. Dev of the population: 5.5 (σ)

Average points per game scored by pros: 12 (μ)

Our null hypothesis: H_0: $\mu = 12$

Alternative hypothesis: H_a: $\mu \neq 12$

$$z = \frac{10.7 - 12}{\frac{5.5}{\sqrt{36}}} = -1.42$$

For a two-tailed test at the 5% level, each tail represents 2.5% of the area under the curve. The critical value for the `z-test` at the 95% level is ±1.96. Because the absolute value for –1.42 is lower than the absolute value for –1.96, I fail to reject the null hypothesis that the difference in means is zero and conclude that statistically speaking, 10.7 is not much different from the pros' average of 12 points scored per game.

Figure 6.22 presents the `z-test` graphically. Note that –1.42 does not fall in the gray-shaded area, which constitutes the rejection region. Thus, I cannot reject the null hypothesis that the mean difference is equal to 0.

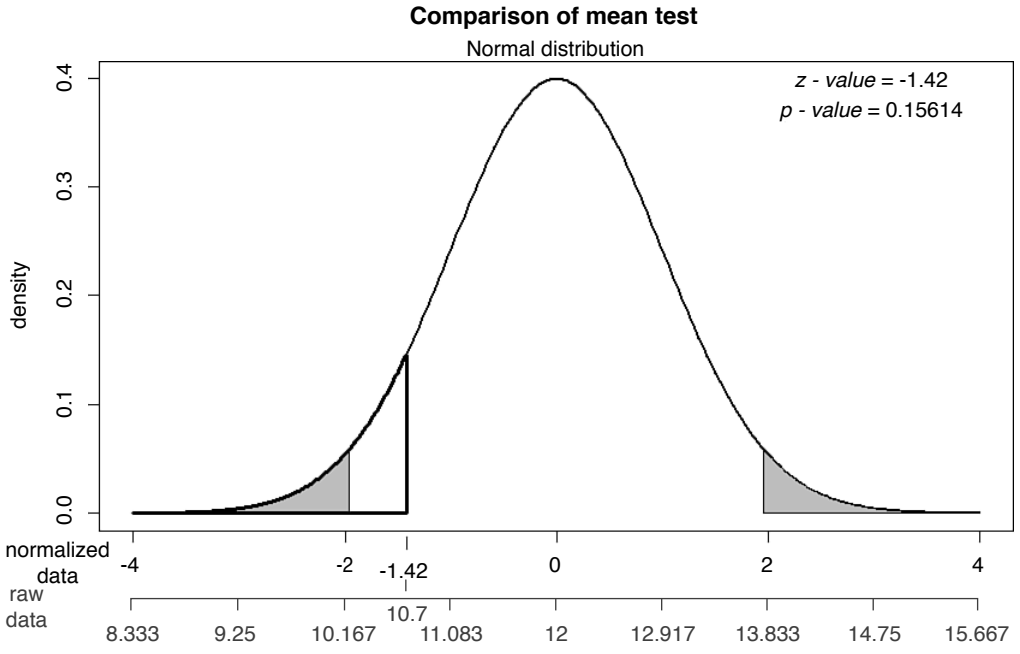


Figure 6.22 Two-tailed test for basketball tryouts

All statistical software and spreadsheets can calculate the `p-value` associated with the observed mean. In this example using a two-tailed test, the `p-value` associated with the average score of 10.7 points per game is 0.156, if the mean of the distribution is 12 and the standard error of the mean is $\frac{5.5}{\sqrt{36}} = 0.912$

Left Tail Between the Legs

A university basketball team might become a victim of austerity measures. Forced to reduce the budget deficit, the university is considering cutting off underperforming academic and nonacademic programs. The basketball team has not done well as of late. Hence, it has been included in the list of programs to be terminated.

The coach, however, feels that the newly restructured team has the potential to rise to the top of the league and that the bad days are behind them. Disbanding the team now would be a mistake. The coach convinces the university's vice president of finance to have the team evaluated by a panel of independent coaches to rank the team on a scale of 1 to 10. It was agreed that if the team received the average score of 7 or higher, the team may be allowed to stay for another year, at which time the decision will be revisited.

A panel of 20 independent coaches was assembled to evaluate the team's performance. After reviewing the team's performance, the panel's average score equaled 6.5 with a standard deviation of 1.5. The VP of finance now has to make a decision. Should the team stay or be disbanded?

The VP of finance asked a data analyst in her staff to review the stats and assist her with the decision. She was of the view that the average score of 6.5 was too close to 7, and that she wanted to be sure that there was a statistically significant difference that would prompt her to disband the basketball team.

The analyst decided to conduct a one-sample mean test to determine whether the average score received was 7 or higher.

The null hypothesis: $H_0: = \mu \geq 7.$

The alternative hypothesis: $H_a: = \mu < 7.$

$$z = \frac{6.5 - 7}{\frac{1.5}{\sqrt{20}}} = -1.491$$

Based on the test, the analyst observed that the `z-value` of –1.49 does not fall in the rejection region. Thus, he failed to reject the null hypothesis, which stated that the average score received was 7 or higher. The VP of finance, after reviewing the findings, decided not to cut funding to the team because, statistically speaking, the average ranking of the basketball team was not different from the threshold of 7.

The test is graphically illustrated in Figure 6.23.

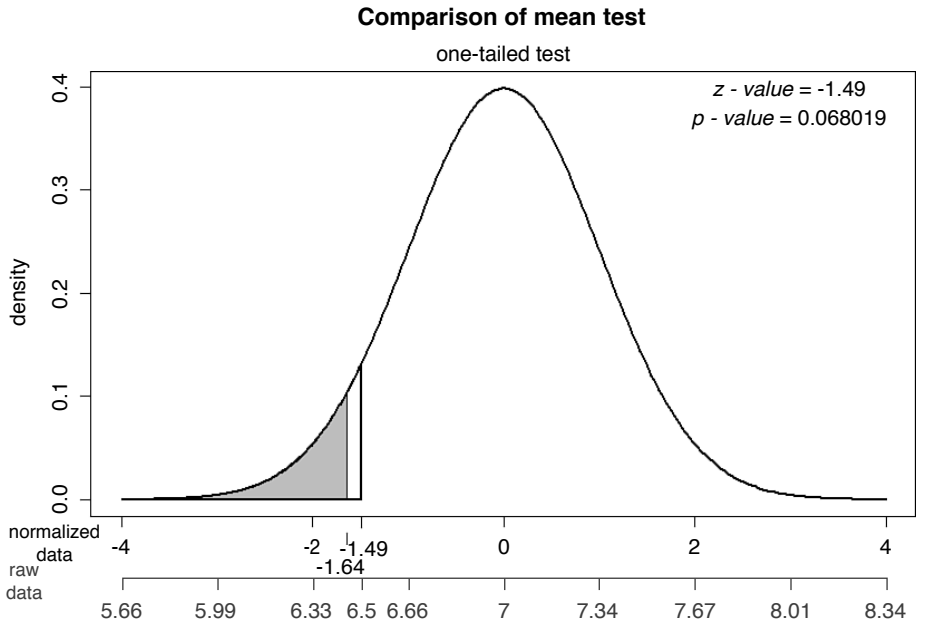


Figure 6.23 Basketball tryouts, left-tail test

Lastly, consider hotdog vendors outside a basketball arena where the local NBA franchise plays. It has been known that when the local team was not winning, the vendors would sell on average 500 hotdogs per game with a standard deviation of 50. Assume now that the home team has been enjoying a winning streak for the last five games that is accompanied with an average sale of 550 hotdogs. The vendors would like to determine whether they are indeed experiencing higher sales. The hypothesis is stated as follows:

The null hypothesis: $H_0: \mu \leq 500$

Alternative hypothesis: $H_a: \mu > 500$

The `z-value` is calculated as follows:

$$z = \frac{550 - 500}{\left(\frac{50}{\sqrt{5}}\right)} = 2.24$$

I see that the `z-value` for the test is 2.24 and the corresponding `p-value` is 0.0127, which is less than 0.05. I can therefore reject the null hypothesis and conclude that there has been a statistically significant increase in hotdog sales.

Given that I only had five observations, it would have been prudent to use the t-distribution instead, which is more suited to small samples. The test is illustrated in Figure 6.24.

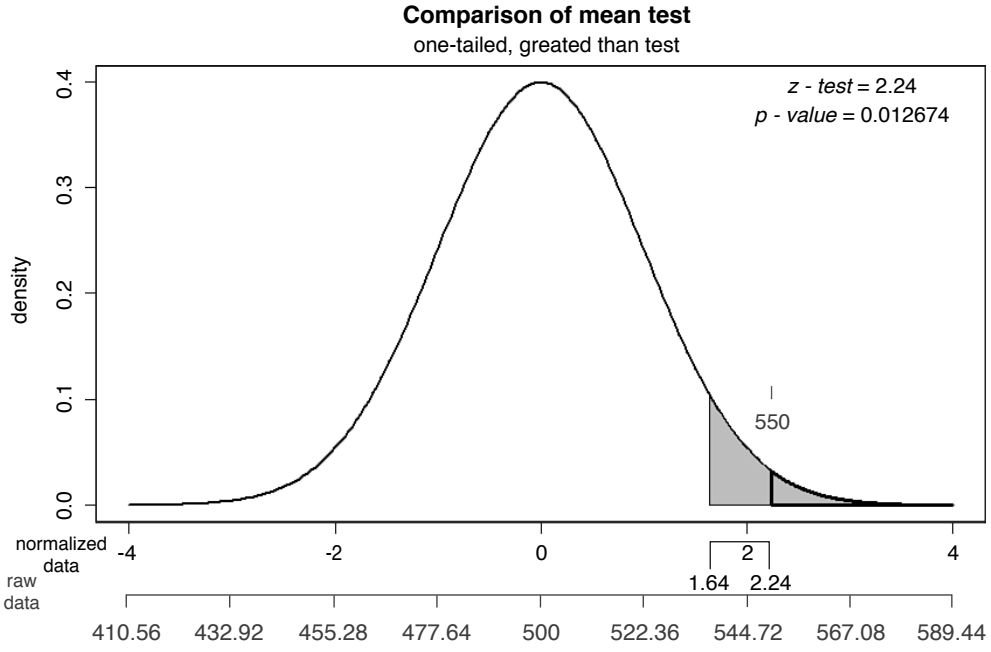


Figure 6.24 Hotdog sales

Comparing Means with Unknown Population SD

We use the t-distribution in instances where we do not have access to population standard deviation. The test statistic is shown in Equation 6.5:

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Equation 6.5

Note that σ (population standard deviation) has been replaced by *s*(sample standard deviation).

Consider the case where a large franchise wants to determine the performance of a newly opened store. The franchise surveyed a sample of 35 existing stores and found that the average weekly sales were $166,000 with a standard deviation of $25,000. The new store on average reported a weekly sale of $170,000. The managers behind the launch of the new store are of the view that the new store represents the new approach to retailing, which is the reason why the new store sales are higher than the existing store. Despite their claim of effectively reinventing the science of retailing, the veteran managers maintain that the new store is reporting slightly higher sales because of the novelty factor, which they believe will soon wear off. In addition, they think that statistically speaking, the new store sales are no different from the sample of existing 35 stores.

The question, therefore, is to determine whether the new store sale figures are different from the sales at the existing stores. Because the franchise surveyed 35 of its numerous stores, we do not know the standard deviation of sales in the entire population of stores. Thus, I will rely on t-distribution, and not Normal distribution.

Average sales per week for the 35 stores: \$166,000 (μ)

Std. Dev of the weekly sales: \$25,000 (s)

Average sales reported by the new store: \$170,000 ($\overline{x}$)

Our null hypothesis: H_0: $\mu = 166000$

Alternative hypothesis: H_a: $\mu \neq 166000$

Because we are not making an assumption about the sales in the new store being higher or lower than the average sales in the existing stores, we are using a two-tailed test. The purpose is to test the hypothesis that the new stores sales are different from that of the existing store sales. Mathematically:

$$t = \frac{170000 - 166000}{\frac{25000}{\sqrt{35}}} = 0.947$$

The estimated value of the `t-statistics` is 0.947. Figure 6.25 shows a graphical representation of the test.

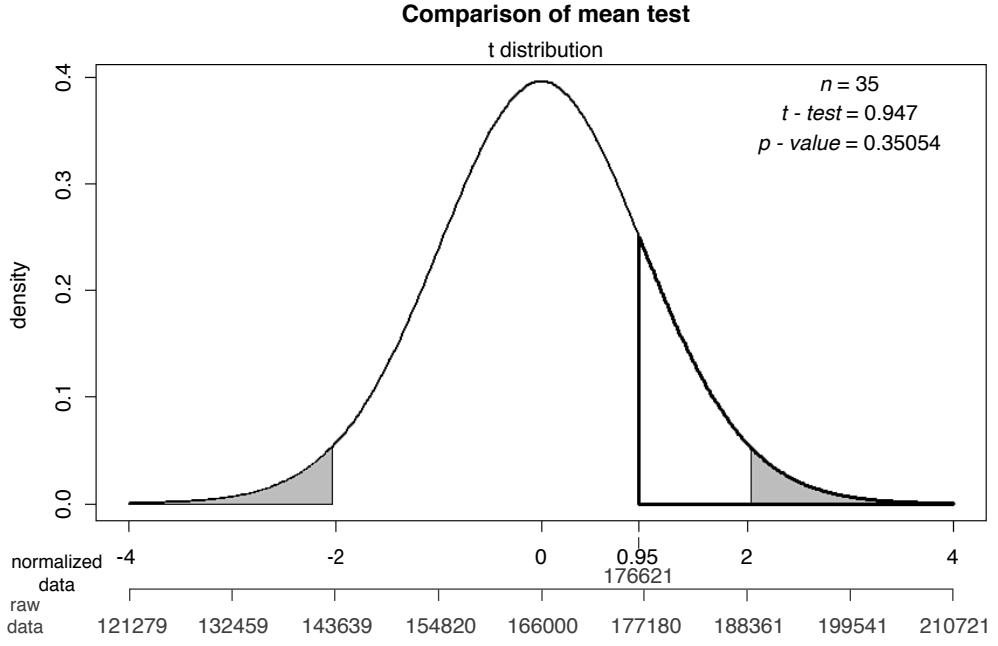


Figure 6.25 Retail sales and hypothesis testing

We see that the new store sales do not fall in the rejection region (shaded gray). Furthermore, the `p-value` for the test is 0.35, which is much higher than the threshold value of 0.05. We therefore fail to reject the null hypothesis and conclude that the new store sales are similar to the ones reported for the 35 sample stores. Thus, the new store manager may not have reinvented the science of retailing.

Comparing Two Means with Unequal Variances

In most applied cases of statistical analysis, we compare the means for two or more groups in a sample. The underlying assumption in this case is that the two means are the same and thus the difference in means equals 0. We can conduct the test assuming the two groups might or might not have equal variances.

I illustrate this concept using data for teaching evaluations and the students' perceptions of instructors' appearance. Recall that the data covers information on course evaluations along with course and instructor characteristics for 463 courses for the academic years 2000–2002 at the University of Texas at Austin.

You are encouraged to download the data from the book's website. A breakdown of male and female instructors' teaching evaluation scores is presented in Table 6.5.

```
t.mean<-tapply(x$eval,x$gender, mean)
t.sd<- tapply(x$eval,x$gender, sd)
round(cbind(mean=t.mean, std. dev.=t.sd),2)
```

Table 6.5 Teaching Evaluation for Male and Female Instructors

	mean	std.dev.
Male	4.07	0.56
Female	3.90	0.54

We notice that the teaching evaluations of male instructors are slightly higher than that of the female instructors. We would like to know whether this difference is statistically significant.

Hypothesis:

$$H_0: x_1 = x_2$$
$$H_a: x_1 \neq x_2$$

I conduct the test to determine the significance in the difference in average values for a particular characteristic of two independent groups, as shown in Equation 6.6.

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ **Equation 6.6**

The shape of the t-distribution depends on the degrees of freedom, which according to Satterthwaite (1946)[5] are calculated as shown in Equation 6.7:

$$dof = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Equation 6.7

Substituting the values in the equation, I have the following:

$s_1 = .56$

$s_2 = .54$

$n_1 = 268$

$n_2 = 195$

$x_1 = 4.07$

$x_2 = 3.90$

Subscript 1 represents statistics for males, and subscript 2 represents statistics for females. The results are as follows: $dof = 426$ and $t = 3.27$. The output from R is presented in Figure 6.26.

```
> t.test(eval~gender, alternative='two', conf.level=.95,
+         var.equal=FALSE, data=TeachingRatings)

        Welch Two Sample t-test

data:  eval by gender
t = 3.2667, df = 425.756, p-value = 0.001176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06691754 0.26909088
sample estimates:
  mean in group male mean in group female
            4.069030              3.901026


> pt(c(3.267), df=425.76, lower.tail=FALSE)*2
[1] 0.001174933
```

Figure 6.26 Output of a t-test in R

Figure 6.27 shows the graphical output:

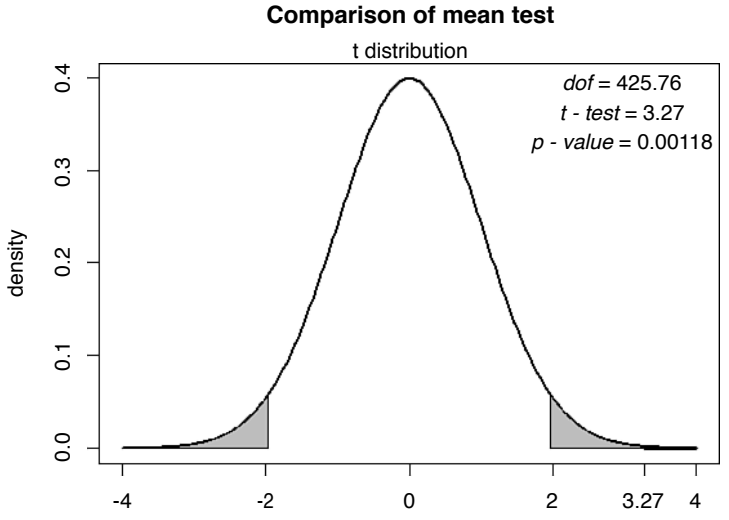


Figure 6.27 Graphical depiction of a two-tailed `t-test` for teaching evaluations

I obtain a `t-value` of 3.27, which falls in the rejection region. I also notice that the `p-value` for the test is 0.0018, which suggests rejecting the null hypothesis and conclude that the difference in teaching evaluation between male and female instructors is statistically significant at the 95% level.

Conducting a One-Tailed Test

The previous example tested the hypothesis that the average teaching evaluation for males and females was not the same. Now, I adopt a more directional approach and test whether the teaching evaluations for males were higher than that of the females: H_a: $x_1 > x_2$.

Given that it is a one-sided test, the only thing that changes from the last iteration is that the rejection region is located only on the right side (see Figure 6.28). The t-value and the associated degrees of freedom remain the same. What changes is the `p-value`, because the rejection region, representing 5% of the area under the curve, lies to only the right side of the distribution. The probability value will account for the possibility of getting a `t-value` of 3.27 or higher, which is different from the two-tailed test where I calculated the `p-value` of obtaining a `t-value` of either lower than –3.27 or greater than 3.27.

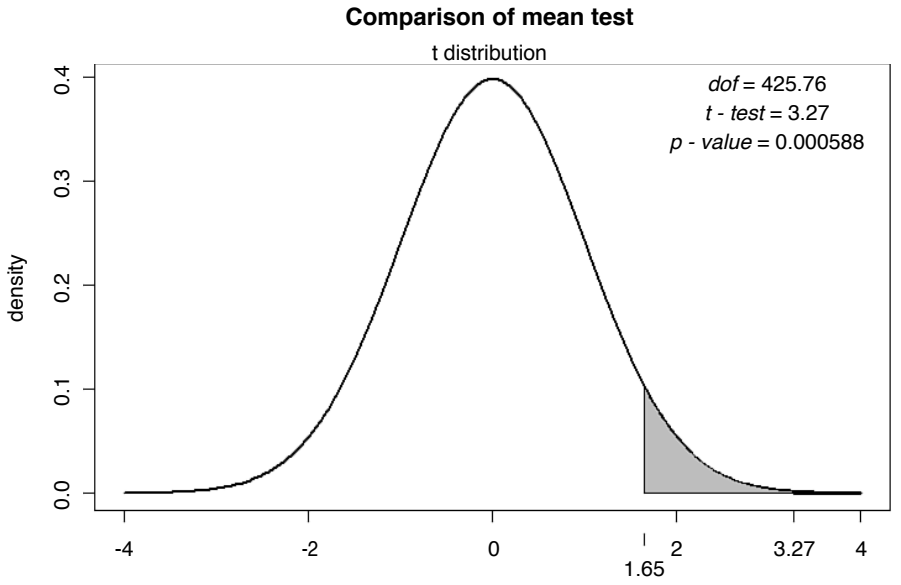


Figure 6.28 Teaching evaluations, right-tailed test

The resulting `p-value` is 0.00058, which is a lot less than 0.05, our chosen threshold to reject the null hypothesis. I thus reject the null and conclude that male instructors indeed receive on average higher teaching evaluations than female instructors do.

Figure 6.29 shows the output from R for a one-tailed test.

```
> t.test(eval~gender, alternative='greater', conf.level=.95,
+         var.equal=FALSE, data=TeachingRatings)

        Welch Two Sample t-test

data:  eval by gender
t = 3.2667, df = 425.756, p-value = 0.000588
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0832263       Inf
sample estimates:
  mean in group male mean in group female
            4.069030              3.901026


> pt(c(3.267), df=425.76, lower.tail=FALSE)
[1] 0.0005874666
```

Figure 6.29 R output for teaching evaluations, right-tailed test

Comparing Two Means with Equal Variances

When the population variance is assumed to be equal between the two groups, the sample variances are pooled to obtain an estimate of σ. Use Equation 6.8 to get the standard deviation of the sampling distribution of the means:

$$sdev = \sqrt{\frac{vpool * (n_1 + n_2)}{n_1 * n_2}} \quad \textbf{Equation 6.8}$$

Equation 6.9 provides the pooled estimate of variance:

$$vpool = \frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{n_1 + n_2 - 2} \quad \textbf{Equation 6.9}$$

Get the test statistics via Equation 6.10:

$$t = \frac{x_1 - x_2}{sdev} \quad \textbf{Equation 6.10}$$

I use the same example of teaching evaluations to determine the difference between the evaluation scores of male and female instructors assuming equal variances. The calculations are reported as follows:

$$vpool = \frac{.557^2 (268 - 1) + .539^2 (195 - 1)}{268 + 195 - 2} = .302$$

$$sdev = \sqrt{\frac{.302(268 + 195)}{268 * 195}} = .052$$

$$t = \frac{4.069 - 3.901}{.052} = 3.250$$

The degrees of freedom for equal variances are given by $dof = n_1 + n_2 - 2$.

Figure 6.30 shows the R output:

```
        Two Sample t-test

data:  eval by gender
t = 3.2499, df = 461, p-value = 0.001239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06641797 0.26959045
sample estimates:
  mean in group male mean in group female
            4.069030             3.901026
```

Figure 6.30 R output for equal variances, two-tailed test

Figure 6.31 presents the graphical display.

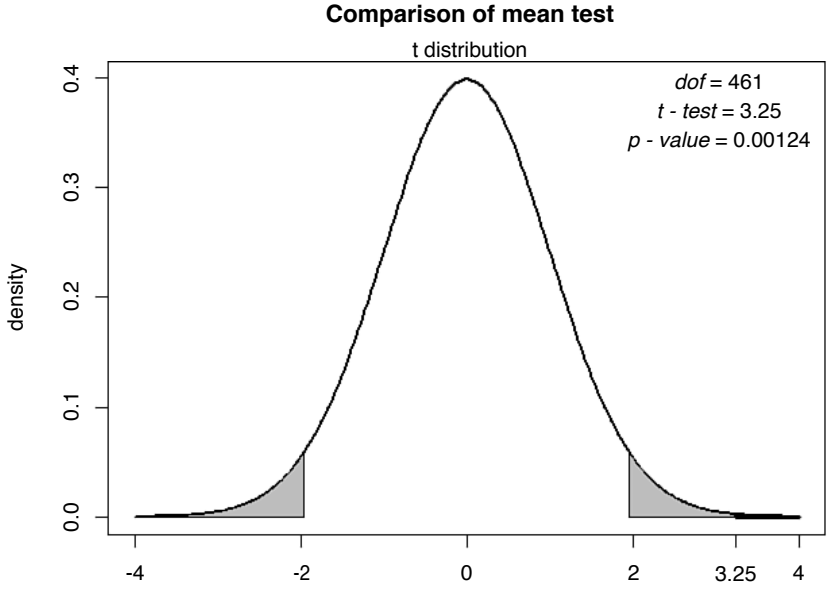


Figure 6.31 Graphical output for equal variances, two-tailed test

Note that the results are similar to what I obtained earlier for the test conducted assuming unequal variances. The `t-value` is 3.25 and the associated `p-value` is 0.00124. I reject the null hypothesis and conclude that the average teaching evaluations for males are different from that of the females. These results are statistically significant at the 95% (even 99%) level.

I can repeat the analysis with a one-tailed test to determine whether the teaching evaluations for males are statistically higher than that for females. I report the R output in Figure 6.32. The associated `p-value` for the one-tailed test is 0.0006194, which suggests rejecting the null hypothesis and conclude that the teaching evaluations for males are greater than that of the females.

```
        Two Sample t-test

data:  eval by gender
t = 3.2499, df = 461, p-value = 0.0006194
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.08280296        Inf
sample estimates:
  mean in group male mean in group female
            4.069030              3.901026


> pt(c(3.2498), df=461, lower.tail=FALSE)
[1] 0.0006196742
```

Figure 6.32 R output for equal variances, right-tailed test

Figure 6.33 shows the graphical display.

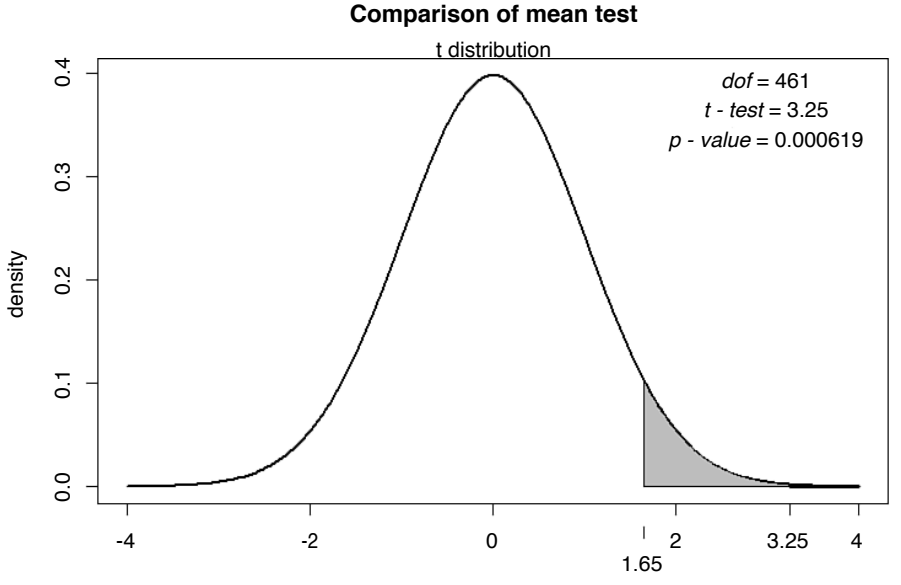


Figure 6.33 Graphical output for equal variances, right-tailed test

Worked-Out Examples of Hypothesis Testing

The best way of mastering a new concept is to practice the concept with more examples. Here I present additional worked-out examples of the `t-test`. I encourage you to repeat the analysis presented in these examples with a handheld calculator to get a real feel of the concepts.

Best Buy–Apple Store Comparison

Assuming unequal variances, let us see whether the average daily sales between fictional versions of a Best Buy (BB) outlet and an Apple Store (AS) are statistically different. Here are some cooked up numbers.

- **BB:** Average daily sales $110,000, SD $5000, sales observed for 65 days
- **AS:** Average sales $125,000, SD $15,000, sales observed for 45 days

Recall that we conduct a `t-test` to determine the significance in the difference in means for a particular characteristic of two independent groups. Equation 6.6 presents the `t-test` formula for **unequal variances**.

Subscript 1 represents statistics for Best Buy, and subscript 2 represents statistics for the Apple Store. The shape of the t-distribution depends on the degrees of freedom, which was presented earlier in Equation 6.7.

The *dof* are needed to determine the probability of obtaining a t-value as extreme as the one calculated here. When I plug in the numbers in the equations, I obtain the following results:

t = –6.46

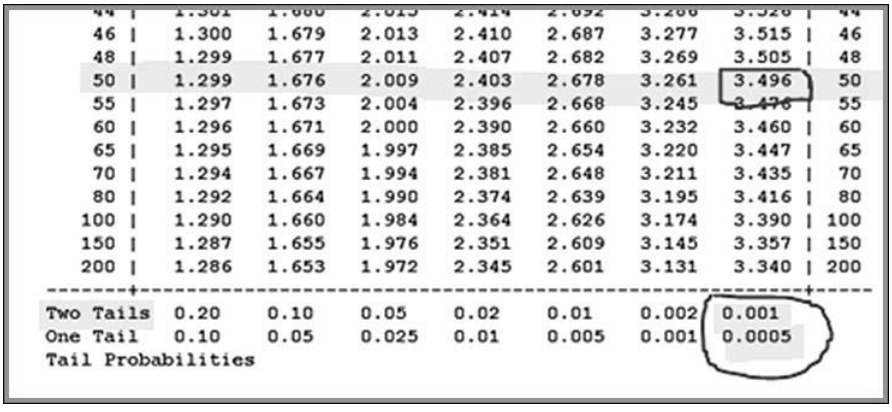
dof = 50.82

The absolute `t-stat` value of –6.4 is significantly greater than the absolute value for +/–1.96, suggesting that the Best Buy outlet and the Apple Store sales are significantly different (using a two-tailed test) for large samples at the 95% confidence level. Let us obtain the critical value for the t-test that is commensurate with the appropriate sample size.

If I were to consult the t-table for *dof* = 50, (the closest value to 50.8) the largest t-value reported is 3.496, which is less than the one I have obtained (–6.46). Note that I refer here to the absolute value of –6.46. The corresponding probability value from the t-table for a two-tailed test for the maximum reported `t-test` of 3.496 is 0.001 (see the highlighted values in the image from the t-table in Figure 6.34). Thus, I can conclude that the probability of finding such an extreme t-value is less than 0.001% (two-tailed test).

I would now like to test whether Best Buy Store sales are lower than that of the Apple Stores. A one-tailed test will help us determine whether the average sales at the fictional versions of BestBuy are lower than at the Apple Store. The probability to obtain a t-value of –6.46 or higher is 0.0005% (see Figure 6.34). This suggests that the fictional BestBuy average daily sales are significantly lower than that of the Apple Store.

A `p-value` of less than 0.05 leads us to reject the null hypothesis that $x_1 > x_2$ and conclude that the Best Buy sales are lower than that of the Apple Store at the 95% confidence level.



46	1.300	1.679	2.013	2.410	2.687	3.277	3.515	46
48	1.299	1.677	2.011	2.407	2.682	3.269	3.505	48
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496	50
55	1.297	1.673	2.004	2.396	2.668	3.245	3.476	55
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460	60
65	1.295	1.669	1.997	2.385	2.654	3.220	3.447	65
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435	70
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416	80
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390	100
150	1.287	1.655	1.976	2.351	2.609	3.145	3.357	150
200	1.286	1.653	1.972	2.345	2.601	3.131	3.340	200
Two Tails	0.20	0.10	0.05	0.02	0.01	0.002	0.001	
One Tail	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Tail Probabilities								

Figure 6.34 T-distribution table

Assuming Equal Variances

I repeat the preceding example now assuming equal variances. Equation 6.8 provides the standard deviation of the sampling distribution of the means. Equation 6.9 describes the pooled estimate of variance, whereas Equation 6.10 describes the test statistics.

Here is the R code:

```
vpool= (s1^2*(n1-1)+s2^2*(n2-1))/(n1+n2-2);
sdev = sqrt(vpool*(n1+n2)/(n1*n2))
t = (x1-x2)/sdev ;
dof= n1+n2-2
t =  -7.495849
dof =108
vpool =  106481481
sdev = 2001.108
```

Notice that when I assume equal variances, I get an even stronger `t-value` of –7.49, which is again much larger than the absolute value for 1.96. Hence, I can conclude both for one- and two-tailed tests that Best Buy sales (assumed values) on average are lower than that of the average daily (fictional) sales for the Apple Store.

The closest value on the t-table for 108 (*dof*) is 100. The highest value reported along the 100 *dof* row is 3.390, which is lower than the absolute value of –7.49, which I estimated from the test. This suggests that the probability for a two-tailed test will be even smaller than 0.001 and for a one-tailed test will be less than 0.0005, allowing us to reject the null hypothesis.

Comparing Sales from an Urban and Suburban Retailer

A franchise operates stores in urban and suburban locations. The managers of two stores are competing for promotion. The manager at the suburban store is liked by all while the manager in the downtown location has a reputation of being a hot-head. The stores' weekly sales data over a 50-week period are as follows:

Downtown Store	Suburban Store
Average weekly sales = $800,000	Average weekly sales = $780,000
std dev = $100,000	std dev = $30,000
$n_1 = 50$	$n_2 = 50$

Hypothesis:
H_0: $x_1 = x_2$
H_a: $x_1 \neq x_2$
Assuming unequal variances:

$$t = \frac{20000}{\sqrt{\frac{100000^2}{50} + \frac{30000^2}{50}}}$$

The calculations return t = 1.35 and the *dof* = 57.75. The `p-value` for the test from the t-distribution table is approximately equal to 0.15. I therefore fail to reject the null hypothesis and conclude that both stores on average generate the same revenue. Thus, the manager of the suburban store, who happens to be a nice person but appears to be selling $20,000 per week less in sales, could also be considered for promotion because the `t-test` revealed that the difference in sales was not statistically significant. Also remember that t = 1.35 should have made the conclusion easier because the calculated `t-value` is less than 1.96 for a two-tailed test.

Exercises for Comparison of Means

Using the teaching ratings data, answer the following questions:

1. Determine the mean and standard deviation for course evaluations for minority and non-minority instructors. Determine whether the instructors belonging to minority groups are more or less likely to obtain a course evaluation of 4.1 or higher.
2. Determine the mean and standard deviation for course evaluations for upper- and lower-level courses. Determine whether the probability of obtaining a below-average course evaluation is higher for lower-level courses. Use 3.999 as the average course evaluation.
3. Determine whether tenured professors receive above-average course evaluations.

Regression for Hypothesis Testing

So far, I have relied on the traditional tools for hypothesis testing that are prescribed in texts for statistical analysis and data science. I would argue that regression analysis, which I explain in

detail in Chapter 7, "Why Tall Parents Don't Have Even Taller Children," could also be used to compare means of two or more groups. I favor regression analysis over other techniques primarily because of the simplicity in its application, which is a desired feature for most data scientists.

Let us focus on the teaching ratings example where we determine whether the average teaching evaluation differed for males and females. I have noted earlier that the average teaching evaluation for female instructors was 3.90 and for males 4.07.

Assuming equal variances, I conducted a `t-test` and concluded that a statistically significant difference in teaching evaluations existed for males and females (see Figure 6.35).

```
> t2 <- t.test(eval~gender,var.equal = TRUE, data=x); t2

 Two Sample t-test

data:  eval by gender
t = 3.2499, df = 461, p-value = 0.001239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06641797 0.26959045
sample estimates:
  mean in group male mean in group female
            4.069030              3.901026
```

Figure 6.35 Equal variances two-tailed `t-test` to determine gender-based differences in teaching evaluations

The `p-value` of 0.0012 suggests that I can reject the null hypothesis of equal means and conclude that the average teaching evaluations between male and female instructors differ. Now let us attempt the same problem using a regression model. Figure 6.36 presents the output from the regression model.

Note that the column `tvalue` under Coefficients reports 3.25 for the row labelled as gen2male. This statistic is identical to the `t-value` obtained in the traditional `t-test` assuming equal variances. Furthermore, the column labelled `Pr(>|t|)` reports 0.00124 as the `p-value` for gen2male, which is again identical to the `p-value` reported in the `t-test`. Thus, we can see that when we assume equal variances, a regression model generates identical results.

```
> summary(lm(eval~gen2, data=x))

Call:
lm(formula = eval ~ gen2, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96903 -0.36903  0.03097  0.43097  0.99897

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.90103    0.03933   99.19  < 2e-16 ***
gen2male     0.16800    0.05169    3.25  0.00124 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5492 on 461 degrees of freedom
Multiple R-squared:  0.0224, Adjusted R-squared:  0.02028
F-statistic: 10.56 on 1 and 461 DF,  p-value: 0.001239
```

Figure 6.36 Regression model output for gender differences in teaching evaluations

The real benefit of the regression model emerges when we compare the means for more than two groups. `t-tests` are restricted to comparison of two groups. Regression models are useful when the null hypothesis states that the average values are the same for multiple groups.

To illustrate this point, I have categorized the age variable into a factor variable with three categories namely young, mid-age, and old. I would like to know whether a statistically significant difference exists between the teaching evaluation scores for young, mid-age, and old instructors. Again, I estimate a simple regression model (see Figure 6.37) to test the hypothesis. Figure 6.37 shows the results, and the R code follows.

```
x$f.age <- cut(x$age, breaks = 3)
x$f.age <- factor(x$f.age,labels=c("young", "mid age", "old"))
cbind(mean.eval=tapply(x$eval,
x$f.age,mean),observations=table(x$f.age))
plot(x$age,x$eval,pch=20)
summary(lm(eval~f.age, data=x))
```

```
> summary(lm(eval~f.age, data=x))

Call:
lm(formula = eval ~ f.age, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.78125 -0.38125  0.01875  0.46396  1.01875

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.00435    0.04361  91.831   <2e-16 ***
f.agemid age  0.03169    0.05727   0.553    0.580
f.ageold     -0.12310    0.07568  -1.626    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5533 on 460 degrees of freedom
Multiple R-squared:  0.00997, Adjusted R-squared:  0.005665
F-statistic: 2.316 on 2 and 460 DF,  p-value: 0.09981
```

Figure 6.37 Regression model output for teaching evaluations based on age differences

I see that the values reported under the column labelled `Pr(>|t|)` in Figure 6.37 for the two categories of age; namely *f.age mid age* and *f.age old* are greater than .05. I therefore fail to reject the null hypothesis and conclude that average teaching evaluations do not differ by age in our sample.

Note that mid-age and old-age are represented in the model, whereas the category young is missing from the model. In regression models, when factor variables are used as explanatory variables, one arbitrarily chosen category, in this case young, is omitted from the output, and is used as the base against which other categories are compared.

The purpose here is not to explain the intricacies of regression models. Instead, my intent is to indicate a possible use of regression models for hypothesis testing. I do, however, explain the workings of regression models in Chapter 7.

Analysis of Variance

Analysis of variance, ANOVA, is the prescribed method of comparing means across groups of three or more. The null hypothesis in this case states that the average values do not differ across the groups. The alternative hypothesis states that at least one mean value is different from the rest.

I use the F-test for ANOVA. If the probability (`p-value`) associated with the F-test is greater than the threshold value, which is usually .05 for the 95% confidence level, we fail to reject the null hypothesis. In instances where the probability value for the F-test is less than .05, we reject the null hypothesis. In such instances, we conclude that at least one mean value differs from the rest.

I will repeat the comparison of means for the three age groups using the ANOVA test. The R code and the resulting output (see Figure 6.38) follow.

```
> summary(aov(eval~f.age, data=x))
             Df Sum Sq Mean Sq F value Pr(>F)
f.age         2   1.42  0.7090   2.316 0.0998 .
Residuals   460 140.82  0.3061
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6.38 ANOVA output for influence of age on teaching evaluations

Note that the value reported under `Pr(>F)` is 0.0998, which is greater than 0.05. Thus, we fail to reject the null hypothesis and conclude that the teaching evaluations do not differ by age groups.

Let us test the average teaching evaluations for a discretized variable for beauty, which in raw form is a continuous variable. I convert the continuous variable into three categories namely: low beauty, average looking, and good looking. The R code and the resulting output (see Figure 6.39) follow.

```
x$f.beauty<-cut(x$beauty, breaks=3)
x$f.beauty<-factor(x$f.beauty, labels=c("low beauty", "average
            looking", "good looking"))
cbind(mean.eval=tapply(x$eval,x$f.beauty,mean),
            observations=table(x$f.beauty))
summary(aov(eval~f.beauty, data=x))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
f.beauty      2    2.2  1.1013   3.618 0.0276 *
Residuals   460  140.0  0.3044
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6.39 ANOVA output for influence of beauty on teaching evaluations

The probability value associated with the F-test is 0.0276, which is less than .05, our threshold value. I therefore reject the null hypothesis and conclude that teaching evaluations differ by students' perception of instructors' appearance.

Significantly Correlated

Often we are interested in determining the independence between two categorical variables. Let us revisit the teaching ratings data. The university administration might be interested to know whether the instructor's gender is independent of the tenure status. This is of interest because in the presence of a gender bias, we might find that a larger proportion of women (or men) have not been granted tenure. A chi-square test of independence can help us with this challenge.

The null hypothesis (H_0) states that the two categorical variables are statistically independent, whereas the alternative hypothesis (H_a) states that the two categorical variables are statistically dependent. The test statistics is expressed shown in Equation 6.11.

$$\chi^2 = \Sigma\left(\frac{\left(f_o - f_e\right)^2}{f_e}\right) \quad \textbf{Equation 6.11}$$

Where f_o is the observed frequency, and f_e is the expected frequency. We reject the null hypothesis if the `p-value` is less than the threshold for rejection ($1\text{-}\alpha$) and the degrees of freedom.

Let us test the independence assumption between gender and tenure in the teaching ratings data set. My null hypothesis states that the two variables are statistically independent. I run the test in R and report the results in Figure 6.40. Because the `p-value` of 0.1098 is greater than 0.05, I fail to reject the null hypothesis that the two variables are independent and conclude that a systematic association *does* exist between gender and tenure.

```
> t1<-table(x$gender,x$tenure);t1

         no yes
  male   52 216
  female 50 145
> round(prop.table(t1,1)*100,2)

           no   yes
  male   19.40 80.60
  female 25.64 74.36
> chisq.test(t1, correct=F)

 Pearson's Chi-squared test

data:  t1
X-squared = 2.5571, df = 1, p-value = 0.1098
```

Figure 6.40 Pearson's chi-squared test to determine association between gender and tenure status of instructors

We can easily reproduce the results in a spreadsheet or statistics software. The f_e in the formula is calculated as follows:

1. Determine the row and column totals for the contingency table (t1 in the last example: see the following code)
2. Determine the sum of all observations in the contingency table
3. Multiply the respective row and column totals and divide them by the sum of all observations to obtain f_e.

The R code required to replicate the programmed output follows.

```
t1<-table(x$gender,x$tenure);t1
round(prop.table(t1,1)*100,2)
r1<-margin.table(t1, 1) #  (summed over rows)
c1<-margin.table(t1, 2) #  (summed over columns)
r1;c1
e1<-r1%*%t(c1)/sum(t1);e1
t2<-(t1-e1)^2/e1;t2;sum(t2)
qchisq(.95, df = 1)
1-pchisq(sum(t2),(length(r1)-1)*(length(c1)-1))
```

Summary

Let me hypothesize in the concluding section of this chapter that you are now at least familiar with the statistical concepts about testing assumptions and hypotheses. The process of stating one's assumptions and then using statistical methods to test them is at the core of statistical analysis. I would like to conclude this section with a warning or two about the limitations of statistical analysis. As budding data scientists, you may naively assume that the techniques you have learned can be applied to all problems. Such a conclusion would be erroneous.

Recall the story of European settlers who spotted a black swan in Western Australia that immediately contradicted their belief that all swans were white. The settlers could have treated the black swan as an outlier, a data point that is very different from the rest of the observations. They could have ignored this one observation. But that would have been a mistake, because in this particular case, a black swan challenged the existing knowledge base.

Let me explain this with an example of when an outlier/s might be ignored. Assume you are working with the housing sales data where the average sale price in the neighborhood is around $450,000. However, you may have a couple or more housing units in the same data set that sold for more than two million dollars each. Given the nature of the housing stock in the neighborhood, you might conclude that a very small number of housing units in the area are much larger in size than the rest of the housing stock and hence have transacted for a larger amount. Because you are interested in forecasting the average price of an average house in the neighborhood, you might declare the very expensive transactions as outliers and exclude those from the analysis.

Now let us assume that you were (in a previous life) Charles Darwin's assistant and assigned the task to document the colors of swans found on the planet. As you landed in Western Australia with the rest of the settlers, you also spotted a black swan. Would you have treated the black swan as an outlier? The answer is emphatically no. Just one out of ordinary outcome or observation that could not be foreseen based on our prior body of knowledge is not an outlier, but the most important observation to ponder in detail.

Similarly, I would like to draw your attention again to Professor Jon Danielsson's estimation that an S&P 500 single-day decline of 23% in 1987 would happen once out of every 12 universes. We know that financial market meltdowns of similar proportions happen at a more rapid frequency than the statistical models would allow us to believe. Our continued reliance on the Gaussian distribution, which we refer to as the Normal distribution, erroneously lead us to believe that natural phenomenon can be approximated using the Normal distribution. This erroneous assumption is behind our poor risk perception of natural disasters and overconfidence in financial markets.

I submit that a data scientist is not one who believes the use of algorithms and statistical methods will provide him or her with "the" answer. Instead, I believe a data scientist is one who is fully cognizant of one's innate inability in predicting the future. A data scientist is one who appreciates the analytics will deliver an informed possible view of the future out of many other possible incarnations. A data scientist is one who never becomes a victim of compound ignorance; that is, the state when one is ignorant of one's own ignorance.

Endnotes

1. Wente, M. (2008, November 22). "A black swan comes home to roost." *The Globe and Mail*. Retrieved from http:// www.theglobeandmail.com/news/national/a-black-swan-comes-home-to-roost/article 716947/.
2. Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House Publishing Group.
3. Danielsson, J. (2011). *Financial Risk Forecasting: The Theory and Practice of Forecasting Market Risk with Implementation in R and Matlab.* John Wiley & Sons.
4. http://www.basketball-reference.com/leaders/pts_per_g_career.html
5. Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2: 110–114.

This page intentionally left blank

Index

B

C

D

E

F

G

H

I

J

K

L

M

N

P

Q

R

S

T

U

V

W

X

Y

Z