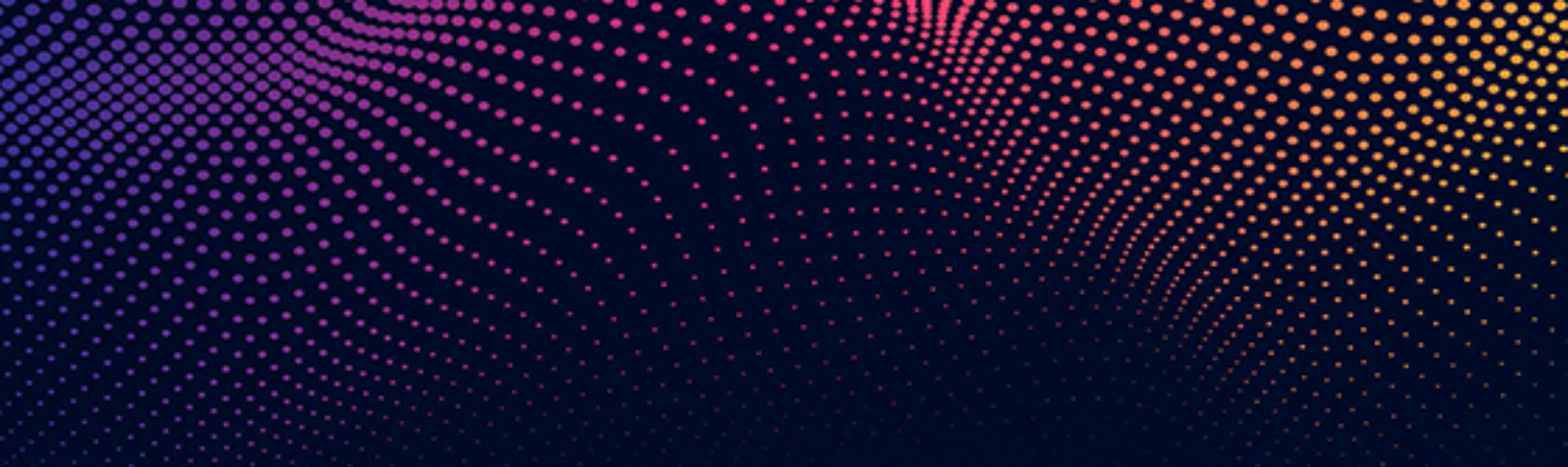




Statistics and Data Foundations for AI

Tamraparni Dasu
Geetha Murthy

A **Chapman & Hall** Book



CRC Press
Taylor & Francis Group



Statistics and Data Foundations for AI

Tamraparni Dasu
Geetha Murthy



A Chapman & Hall Book



CRC Press
Taylor & Francis Group

“Statistics and Data Foundations for AI provides a rigorous yet accessible introduction to the statistical theory and data analysis essential for understanding and using modern AI. The book is grounded in real-world examples from authoritative sources such as NASA and NOAA, and thereby connects core statistical concepts directly to contemporary AI applications. Its distinctive emphasis on data quality, ethics, bias, and the limitation of AI equips readers to engage critically and responsibly with AI systems. Suitable for both STEM and non-STEM audiences, the text supports hands-on learning through tools such as R, Excel, and Google Sheets, without assuming prior programming experience.”

Sivaramakrishnan “Bala” Balachandar, Chairman and William F. Powers Professor of Mechanical and Aerospace Engineering at the University of Florida

“This slender volume creatively fills the gap between Data Management, Statistics, and Artificial Intelligence, proving once again that less can be more. It embraces what many educators avoid: The use of Large Language Models (here: Google Gemini) as partners in education. Demonstrations and exercises are built on real, large datasets, not on the toy datasets from yesteryear. While being scientifically rigorous, it builds on the reader's intuition, to the point of making “Data Intuition” a primary topic of discussion.”

James Geller, Professor and Chair, Department of Data Science, Ying Wu College of Computing, New Jersey Institute of Technology (NJIT)

“This engaging textbook, richly laden with real world examples and case studies, gives an intuitive yet thorough view of probability and statistics, and of how we can use them to understand the stochasticity in our world, the uncertainty in our knowledge, and the workings of artificial intelligence. This is a timely book that deserves a very wide audience.”

Sanjoy Dasgupta, Professor, Jacobs School of Engineering, University of California San Diego

“This amazing book is written both as a textbook for data science and AI learners and educators, as well as a reference for more experienced users. Using a pedagogical approach targeting adult

learners, the authors have explained the various aspects of AI, statistical foundations of AI and LLMs, and provided examples, practice exercises, and links to additional learning from current sources. The book provides thoughtful questions for discussion, illustrates the strengths and weaknesses of AI, and engages the reader with descriptions and links to applications from a wide variety of fields. It is well-researched and written in accessible yet scholarly language. As a long-time teacher of statistics and now engaged in teaching with AI, I found the book extremely valuable. I would highly recommend this book for college instructors, individuals who are self-teaching AI use, and researchers seeking to extend their knowledge of AI applications in their work. Congratulations to Dasu and Murthy on this excellent book!”

Rene Parmar, Dean, School of Education, SUNY New Paltz

“At a time when artificial intelligence is often perceived as a “black box,” *Statistics and Data Foundations for AI* authored by Tamraparni Dasu and Geetha Murthy offers the essential map and compass needed to explore its inner mechanics. By positioning data as the fuel that drives every AI model—from training and learning to inference—the authors establish a robust foundation in data collection, storage, and management, addressing core concepts that are frequently underrepresented in traditional curricula.

What distinguishes this text is its forward-thinking pedagogy. The Data Debate features encourage students to engage deeply and thoughtfully with complex statistical concepts through structured, nonjudgmental discourse. Complementing these are the GeminiBox sections, which exemplify active learning in action—guiding students to craft precise AI prompts and, importantly, to evaluate model outputs critically for accuracy and potential “hallucinations.” Together, these elements cultivate both technical proficiency and data literacy essential for working with modern AI tools.

The book succeeds in bridging theory with real-world relevance. From analyzing multivariate factors influencing hurricane trajectories to addressing ethical considerations and sampling bias, the material remains grounded in authentic data and contemporary challenges. The integration of rich visuals, real datasets, and well-designed summative projects ensures that students not only understand core principles but can apply them to meaningful, data-driven inquiries.

Finally, the inclusion of a 125-year historical perspective on the co-evolution of AI and hardware is a masterstroke. It provides invaluable context, reminding readers that today's AI revolution is the product of a century of innovation in computation and data science. *Statistics and Data Foundations for AI* is an indispensable resource for educators, students, and practitioners seeking to understand not just where AI stands today, but the data and statistical foundations that will shape its future.”

Kiron Sharma, Professor of Computer Science, Interim Chair, Dept. Computer Science & Mathematics, Olsen College of Engineering and Science, Farleigh Dickinson University

Statistics and Data Foundations for AI is an interdisciplinary approach to statistical concepts and data foundations of AI with real-world illustrative examples from authoritative sources such as NASA, NOAA and the United States Census Bureau. Co-authored by a data science research expert and an experienced educator, the book serves as a prequel to an AI and machine learning course.

Given the interdependence of data and AI, understanding data and using it responsibly to create and interact with AI tools requires a high level of statistical skill and data intuition. The book includes topics such as data management, exploratory data analysis, sampling, probability theory, hypothesis testing, multivariate analysis, data quality, ethics, data privacy, and responsible use of AI. Every key statistical concept is presented in the context of how it is used by AI applications in areas such as sports, fashion, climate science, environmental science, health, medicine, and space exploration. The book makes AI relatable to everyday life so that it is no longer an abstraction. Instructor resources, supplementary materials, further reading, and debate topics enable advanced study and deeper thinking.

Statistics and Data Foundations for AI is intended for undergraduate and graduate students, and practitioners interested in learning statistical foundations in relation to data and AI with application to real-world problems. The content is accessible to learners from a wide variety of backgrounds (STEM and non-STEM) without sacrificing rigor.

Statistics and Data Foundations for AI

Statistics and Data Foundations for AI is an interdisciplinary approach to statistical concepts and data foundations of AI with real-world illustrative examples from authoritative sources such as NASA, NOAA and the United States Census Bureau. Co-authored by a data science research expert and an experienced educator, the book serves as a prequel to an AI and machine learning course.

Given the interdependence of data and AI, understanding data and using it responsibly to create and interact with AI tools requires a high level of statistical skill and data intuition. The book includes topics such as data management, exploratory data analysis, sampling, probability theory, hypothesis testing, multivariate analysis, data quality, ethics, data privacy, and responsible use of AI. Every key statistical concept is presented in the context of how it is used by AI applications in areas such as sports, fashion, climate science, environmental science, health, medicine, and space exploration. The book makes AI relatable to everyday life so that it is no longer an abstraction. Instructor resources, supplementary materials, further reading, and debate topics enable advanced study and deeper thinking.

Statistics and Data Foundations for AI is intended for undergraduate and graduate students, and practitioners interested in learning statistical foundations in relation to data and AI with application to real-world

problems. The content is accessible to learners from a wide variety of backgrounds (STEM and non-STEM) without sacrificing rigor.

Statistics and Data Foundations for AI

Tamraparni Dasu and Geetha Murthy



CRC Press

Taylor & Francis Group

Boca Raton London New York

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

First edition published 2026

by CRC Press

2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press

4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

© 2027 Tamraparni Dasu and Geetha Murthy

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

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

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

For Product Safety Concerns and Information please contact our EU representative GPSR@taylorandfrancis.com. Taylor & Francis Verlag GmbH, Kaufingerstraße 24, 80331 München, Germany.

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

Library of Congress Cataloging-in-Publication Data

Names: Dasu, Tamraparni author | Murthy, Geetha author

Title: Statistics and data foundations for AI/Tamraparni Dasu and Geetha Murthy.

Description: First edition. | Boca Raton FL: CRC Press, 2026. | Includes bibliographical references.

Identifiers: LCCN 2026015351 (print) | LCCN 2026015352 (ebook) | ISBN 9781041006428 hardback | ISBN 9781041006411 paperback | ISBN 9781003610922 ebook

Subjects: LCSH: Artificial intelligence--Statistical methods | Statistics--Data processing

Classification: LCC Q335. D3634 2026 (print) | LCC Q335 (ebook)

LC record available at <https://lcn.loc.gov/2026015351>

LC ebook record available at <https://lcn.loc.gov/2026015352>

ISBN: 978-1-041-00642-8 (hbk)

ISBN: 978-1-041-00641-1 (pbk)

ISBN: 978-1-003-61092-2 (ebk)

DOI: [10.1201/9781003610922](https://doi.org/10.1201/9781003610922)

Typeset in Latin Modern font

by KnowledgeWorks Global Ltd.

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

*To our families for their constant love, support
and encouragement.*

Contents

[Preface](#)

[QR Code to Access the URLs in the Textbook](#)

[Authors](#)

[1 Statistics, Data and AI](#)

[1.1 Why Statistics and Data Foundations?](#)

[1.2 Special Features of the Book](#)

[1.3 Statistics, Data and AI](#)

[1.3.1 About Data, Chapter 2: Monitoring weather data with AI](#)

[1.3.2 Exploratory Data Analysis, Chapter 3: How soccer clubs use AI](#)

[1.3.3 Sampling, Chapter 4: AI speeds up financial trading](#)

[1.3.4 Probability, Chapter 5: AI and sports-betting](#)

[1.3.5 Hypothesis testing, Chapter 6: How AI tests user preferences](#)

[1.3.6 Variable relationships, Chapter 7: AI detects wildfires early](#)

[1.3.7 Data quality, Chapter 8: AI-powered autonomous cars and bad data](#)

[1.3.8 What's that AI?, Chapter 9: Digital Public Infrastructure \(DPI\)](#)

[1.3.9 *Responsible use of data and AI*, Chapter 10: AI for good](#)

[1.4 Data Tools and How-to](#)

[1.5 What's Next?](#)

[1.6 Read & Reflect](#)

[1.7 Exercises](#)

2 About Data

[2.1 Types of Data](#)

[2.1.1 Data types based on usage](#)

[2.1.2 Data types based on how obtained](#)

[2.2 Data Collection](#)

[2.2.1 Manual data collection methods](#)

[2.2.2 Automated data collection methods](#)

[2.3 Storing, Managing and Accessing Data](#)

[2.4 Conclusion](#)

[2.5 Read & Reflect](#)

[2.6 Exercises](#)

[2.7 Summative Project](#)

3 Exploratory Data Analysis

[3.1 Data Sets](#)

[3.1.1 Real-world example: Palmer penguins data](#)

[3.1.2 A simple example: Middle school student heights](#)

[3.2 Exploratory Data Analysis](#)

[3.3 Descriptive Statistics: Typical values](#)

[3.3.1 Robots and robust statistics](#)

[3.4 Descriptive Statistics: Data spread](#)

[3.4.1 Rank based statistics](#)

[3.4.2 Deviation based statistics](#)

- [3.5 Descriptive Statistics: Shape of the data spread](#)
- [3.6 Descriptive Statistics: Qualitative variables](#)
- [3.7 EDA of Palmer Penguins Data with Descriptive Statistics](#)
 - [3.7.1 Data quality issues](#)
- [3.8 Exploring Data through Visualization](#)
 - [3.8.1 Circle \(pie\) charts](#)
 - [3.8.2 Bar charts](#)
 - [3.8.3 Histograms](#)
 - [3.8.4 Boxplots](#)
 - [3.8.5 Dotplots and scatterplots](#)
 - [3.8.6 Time series](#)
 - [3.8.7 Heatmaps](#)
 - [3.8.8 Choice of graph](#)
- [3.9 EDA and Big Data](#)
- [3.10 EDA and AI](#)
- [3.11 Read & Reflect](#)
- [3.12 Exercises](#)
- [3.13 Summative Project](#)

[4 Sampling: Less Is More](#)

- [4.1 Why Sample?](#)
- [4.2 Illustrative Examples of Sampling Applications](#)
 - [4.2.1 Sampling in environmental sciences](#)
 - [4.2.2 AI and sampling](#)
 - [4.2.3 Monitoring quality control through sampling](#)
- [4.3 Types of Sampling](#)
 - [4.3.1 Hands-on sampling practice: Palmer penguins](#)
 - [4.3.2 To replace or not?](#)
- [4.4 Sampling Variability and Distributions](#)

- [4.4.1 Effect of sample size](#)
- [4.4.2 Statistical bias](#)
- [4.5 Bias and Ethics of Sampling](#)
- [4.6 Read & Reflect](#)
- [4.7 Exercises](#)
- [4.8 Summative Projects](#)

5 Probability in the Age of AI

- [5.1 Uncertainty, probability and AI](#)
- [5.2 Events and Outcomes](#)
 - [5.2.1 Probabilities of events](#)
- [5.3 Probability Distributions](#)
 - [5.3.1 Random variables](#)
 - [5.3.2 Probabilities of random variables](#)
 - [5.3.3 Joint probabilities](#)
 - [5.3.4 Marginal probabilities](#)
 - [5.3.5 Conditional probabilities](#)
 - [5.3.6 Bayes' rule](#)
 - [5.3.7 Independent events](#)
- [5.4 Common Probability Distributions](#)
 - [5.4.1 Empirical distributions: learning from data](#)
 - [5.4.2 Probability distributions and AI](#)
 - [5.4.3 Theoretical probability distributions](#)
 - [5.4.4 Illustrative example: Blue Ghost and the Binomial distribution](#)
 - [5.4.5 Binomial distribution](#)
 - [5.4.6 Illustrative example: Hurricanes and the Normal distribution](#)
 - [5.4.7 Normal probability distribution](#)

[5.5](#) [Margin of Error and Confidence Intervals](#)

[5.5.1](#) [More on confidence intervals](#)

[5.6](#) [Read & Reflect](#)

[5.7](#) [Exercises](#)

[5.8](#) [Summative Project:](#)

[5.9](#) [Appendix](#)

[5.9.1](#) [Sets of outcomes](#)

[5.9.2](#) [Conditional probability distribution of \$Y|X = CC\$](#)

[6](#) [Data-Driven Hypotheses Testing](#)

[6.1](#) [Automation and Hypothesis Testing](#)

[6.1.1](#) [Process of hypothesis testing: Key steps](#)

[6.1.2](#) [What is a \$p\$ -value?](#)

[6.2](#) [Tests Based on Type of Hypothesis](#)

[6.2.1](#) [Two-tailed test](#)

[6.2.2](#) [One-tailed test](#)

[6.2.3](#) [Illustrative Example: Testing adoption rates of puppies at rescue centers](#)

[6.3](#) [Tests Based on Sample Size](#)

[6.3.1](#) [Mars rover Perseverance: Small samples and Student's \$t\$ -test](#)

[6.3.2](#) [BirdCast AI: Large samples and the Z-score test](#)

[6.3.3](#) [A/B testing: Tests based on more than one sample](#)

[6.4](#) [Could We Have Made a Mistake? Type 1 and Type 2 errors](#)

[6.4.1](#) [Power of a test](#)

[6.4.2](#) [Relationship to Type 2 error](#)

[6.5](#) [Nonparametric Tests for Empirical Distributions](#)

[6.5.1](#) [StreetStyle, AI tool for fashion: Sign test for the median](#)

- [6.5.2 Illustrative example: Testing if Kathakali headpieces typically weigh 1.75 pounds](#)
- [6.6 Pitfalls of Hypothesis Testing: Data Dredging](#)
- [6.7 Read & Reflect](#)
- [6.8 Exercises](#)
- [6.9 Summative Project](#)

7 Variable Relationships: the Full Picture

- [7.1 Illustrative Example: Penguins, the more variables the merrier](#)
 - [7.1.1 Illustrative example: AI tracks climate change](#)
 - [7.1.2 Illustrative example: AI deployed in drones helps farmers](#)
- [7.2 Independence of Categorical Variables](#)
 - [7.2.1 AI, fake handbags and the Chi-squared \(\$\chi^2\$ \) test of independence](#)
 - [7.2.2 Step-by-step example of \$\chi^2\$ test of independence](#)
- [7.3 Analysis of Variance for Continuous and Categorical Variables](#)
 - [7.3.1 Hydroponic agritech AI and ANOVA](#)
- [7.4 ANOVA with Excel](#)
 - [7.4.1 One-way ANOVA with Excel: Palmer penguins](#)
 - [7.4.2 Two-way ANOVA with Excel: Hydroponic fruit](#)
- [7.5 Manual Computation of a Two-Way ANOVA](#)
 - [7.5.1 ANOVA: Expected values](#)
 - [7.5.2 ANOVA: Sum of squares of deviations](#)
 - [7.5.3 ANOVA: Sampling distribution](#)
 - [7.5.4 ANOVA: Conclusion](#)
- [7.6 Relationships Between Continuous Variables](#)
 - [7.6.1 Linear relationship: Correlation](#)
 - [7.6.2 Line of best fit: Ordinary Least Square regression](#)
 - [7.6.3 Relationship between correlation \(\$\rho\$ \) and slope \(\$\beta\$ \)](#)

- [7.6.4 Beware: Correlation is not causation](#)
 - [7.7 Time and Space Are Special](#)
 - [7.7.1 Time-dependent data](#)
 - [7.7.2 Time series](#)
 - [7.7.3 Spatial data](#)
 - [7.7.4 Spatio-temporal data](#)
 - [7.8 Read & Reflect](#)
 - [7.9 Exercises](#)
 - [7.10 Summative Project](#)
 - [7.10.1 Chi-squared test of independence: How does time spent on homework impact your grade?](#)
 - [7.10.2 One-Way ANOVA: Is there a significant difference between the mean bill lengths of different species of penguins?](#)

8 Data Quality and AI

- [8.1 Data Quality: Has your data gone bad?](#)
 - [8.1.1 Data pipeline and data quality](#)
- [8.2 Monsters in Your Data: Types of data glitches](#)
 - [8.2.1 Missing and mangled data](#)
 - [8.2.2 Incomplete data](#)
 - [8.2.3 Inconsistent data](#)
 - [8.2.4 Illustrative example: Orbiter lost due to inconsistent data](#)
 - [8.2.5 Anomalies](#)
 - [8.2.6 Duplicates and near-duplicates](#)
 - [8.2.7 Other data quality issues](#)
- [8.3 Measuring Data Quality](#)
- [8.4 Statistical Distortion and Data Cleaning](#)
- [8.5 Read & Reflect](#)

- [8.6 Exercises](#)
- [8.7 Summative Project](#)

[9 What's that AI?](#)

- [9.1 What's AI, anyway?](#)
 - [9.1.1 AI and what we have learned so far](#)
- [9.2 Illustrative Example: Build an AI classifier with probabilities](#)
- [9.3 Illustrative Example: Large Language Models and Probabilities](#)
- [9.4 Types of AI](#)
 - [9.4.1 Symbolic AI](#)
 - [9.4.2 Machine learning](#)
 - [9.4.3 Natural language processing](#)
 - [9.4.4 Robotics](#)
 - [9.4.5 Generative AI](#)
 - [9.4.6 Agentic AI](#)
 - [9.4.7 AGI](#)
 - [9.4.8 Web search](#)
- [9.5 AI: Power and Responsibility](#)
 - [9.5.1 AI and education](#)
 - [9.5.2 AI and employment](#)
 - [9.5.3 AI and architecture](#)
 - [9.5.4 AI and healthcare](#)
 - [9.5.5 Community and AI](#)
 - [9.5.6 AI is not always right](#)
- [9.6 Read & Reflect](#)
- [9.7 Exercises](#)
- [9.8 Summative Projects](#)

[10 Responsible Use of Data and AI](#)

- [10.1](#) [Types of bias](#)
 - [10.1.1](#) [Bias in training data](#)
 - [10.1.2](#) [Bias in algorithms due to feedback](#)
 - [10.1.3](#) [Bias due to measurement proxies](#)
- [10.2](#) [Misinformation and Fakery](#)
- [10.3](#) [Data Privacy and AI](#)
- [10.4](#) [Ethics in AI](#)
- [10.5](#) [What Is Being Done](#)
- [10.6](#) [Our Responsibility as AI Consumers](#)
- [10.7](#) [Conclusion](#)
- [10.8](#) [Read & Reflect](#)
- [10.9](#) [Exercises](#)
- [10.10](#) [Summative Project](#)

Bibliography.

Preface

Statistics and Data Foundations for AI grew from the discussions that the authors, long-time friends with shared interests, both educators and scientists, had about the level of preparation of college students for an education in Artificial Intelligence (AI) which will soon be a key requirement for the workforce. AI is increasingly integrated into productivity tools at the workplace. In daily life, GPS navigation, self-driving cars and AI chatbots that can plan itineraries and provide counseling are already a reality. Understanding the underpinnings of AI is therefore essential to use it effectively to its full potential, and to handle its peril and its promise.

For many, AI is an abstraction. However, a majority of the key concepts of AI have their foundations in classical statistics including probability theory and statistical inference. AI applications are trained on massive amounts of data, and in turn generate data in the form of text, images, analyses and predictive models. Therefore, a deep understanding and facility with data is important as well.

The main purpose of the book is to act as a bridge between classical statistics, data science and emerging AI technologies. Every key concept is explained in the context of AI applications and real-world data examples from authoritative sources such as NASA (National Aeronautics and Space Administration) and NOAA (National Oceanic and Atmospheric Administration). In addition, given the primacy of data for AI, the authors

take the unique approach of devoting an entire chapter to the technical and quantitative aspects of data quality. While the pursuit of AI has become a modern-day gold rush, it is also important for students to be aware of the ethical concerns, biases and potential catastrophic errors that accompany such a rapid development.

To make AI an immersive experience, the book features the use of AI as a learning tool in the form of prompts to AI-agents (Google's Gemini, and OpenAI's ChatGPT in one instance) to delve deeper into any topic for further study and research. The authors draw on real-world AI applications from many disciplines including sports, fashion, and the arts and make the content accessible and relatable to students from different academic backgrounds. The open-ended nature of many discussions and exercises allows the book to be taught and read at multiple levels of depth and rigor. In addition, the book provides a strong grounding for advanced courses in AI and machine learning.

The book is suitable for undergraduate and graduate students, and practitioners from both STEM and non-STEM backgrounds. It is comprehensive and rigorous with a strong focus on foundational concepts.

With the emergence of AI, basic data literacy is no longer sufficient. The authors believe that in order to thrive in an AI-powered world, it is essential to develop critical 21st Century skills. This book provides those skills.

Tamraparni Dasu & Geetha Murthy

QR Code to Access the URLs in the Textbook

Statistics and Data Foundations for AI includes almost 300 URLs. Some are associated with datasets, some with the output of GeminiBox prompts, and some point to resources such as articles for further study and research. We have provided a QR code to facilitate access to the URL list to users of the print version of the textbook. The URL list is organized in order of appearance in the textbook, annotated with the chapter and section numbers. The URLs are denoted by a blue  symbol.



https://docs.google.com/document/u/0/d/1LdwqUpZmhLkxnoaRbL4IH_wUJ9nkUxHGgMoKcFHIBZ0/mobilebasic?pli=1

QR code to access URLs in Statistics and Data Foundations for AI.

Authors

Dr. Tamraparni Dasu (Ph.D. Mathematical Statistics, University of Rochester, 1991) is a research scientist and Data Science expert specializing in computational statistics, machine learning and data quality. She retired as Lead Inventive Scientist after 31 years at AT & T Bell Laboratories and now teaches Data Mining, Machine Learning and AI as an adjunct professor at Fairleigh Dickinson University, New Jersey. Dr. Dasu has published extensively in top-tier journals and research conferences such as SIGMOD, KDD and VLDB and authored the field's first technical book on data quality, *Exploratory Data Mining and Data Cleaning*, John Wiley (2003), with Dr. Theodore Johnson. As an educator, Dr. Dasu is committed to mentoring the next generation of quantitative thinkers, computer scientists and data scientists.

Dr. Geetha Murthy (Ed.D. Instructional Leadership, St. John's University, Queens, New York, 2015) is a highly experienced educator and administrator whose research focused on describing and dismantling self-limiting beliefs in students that inhibit them from pursuing STEM education/careers. Most recently, Dr. Murthy served as the K-12 director of mathematics for Herricks School District in Long Island, NY, a high-performing public school district. Under her leadership, the district made significant progress in increasing student achievement through initiatives that focused on systemic changes to create and sustain greater equity, access

and success for all students. Dr. Murthy is passionate about promoting 21st Century Skills in teaching and learning.

The authors believe that statistics and data foundations are essential in an increasingly AI-native world and should be accessible to students and practitioners from varied backgrounds. The authors have a unique combination of complementary skills, a deep understanding of the core set of fundamental concepts and practice of statistics and data science, and extensive experience and insights into foundational education, that makes this a canonical book for the study of statistics and data in relation to AI.

Statistics, Data and AI

AI has become a ubiquitous part of every aspect of our lives. We use AI when we dictate text messages to our smartphone, when we ask our navigation system for directions, and when we use AI tools like Gemini, ChatGPT or Co-pilot to plan travel, write summaries and draft emails. We rely increasingly on AI assistants such as chatbots to find information, to reach customer service, to get counseling and even health advice.

Data is used to train, manage, monitor and deploy AI, and AI in turn generates data in the form of content as well as logs. Learning how to manage and understand data, to use it responsibly, to develop an intuition for its usability and limitations, are essential for living in an AI world. These data skills together with key statistical concepts constitute the first step to learning AI. The objective of this book is to deconstruct complex, foundational concepts of statistics and data without sacrificing rigor, and to place them in the context of real-world AI applications in order to make them widely accessible.



FIGURE 1.1: Left: Australian marine scientists use drones to collect data and monitor coral reefs. (Image: Australian Institute of Marine Science.) Right: Sports analytics use AI and real-time data collected through sensors to scout, coach, match up and improve the performance of athletes. (Image: Freepik.)

1.1 Why Statistics and Data Foundations?

Many aspects of our life depend on AI, and going forward AI will be an inextricable part of our daily life, like cars or electricity. In addition to our personal use, industries and institutions too rely on AI. Social media companies use AI to tailor recommendations about movies, food, clothes and restaurants based on our browsing history. AI is used by scientists to launch rockets, design self-driving cars, discover new drugs, diagnose tumors and to monitor hurricanes. Sports, fashion, marketing and agriculture industries use AI-powered tools extensively to manage their businesses, inventory, and supply chains. Even museums and libraries use AI to organize and preserve artworks, artifacts and books.

AI is powered by data. Self-driving cars are trained on thousands of images of other vehicles, pedestrians, traffic rules, speed limits and weather

conditions. Sports clubs use data extensively to research strategies, evaluate athletes and engage viewers, e.g., through sports betting. AI applications use image recognition to transcribe and understand documents from hundreds of years ago. Such documents are hard to decipher because of deterioration in quality, obscure text and notation, and damage to documents, as explained in the article *Using AI to decipher historical manuscripts*, in [Section 1.6](#).

AI itself generates data. Generative AI can write novels, make videos and paint pictures. It can have philosophical conversations, write code and answer any question posed to it. In addition, every action performed by AI leaves a digital trace in the form of *system logs* that can be used to analyze, audit and improve AI.

The need for strong statistical foundations and data intuition in order to understand emerging AI technologies has been emphasized by experts and leading organizations, e.g., the *American Statistical Association (ASA)*, referenced in [Section 1.6](#). This book aims to fulfill that need. [Figure 1.2](#) summarizes the main goals of the book and special features that enable achieving that objective.

- *Bridge between Statistics and AI applications*: The book connects classical statistical concepts to AI by demonstrating through real-world examples how such concepts are used by AI applications. By drawing from a wide spectrum of applications in sports, fashion, climate science, environmental science, health, medicine, and space exploration, to name a few, AI is made relatable to everyday life and is no longer an abstraction. The examples used in this book cite authoritative sources such as (National Oceanic and Atmospheric Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), Centers for Disease Control and Prevention (CDC), Audubon Society, and the Pew

Research Center. The content is accessible to learners from a wide variety of backgrounds (STEM and non-STEM) without sacrificing rigor.

- *Precursor to AI and Machine Learning courses:* The book provides a comprehensive discussion of fundamental concepts of statistics and data science such as data management, exploratory data analysis, sampling, probability theory, hypothesis testing, and multivariate analysis. The book can be used at different levels, including as a rigorous statistical course that serves as a prequel to AI and machine learning courses.
- *Foundation for strong data intuition with emphasis on data quality:* According to AI experts, data is the new frontier of AI. Algorithms and computing infrastructure that support AI, both hardware and software, have advanced at a rapid rate, but the data on which AI is trained has lagged behind. A strong focus on the quality of data, as well as an intuition for data in the context of a specific problem, is critical to using AI effectively. In this book, we devote an entire chapter to data quality while emphasizing its importance throughout the text.
- *Guide to ethical and responsible use of data and AI:* AI has been known to provide incorrect or imaginary responses called *hallucinations*, as well as perpetuate and amplify errors and biases embedded in its training data. AI has sped up the creation and transmission of misinformation, fake videos and polarizing content with serious real-life consequences. This book emphasizes the need for careful evaluation of AI output and content, and screening it for intentional and unintentional bias through AI audits. Preserving the privacy of individuals by securing their data is another important aspect of the ethical and responsible use of data and AI.

About *Statistics and Data Foundations for AI*

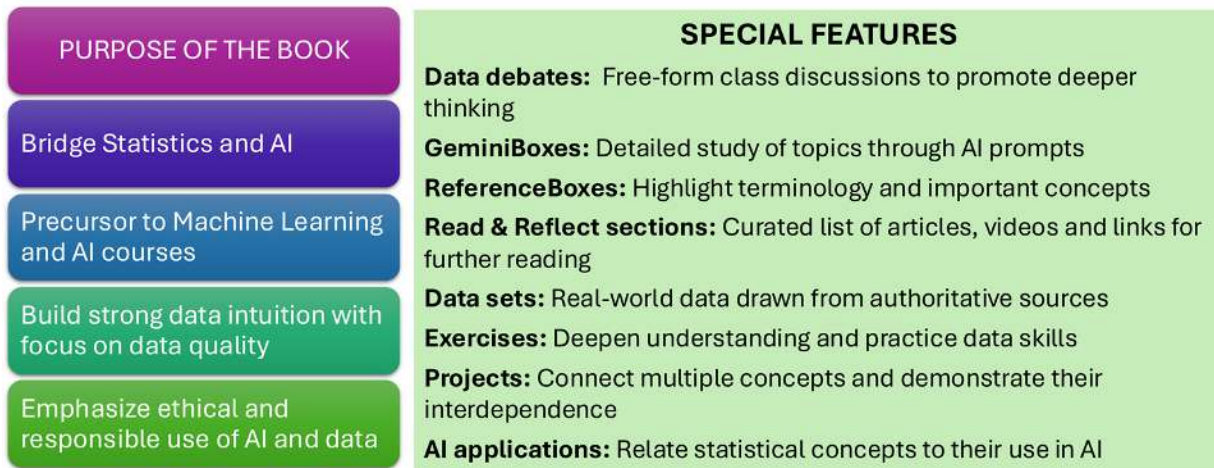


FIGURE 1.2: The purpose of the book *Statistics and Data Foundations for AI*, and special features that enable fulfilling the goals. [↩](#)

1.2 Special Features of the Book

We have organized the content of the book in such a manner as to reiterate concepts already learned, and to connect them to real-world applications, particularly related to AI. The book features special pedagogical elements that enable delving deeper into the material in an interactive way, to build upon and extend understanding of concepts, and to demonstrate how they are connected to material discussed in earlier chapters. These features are:

1. *Data Debate:* Statistical and data concepts often require deeper thinking to fully grasp their meanings and applications. Data debates are an opportunity for students to actively engage in classroom discussion and share their thinking and potentially conflicting points of view in a non-

judgmental environment, facilitated by the instructor. There is no one correct answer or solution. [Data Debate 1.3.1](#) is an example.

2. *GeminiBox*: The motto “Learning by Doing” inspired our GeminiBox which teaches how to write *prompts* for AI tools. Gemini is an AI assistant developed by Google. A GeminiBox demonstrates how to gather information related to specific questions by using an appropriate prompt. We present a prompt as well as a link to its output, typically to foster curiosity and extend a concept beyond what is taught and discussed in the classroom. Using the prompts in the GeminiBoxes and examining the output is mandatory.

GeminiBoxes are an opportunity to teach students to think analytically, to question and validate AI and its output, e.g., to identify *hallucinations*, AI's tendency to fabricate factually incorrect answers. For example, in order to learn more about data literacy and historical manuscripts, the following GeminiBox can be used. The link points to the output of the prompt. Note that the same prompt might return different outputs each time. Students must review the output for accuracy. We encourage students to use the GeminiBox prompts as they progress through the book in order to get the most up-to-date output because the information available on the internet is constantly refreshed.

GeminiBox 1.2.1: Data literacy and deciphering archives

“How is data literacy used in deciphering historic documents?” 

3. *ReferenceBox*: ReferenceBoxes contains *quick-reference material* such as terminology, definitions, and scientific and statistical vocabulary that is essential to master the concepts in a chapter. Students should become

familiar with, and use, these terms as they answer questions or participate in discussions in class. ReferenceBoxes are positioned close to related content, and they make it easy to refer to important technical terms.

4. *Read & Reflect*: AI is a fast-moving field, and there is a vast amount of information and knowledge available to educate the public. It is difficult to navigate this ocean of material of varying quality and provenance, written at different levels of technical detail, and intended for a wide range of readers from a lay audience to experts. We have vetted and curated *high quality articles* such as research papers and videos related to concepts presented in each chapter, and we listed them in a special *Read & Reflect* section at the end of each chapter. Instructors may use these for assignments.
5. *Data sets*: To reiterate the “Learning by Doing” theme, we use real-world data in most examples and exercises, *curated from authoritative sources* such as NOAA and from publicly available data resources such as Kaggle. Using real-life data serves two purposes: first, to demonstrate how data and statistical principles and concepts are related to concrete real-life situations; and second, to practice such concepts and skills on real data, to explore its features and flaws, to choose appropriate analytics, and to interpret the results.
6. *Exercises*: At the end of each chapter, we provide a set of exercises created to deepen conceptual understanding and practice data and statistical analysis skills. The focus is to interpret results and gain insights on how fundamental statistical concepts are applied in different fields such as science, business, medicine, and education. Exercises are also an opportunity to learn data skills, and to interpret and

communicate results in the context of a problem discussed in the exercise.

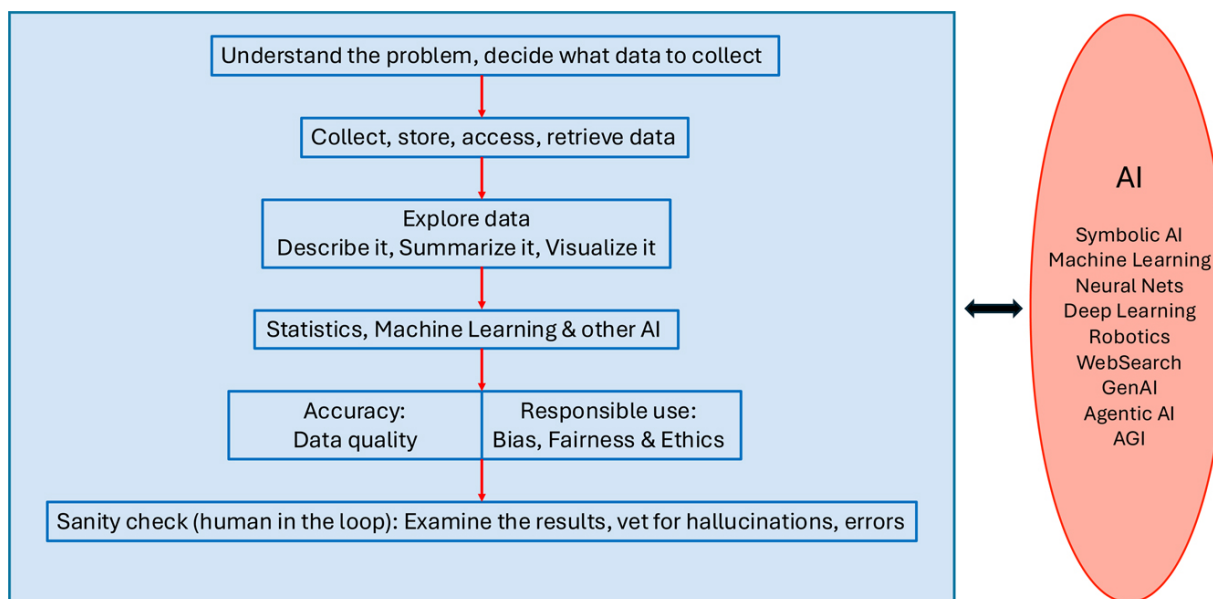
7. *Summative projects*: Projects are summative and bring together numerous concepts discussed in the chapter, and frequently concepts from prior chapters. Projects have multiple parts, and students are required to present a comprehensive response to each part. By working through the summative projects, students will gain further understanding of how different aspects of data and statistical analysis fit together, and build data intuition. Summative projects may be used for student evaluation by the instructor.

1.3 Statistics, Data and AI

AI originated many decades ago, as far as back in 1950. However, the availability of vast amounts of data and the explosion in computing power led to the widespread development and deployment of AI that we see now. However, the foundational concepts of AI are drawn from classical statistics, which came to be known as *Data Science* in recent years when combined with basic data processing and programming skills. A simple introduction to basic statistical concepts, especially for non-STEM majors, can be found in *Statistics*, by Freedman, Pisani and Purves [[17](#)].

[Figure 1.3](#) shows the connection between classical statistics, data and AI. AI builds on fundamental concepts from statistics and uses data for training, while statistical models gather data from AI through feedback loops to improve performance and advance foundational concepts. Since examples are the best way to establish this connection, we start the book with an introduction to real-world examples of topics discussed in each chapter and

illustrate their connection to AI. We will revisit this figure in [Chapter 9](#) and discuss how the statistical concepts discussed in the book are used in AI.



[Extended Description](#)

FIGURE 1.3: Statistics, data and AI are interconnected. AI builds on fundamental concepts from statistics and uses data for training, while statistical models gather data in the form of feedback loops from AI to improve performance and advance foundational concepts. [↗](#)

The following sub-sections present a brief overview of the topics in each chapter of the book along with an illustrative example.

1.3.1 *About Data*, [Chapter 2](#): Monitoring weather data with AI

Data is the fuel AI runs on: it is trained on data, learns from data, and generates data. Data collection, storage, retrieval, and management are fundamental steps in data science and AI, as shown in [Figure 1.3](#). AI is used extensively to collect and manage data for other AI applications to use, such

as continuous data collection through sensors on spacecraft. [Chapter 2](#) discusses different types of data, their storage and retrieval, and how data is used by AI.

As an example, NOAA deploys satellites as a part of the Joint Polar Satellite System (JPSS), [Figure 1.4](#), left, to orbit the earth multiple times a day collecting data for AI applications that monitor weather and climate. The project is described in *Fact Sheet about the Joint Polar Satellite System to Monitor Weather and Climate* in the [Section 1.6](#). Such satellites, in addition to collecting external data, also monitor their own movements, internal environment, and functionality of key systems through *telemetry* data from various sensors and performance metrics. Data collection and management on such a large scale requires AI-powered tools.

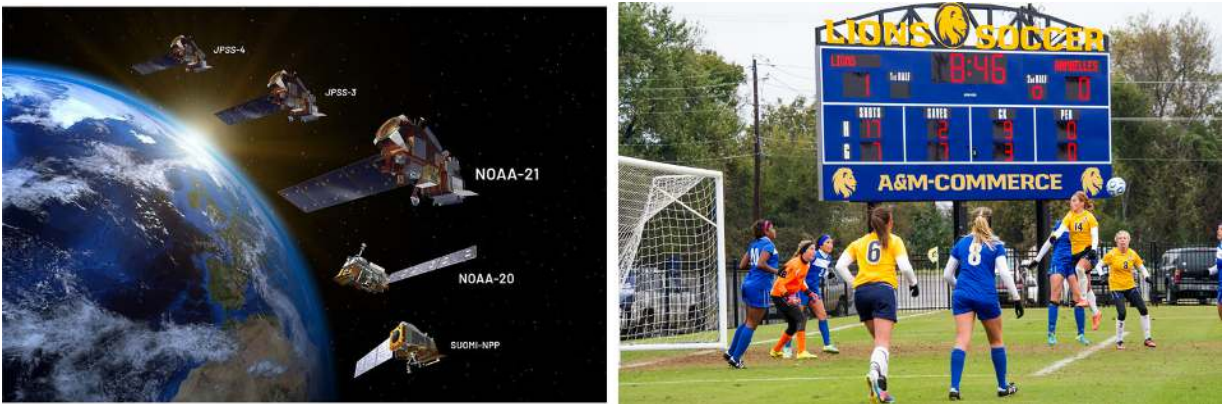


FIGURE 1.4: Left: NOAA's satellite system, JPSS, circles the earth several times a day collecting data. AI applications analyze and manage the data, as well as use it for predictive purposes such as hurricane warnings. Right: AI enables real-time exploratory data analysis (EDA) of vast amount of sports data changing the way sports are played and athletes are measured, as well as driving massive user engagement in the form of sports betting and fantasy sports. [↗](#)

1.3.2 *Exploratory Data Analysis*, [Chapter 3](#): How soccer clubs use AI

Exploring data helps in understanding the objects about which the data are collected, as well as the strengths and weaknesses of the data, e.g., satellite trajectories. Is the data relevant? Is it clean? Is it complete? Is it timely? How should it be prepared so that AI applications can ingest it and process it seamlessly? Such a fundamental analysis of data outside the context of a particular task such as prediction, is called *exploratory data analysis (EDA)*.

AI has powered the use of real-time EDA in many fields, from marketing to consumer recommendations. Fantasy sports and sports betting, have become juggernauts of user engagement through real-time EDA of sports data, even as games are in progress. EDA is also used to gain a deep understanding of the sport, [Figure 1.4](#), right, analyzing opponent strategies, player capabilities, coaching tactics, and user engagement. A *Forbes* article, *Unleashing The Power Of AI To Create A New Era In Soccer Excellence*, listed in [Section 1.6](#) offers further reading on this fascinating topic.

Data Debate 1.3.1 [↩](#) *How would you use EDA to help your college basketball team?*

1.3.3 *Sampling*, [Chapter 4](#): AI speeds up financial trading

Sampling is the process of using a subset to make inferences about a larger population, e.g., learning about migration patterns of whales by tagging or observing a select group of them. [Chapter 4](#) discusses different types of sampling and the scenarios in which they are used. Sampling is used extensively to test soil, air and water for contaminants, and to study the spread of disease.

Sampling plays a significant role in AI. For example, samples of human speech are used to train and generate synthetic voices for AI chatbots. In general, AI algorithms are more efficient when trained on properly chosen data samples. It is often not possible to use all the available data while making real-time decisions, such as predicting the behavior of complex financial instruments [Figure 1.5](#), left, or guiding a self-driving car. The massive amounts of data available in these situations might take too long to process and analyze, resulting in latency in reaction time which could lead to huge financial losses, or worse, damage to lives and property.



FIGURE 1.5: Left: Sampling is used by AI tools to make real-time decisions in the finance industry (Source: Freepik). Right: Sports betting uses AI to compute hundreds of thousands of real-time probabilities. Photo credit: Baishampayan Ghose. [↗](#)

1.3.4 *Probability*, [Chapter 5](#): AI and sports-betting

Probability theory is used to evaluate the chances of a particular event happening. A coin toss is a simple example, where each of the two outcomes, heads or tails, has an equal chance of occurring. Events can also be highly complex, like the probability of a particular film breaking box office records. Numerous factors, such as the cast of the movie, the director,

cinematography, word-of-mouth, marketing and ad campaigns, and possibly even weather conditions, figure in the probability that a movie will be a blockbuster.

Probability theory is critical to AI. For example, AI algorithms use *probabilistic reasoning* to guide movements of robots and autonomous vehicles, e.g., by detecting moving objects and their most probable trajectories in order to avoid collisions. AI is also used in sports betting, [Figure 1.5](#), right, to provide continuously updated odds for different types of events: number of runs scored by a baseball team, the order of finish at the Kentucky Derby, and the likelihood of Michigan Wolverines winning the National Championship in college football.

1.3.5 Hypothesis testing, [Chapter 6](#): How AI tests user preferences

Suppose we want to find the answer to a question, e.g., “does a drug help relieve the symptoms of a disease”, in an objective, data-driven manner rather than based on opinions or convenience. Hypothesis testing provides a rigorous statistical framework for using data from an appropriately chosen sample of individuals to make inferences about the drug's effects on the population in general.

AI has accelerated the process of hypothesis testing. It can generate hundreds of hypotheses, collect enormous amounts of data, and through repetition, output reliable results. Social media companies do this continuously to test user engagement and interest. Should the “like” icon be dark blue or light blue? Should it be a thumbs-up icon or a heart? They use hypothesis testing to make constant changes to their websites and the products and services they recommend, based on your preferences and those of others like you.

Data Debate 1.3.2 Give an example of a hypothesis about campus security that you would like to test.



FIGURE 1.6: Left: A scientist in the process of testing a sample in a lab. (Source: U.S. Navy. Naval Medical Research Center.) Right: Satellites use AI based on geo-spatial multivariate data to detect and contain wildfires. Photo credit: Noah Berger.

1.3.6 Variable relationships, [Chapter 7](#): AI detects wildfires early

Most real-world phenomena are complex, with multiple factors interacting with each other to influence their outcome. For example, to predict a hurricane's path, it is important to collect data about multiple variables such as wind speed, wind direction, temperature shifts, atmospheric pressure, humidity, topological features like mountains, and many others. Collecting all these measurements increases the amount of data considerably, and makes the analyses more computationally challenging. However, in general, the accuracy of predictions and forecasts improves when multiple variables and their interactions are included in the study.

AI is ideal for studying multiple aspects since it excels at dealing with vast amounts of data, and making complex calculations. Google research and the Earth Fire Alliance use an AI tool, *FireSat*, for early detection of wildfires. Current satellite tools cannot detect fires that are small because they use low-resolution images that are updated only a few times a day. FireSat can detect smaller fires sooner, by comparing current images with previous images. In addition, FireSat factors in multiple variables about the weather, environment and topology to detect and forecast the propagation of fire. This provides firefighters and other first responders early warning capabilities and an opportunity to take preventive measures to stop the spread of fires, and save lives and property. FireSat is an example of how AI uses geospatial analysis and multiple variables for forecasting wildfires. The article *FireSat Detects Wildfires Sooner and Alerts First Responders* in [Section 1.6](#) provides further details.

1.3.7 Data quality, [Chapter 8](#): AI-powered autonomous cars and bad data

Incorrect data has serious consequences. What happens if a business ships out products to an incorrect address? The customer will be annoyed because they did not receive the product on time and will want their money back. The person who received the item by mistake may decide to keep it as a freebie. The company will lose revenues, customers, and time trying to track down the incorrect address.

Data is AI's oxygen, without high-quality data AI will be ineffective and possibly harmful. [Figure 1.7](#) shows a vehicle's GPS tracks that seem to veer off into the wilderness for a short stretch, due to bad quality GPS data, discussed in the research paper *Detection of outliers in crowdsourced gps traces* [[23](#)]. This anomaly is hopefully a harmless artifact of bad data, but

what if navigation systems use outdated maps or routes built on bad GPS data? Motorists might end up in the middle of the woods or on newly constructed cul-de-sacs. What if self-driving cars are trained on incomplete data and only recognize people as obstacles, not animals or two-wheeled vehicles? They might plow into dogs, bicycles and motorcycles.



FIGURE 1.7: Left: Examples of weird GPS traces caused by bad data. From *Detection of outliers in crowdsourced GPS traces*. Right: A self-driving car. What if it were trained on outdated maps? [↗](#)

While AI has made considerable progress in analytical algorithms and computational power, it lags in the use of high-quality data. Many AI hallucinations and biases can be traced back to poor-quality data. Recently, there has been a strong push to improve the quality of data used by AI applications, as discussed in the article *Why It's Time for Data-Centric AI*, in [Section 1.6](#). [Chapter 8](#) is unique to this book. It provides a rigorous and quantitative introduction to data quality, including discussion of types of issues, cleaning strategies, quantifying data quality, and the potential consequences of excessive remediation that lead to *statistical distortion* of the data.

1.3.8 *What's that AI?*, [Chapter 9](#): Digital Public Infrastructure (DPI)

While examples of the connection between statistical concepts and AI are discussed throughout the book, [Chapter 9](#) focuses specifically on AI, with a brief overview of AI and how it is used, from symbolic AI to machine learning to recommender systems and natural language processing.

AI is now an important part of daily life, as well as an essential tool used by countries, corporations and institutions to increase efficiency. Developing nations are leaning into AI to leapfrog over intermediate stages of development. They are focused on investment in digital public infrastructure (DPI) to help the public navigate life in a frictionless manner, whether it is interacting with the government or running a street-side food stall. For example, in India there are 22 official languages with distinct scripts and more than 1600 dialects, which make communication and conducting business across the country difficult. The Indian government has used natural language processing, a form of AI, to provide a digital platform, [Bhashini](#), to translate between any two languages or dialects to help its 1.4 billion residents transcend language barriers.

[Figure 1.8](#), left, shows a large multilingual crowd at the Kumbh Mela in Prayagraj, 2019, which attracted around 200 million people.



FIGURE 1.8: AI is being used to enable India's 1.4 billion people (who speak 22 languages and 1600 dialects) to communicate with each other. Photo credit: Michael T Balonek. Right: Bias in mobile games like Pokémon GO. [Pokemon Go](#), [Virginia State Parks](#). ↵

Data Debate 1.3.3 *What is an important problem that you'd like AI to solve?*

1.3.9 *Responsible use of data and AI*, [Chapter 10](#): AI for good

Bias and prejudice seep into technology and data, both intentionally and unintentionally. There have been instances in which machine learning algorithms and AI bots amplify the bias embedded in their training data, or they are manipulated into such behavior by users. An example of unconscious bias is the augmented reality game *Pokémon GO* where the locales for the game elements were crowdsourced, shown in [Figure 1.8](#), right. The contributors were predominantly gamers who played *Ingress*, mainly from affluent suburban neighborhoods. This resulted in the game

elements being inadvertently concentrated in similar neighborhoods, making them inaccessible to minorities from urban areas.

[Chapter 10](#) focuses on ethical challenges in the use of data and AI, including the issues of *data privacy*, *data security* and the spread of misinformation and deepfakes. The chapter concludes with an overview of extant guidelines, policies and measures adopted by countries, international organizations, and corporations to support ethical practice of AI, and the role of individuals to ensure that the power of data and AI is used for the benefit of humanity.

1.4 Data Tools and How-to

Statistical analyses often require applying formulas and performing lengthy mathematical calculations. Stepping through the calculations manually can reinforce the understanding of concepts and formulas. However, given the size of many data sets, manual computations may not be feasible, and it becomes necessary to use advanced calculators and software tools. Popular ones include:

- Graphing calculators, such as the *TI-84 Plus*, are useful in classroom settings for relatively small data sets, and for select statistical functions.
- Web-based calculators are available online, sometimes for free.
- Sophisticated analyses with larger data sets require tools such as *Google Sheets*, a spread-sheet application that has built-in data analysis capabilities that are easy to use, and it features an extension called *XLMiner Analysis Toolpak* that can be downloaded and used to perform

data analysis tasks such as sampling, hypothesis testing, linear regression, and exploratory data analysis.

- Although *Microsoft Excel* is only available with a subscription, many organizations provide free access to this tool. Excel too has many built-in functions and formulas for data analysis and data visualization. In most instances, step-by-step instructions on how to perform a particular analysis are available in the Resources section  of the book.
- Programming languages like *Python* and *R* are not required for this course. However, the Resources section  of this book includes sample code, specifically for figures in the textbook created using *R* code. *R* can be downloaded for free at *R: The R Project for Statistical Computing* [33] . *R for Data Science* by Wickham *et al.* provides a good introduction to programming in *R* [44]. Software tools such as *SPSS* and *Minitab* can also be used for statistical analysis at the discretion of the instructor.

Regardless of how the computations are done, students must interpret the results in the context of the problem, vet them for inconsistencies, and communicate them effectively to others.

1.5 What's Next?

AI is a significant part of our daily life. It is embedded in the GPS navigation of cars, in voice assistants that find recipes, and in chatbots that provide customer service and health counseling. While we use many sophisticated technologies without knowing their inner working, AI is different for the following reasons:

- AI has the power to influence opinions and change behavior of individuals, communities and countries in an unprecedented manner. The speed with which it enables the propagation of information in an easily consumable manner, such as social media reels, real and fake, has already changed the course of history in many ways.
- AI can accelerate creation of scientific innovation and technologies and improve the quality of life of humanity on a global scale: provide state-of-the-art education to children in remote and underserved areas, enable business and commerce at grassroots level, discover drugs to treat cancer, create artificial organs including eyes and heart, sustain the world on natural energy, ..., anything is possible with advances in AI.
- AI has the *potential* to reason and act autonomously. Just like humans, it can do great good, or potentially cause existential-threat-level harm.

For these reasons, it is the responsibility of the public to be familiar with the applications of AI and the foundations on which it is built, namely statistics and data. This book provides that education.

1.6 Read & Reflect

- *American Statistical Association: Role of Statistics and data science in the development of AI.* 
- *On the importance of statistical literacy in today's world of data, AI, and rapid information flow.* (Video Interview: Dr. Steve MacFeely, Chief Statistician at the OECD, former Director of Data and Analytics at the WHO.) 

- *Using AI to decipher historical manuscripts.* 
 - *Fact sheet about the Joint Polar Satellite System to monitor weather and climate.* 
 - *Unleashing the Power of AI to Create a New Era in Soccer Excellence.* 
 - *FireSat detects wildfires sooner and alerts first responders.* 
 - *Why it's time for data-centric AI.* 
-

1.7 Exercises

1. What are your career goals? Choose a career path(s) that you may be interested in and research how data and AI have made an impact on it, and what data skills you may need to succeed in this career.
2. What kinds of data are being collected through your smartphone and social media accounts? What are some data security issues that you may be concerned about?
3. How can AI help students and teachers? In your opinion, should AI be used for teaching and learning? What are the benefits? What are some challenges?
4. What are the environmental challenges with the widespread use of AI? Find articles, blogs or videos from reputable sources to research and write a summary. If you decide to use AI tools, make sure to check the sources and validate the summary and make sure you agree with it!

5. How do you think this course in Statistics and Data Foundations will help you in your studies and/or career? Do you think it is important to understand how AI was developed?
6. The World Economic Forum has published a report, *The future of jobs*, on how AI technologies impact jobs. Study this infographic  and use the interactive web tool *Data Explorer* provided at this link  to investigate trends related to skills and jobs in different sectors and occupations.

2

About Data

Data is a description of an object or process in the real world. It is essential to have good intuition for data in that specific context in order to provide an effective solution. The objective of this chapter is to cultivate such *data intuition*. Modern AI applications handle a great variety of data types such as simple tables of numbers, text, images, audio, video, webpages with multiple nested links, complex graphs of connected objects, and continuous streams of readings from sensors. Each data type requires a different treatment and understanding. Data intuition develops that skill.

We rely on data to make decisions and predictions. For example, while standing in line for ice cream as shown in [Figure 2.1](#), it is natural to count the number of people ahead to predict the wait time, or to turn on notifications and gather data about online deals to get the best possible price on a pair of jeans.



FIGURE 2.1: Data: The number of people ahead of you in an ice cream line to predict wait time. Photo credit: “Ice cream line” by Roman Königshofer.



In general, collecting data is essential for answering questions in an objective manner. Statistics, data science and AI rely on data to perform complex analytical tasks and predictions, e.g., compute distances between galaxies, predict the affect of changes in atmospheric temperature on ocean currents, and assess the probability of occurrence of natural disasters like wildfires. However, data is not just a matrix of rows and columns to be fed in to a software tool. Data is a description of an object or process in the real world and exists in that context. It is essential to have good intuition for data in the context of that specific problem in order to provide an effective solution.

ReferenceBox 2.0.1: Data intuition

Data intuition is the ability to understand data in the context of a real-world problem, assess it for relevance and quality, choose an appropriate analytical strategy for solving the problem, and interpret the results in that specific context to provide an effective solution.

Data consists of information, gathered through observation, measurement, or by asking questions as in surveys. For example, the number of people waiting in line for ice cream is an instance of data collected through observation. In order to find the most useful and relevant answers in the quickest, most economical way, it is important to understand what data to collect, how to collect, how much to collect, and how to organize it. [Section 2.1](#) discusses different types of data.

For the purpose of this book, we focus on *structured data* that consists of well-formatted numbers and characters organized in rows and columns. This is a popular and convenient type of data that is easily accessible by most types of analyses and software tools. However, there are a wide variety of data which require specialized tools and databases. For example, corpora of free-form text, images, emails, webpages, audio and video, connected graphs such as friend networks, DNA sequences and many others. Such data are called *unstructured data*. AI applications make extensive use of unstructured data.

In addition, the objective of this book is to study data as it relates to statistics and AI. Therefore, operational aspects of data management, such as data warehousing, data integration, federation, and optimizing data retrieval, fall outside the purview of this book. However, data quality and the responsible use of data *do* fall within the purview of statistics and AI, to

ensure the accuracy and reliability of analytical results. These topics are covered in detail in [Chapters 8](#) and [10](#).

ReferenceBox 2.0.2: Data and AI

- *Data* is a collection of information gathered by observation, through questions, or by taking measurements either manually or through automated means like sensors and computers.
- *Data literacy* is the ability to understand, manage, analyze and interpret data accurately and responsibly.
- *Data Science* is a cross-disciplinary field that uses methods from statistics, computer science, software engineering and programming for data-driven problem-solving, and pervades many domains of application like climatology, science, and finance.
- *AI* is a technology that uses statistical concepts, data, algorithms, and computation to mimic human intelligence. Examples include self-driving cars, voice assistants, and anomaly detection AI that can identify abnormal cells in medical imaging scans.

Data Debate 2.0.1 *Is the use of data always helpful? Can data be harmful? What should one do to make data a force for good and reduce its potential for harm?*

2.1 Types of Data

In statistics, the type of data determines the type of computation and analysis. For example, the average of a numeric quantity like number of people in an ice cream line is meaningful, but the average of descriptive characteristics like ice cream flavors is not. Data types play an important role in AI since they allow humans to specify important properties of data that help computers to process it efficiently.

There are numerous ways of categorizing data depending on the field of study. *Introduction to Data Mining* [40] has a comprehensive discussion of data categorization. In this book, we describe two ways of determining data types: based on usage (shown in [Table 2.1](#)), and based on how data is obtained.

TABLE 2.1: Examples of different types of data according to how they are used: as identifiers, for comparison and ranking, for counting and for measuring on the real number scale. 

Qualitative Data		Quantitative Data	
Nominal (Identifying)	Ordinal (Comparison)	Discrete (Countable)	Continuous (Every value)
Names of students	Scale of 1-5 stars	No. of people in line	Price
Ice cream flavors	Pass-Fail	No. of soccer goals	Distance
Flight numbers	Early-OnTime-Delayed	No. of TikTok likes	Volume

2.1.1 Data types based on usage

The following data types are based on the way data is used.

ReferenceBox 2.1.1: Data types based on usage

- *Qualitative data* is descriptive information, such as names, labels, tags. It is further categorized as: (i) *Nominal data* used as an identifier, e.g., names, zip codes, movie genres, and flight numbers. (ii) *Ordinal data* used for comparisons, ordering and ranking data. For example, “mild”, “medium”, “hot” and “on fire” are ordinal data used to describe the spiciness of hot sauces in increasing order.
- *Quantitative data* consists of numbers such as counts and measurements that can be used in arithmetic operations, e.g., addition and division. It is further subdivided into two types: (i) *Discrete data* that can only take certain values, and it is countable. (ii) *Continuous data* that can take any value on the real number line, and it is uncountable.

Qualitative data is used to describe the qualities of people, objects, and ideas, and is fundamental to our ability to communicate and understand the world around us. Ice cream flavors, colors of flowers, names of students, satisfaction index (on a scale of 1-5, for example) are all qualitative data. Qualitative data can be used to put things into categories, e.g., “toys”, “cars”, or “pillows”. For this reason, qualitative data is also called *categorical data*.

Qualitative data is of two types: *Nominal data* which is purely for identification, even if it is a number, e.g., an athlete's jersey number. *Ordinal data* has a little more structure, and allows ranking, ordering and classification. For example, a 1-5 customer survey rating index implies that

4 stars are better than 2, but not by how much. The Dewey Decimal Classification (DDC) system used by libraries is a way of classifying books by subject (and other sub-categories) to make them easier to locate.

Quantitative data is numeric and often the result of measurements or counting, e.g., price of coffee, weight of a person, and the minutes until the end of class. Quantitative data is critical for important endeavors, e.g., launching rockets, for monitoring hurricanes, and for tracking the popularity of politicians and entertainers.

Quantitative data is of two types: *Discrete data* is countable. Examples are the number of runs scored in a game of baseball, number of students in a class, and number of stars in the sky. *Continuous data*, on the other hand, can take any value on the real number line including rational and irrational numbers. Examples of continuous data are duration of wait time for ice cream, an athlete's height and weight, proportion of students who play team sports, speed of a tennis serve, and water temperature at the beach.

Google's *Gemini AI* can help delve deeper into different types of data and how they are used in specific fields such as fashion and studying public spaces. First we provide a brief note on how to use this AI-assistant.

A word about Google's Gemini AI

AI tools make it easy to gather, collate, organize, summarize and interpret relevant data and information from millions of sources, a task that is impossible to accomplish manually. Google Gemini prompts are used throughout this textbook to enable deeper understanding and self-learning. Therefore, using the prompts in the Gemini boxes featured in this textbook, and studying the generated content, is mandatory. Note that any reliable AI chatbot can be used in Gemini's place.

An additional advantage of using Gemini is that it automatically provides the most current material, ensuring that students always have access to state-of-the-art information. However, the onus of verifying the output and using it responsibly lies with the users of Gemini and other AI tools. Screening and checking AI output for errors and inconsistencies is essential because these tools are still evolving, and AI has been known to make errors.

ReferenceBox 2.1.2: How to use Gemini AI

Gemini can be accessed either through a web browser at the URL gemini.google.com, or through an app. Gemini prompts can be typed in many human languages, for example, English, Spanish, Chinese, Arabic and Japanese, and several Indian languages such as Hindi, Telugu, Tamil and Kannada.

Gemini, in addition to a response to the prompt, provides a list of the sources from which it has collated the response, and additional sources. This is an excellent feature that helps validate the output and provides opportunities for further research.

Gemini can even help collect data. [Figure 2.2](#) shows the results of uploading an image and prompting Gemini to describe and identify the object in the image. The image is shown in [Figure 2.2](#), left, the prompt is shown in [GeminiBox 2.1.1](#), and the output is shown in [Figure 2.2](#), right.



Image Description and Object Identification

The main object in the image is a plant with **brightly colored flowers** and **green, textured leaves**.

- **Flowers:** The flowers are clustered in dense, rounded heads, or umbels. They display a vibrant mix of **red, orange, and yellow** colors. In many of the flower heads, the individual florets appear to transition in color from yellow in the center to orange and then red at the outer edges.
- **Leaves:** The leaves are **dark green**, have a **serrated or toothed edge**, and appear to have a somewhat **rough or wrinkled texture** with prominent veins.
- **Context:** The plant is shown outdoors, possibly in a garden, with other green foliage slightly blurred in the background, suggesting it is a shrub or bedding plant.

The plant in the image is a type of **Lantana** (likely *Lantana camara*), a common perennial or annual known for its colorful, multi-hued flower clusters.

FIGURE 2.2: Left: Image uploaded to Gemini (Photo: T. Dasu). Right: Gemini output in response to the prompt in [GeminiBox 2.1.1](#). ↗

GeminiBox 2.1.1: Image Data ↗

“Describe the uploaded image and identify the object.”

Data Debate 2.1.1 *What are the qualitative variables in [Figure 2.2](#), right? Is the description of the image generated by Gemini accurate?*

Try the prompt in [GeminiBox 2.1.2](#) to find examples of qualitative data, or click on the link to go directly to the content generated by Gemini in response to the prompt.

GeminiBox 2.1.2: Qualitative Data ↗

“Give examples of qualitative data in social interactions in public spaces.” 🗨️

Data Debate 2.1.2 *What do you think of the output generated by Gemini? Create and submit your own Gemini prompts to learn more about types of*

data.

GeminiBox 2.1.3: Quantitative Data

“Give examples of quantitative data used in fashion.” 

Data Debate 2.1.3 *What are the different types of data you come across in your daily life? What are the hurdles to collecting data? When is it not possible to collect data?*

2.1.2 Data types based on how obtained

An alternate way of defining data types is based on how the data was obtained, whether first-hand through design and intent, and by whom. This type of categorization helps in determining the usability and reliability of the data, the type of preprocessing it might need and the accuracy of the results.

ReferenceBox 2.1.3: Data types based on how obtained

- *Primary data* is collected by the person using the data, for the purpose of a specific task such as predicting ocean currents.
- *Secondary data* is collected by an external source, preferably of repute, like NOAA, NASA or CDC. These organizations collect data for their own projects and research, but they make it available for public use under various *open data* initiatives.
- *Found data* is a third data type based on its origin, so named by Sir D. R. Cox, a famous statistician, who used the phrase “found data” for non-primary data that was not collected as a part of an

experiment or designed with a purpose in mind. It is just happenstance observation data whose origins are sometimes obscure, e.g., a video that is forwarded millions of times on social media, sometimes with alterations like filters or edits by intermediate users.

Primary data is considered the most reliable type of data because it is specifically designed and collected with the problem in mind, for example, a survey of a set of students on campus to identify the most frequently played sport on a campus. On the other hand, primary data might be too specific to be used generally. For example, sports survey data collected from a campus in Alaska might not be useful for predicting sports preferences on a campus in rural India.

A good example of reliable secondary data is a result of government mandates called *Open Data* initiatives that specify that data collected using government funding should be made available to the public. The data is available to the world-wide community for analysis and research, and therefore it is used and vetted quite rigorously. [Figure 2.3](#) shows two sources of high-quality secondary data: Prediction of Worldwide Energy Resources (POWER) provided by NASA (National Aeronautics and Space Administration)  and CDC (Centers for Disease Control) data from a variety of sources, e.g. the National Center for Health Statistics, on topics such as smoking, vaccines and other health trends . Several US federal government agencies maintain high-quality research grade data which they make available to the public.



FIGURE 2.3: Open data initiatives increase access to public information in an effort to make governments more transparent and accountable. They are an example of *secondary data* for external users. Left: Website to access data from NASA's Prediction of Worldwide Energy Resources (POWER) project. Right: Data dashboard for a variety of CDC health topics. [↗](#)

Found data tends to be passively observed data and often falls under the category of Big Data. Data that are automatically recorded by sensors and software platforms, e.g., automobile sensors, clicks on social media, third-party audio/videos, can be termed found data. Such data tends to have quality issues, and is often not usable because there is not enough information about the data. For example, if you are given a spreadsheet with columns of data but no column headers, would you be able to use it?

There are other data types as well, such as *metadata* and *AI-generated content (AIGC)*. Metadata is “data about the data,” that make it easy to understand the data. For example, the Palmer Penguin data set refers to 344 penguins, with the variables species, island, flipper length, etc. Without metadata, most data becomes less reliable.

In addition, AI has given rise to new types of data, AIGC. Output generated by Google Gemini in response to prompts is an example of

AIGC. At this time, AIGC is not well understood, yet it has significant impact. For example, AI-generated videos, called “deepfakes”, featuring celebrities, or videos of one community provoking another community are common on social media. Such fake videos create chaos and unrest. In addition, AI algorithms are prone to making up facts, called “hallucinations.” For example, if a user requests the names of five spy thriller movies, AI might find three real ones and make up the names of two others. We will learn more about AI generated hallucinations and deepfakes [Chapter 9](#).

ReferenceBox 2.1.4: Other types of data

- *Metadata* consists of descriptions of data. Column headers, descriptions of the variables, information about how the data was collected and for what purpose, are all examples of metadata.
- *AI-generated content (AIGC)* is a special type of data generated by AI applications in response to user-provided prompts. The output of the Gemini AI prompts in this textbook is an example of AIGC.

2.2 Data Collection

In the past, data was collected manually through painstaking observation, experimentation and measurement. Dr. Jane Goodall  collected data on chimpanzees and their social structure and behavioral habits through observation over a period of 30 years in East Africa. Gregor Mendel  filled numerous notebooks with measurements, descriptions and comments

about pea plants and their evolution over generations, laying the foundation for modern genetics. Such manual data collection methods tend to be limited by human capacity for observation and recording, resulting in small data sets. In some instances such as the occurrence of rare diseases, the amount of data available is inherently limited.

Just like Mendel and Goodall, modern-day marine biologists and research scientists continue to collect data through observation. [Figure 2.4](#), left, shows a researcher gathering laughing gull chicks for observation and measurement, collecting qualitative data (e.g., species, presence of diseases), and quantitative data (e.g., weight), before the chicks are returned to the wild. However, unlike Mendel and Goodall, these scientists have a mind-boggling array of data collection tools available to them, because trackers and sensors can be attached to any device and connected to a computer for collecting the data. [Figure 2.5](#) shows examples of modern scientific data-collection tools for observation and measurement. On the left hand side of the figure, marine scientists collect image data, and on the right, scientists attach a tag to a whale. The tag will send the GPS coordinates of the whale to the scientists so they can study the effect of human-caused noise on whale behavior.

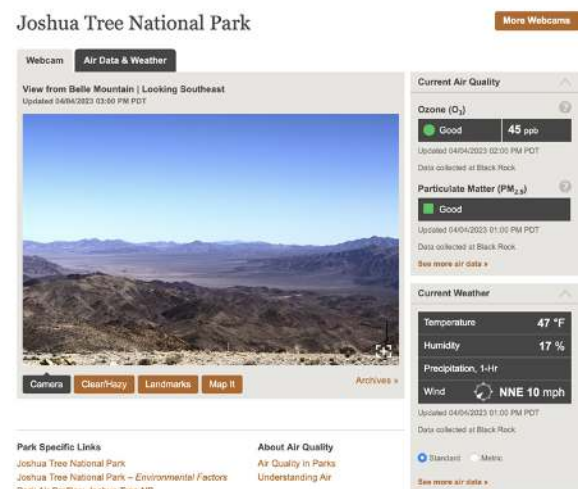


FIGURE 2.4: Left: A researcher from the National Audubon Society gathers laughing gull (*Leucophaeus atricilla*) chicks from Eastern Egg Rock, an island in the U.S. state of Maine. The chicks will be tested for disease, as well as measured and weighed before being returned to the wild. Right: National Environmental Satellite Data and Information Service (NESDIS) provides data to monitor air quality, e.g., at Joshua Tree National Park. [📄](#)



FIGURE 2.5: Left: NOAA's marine scientists collect image data about white sharks in Greater Farallones National Marine Sanctuary. Right: A drone approaches a sei whale to attach an acoustic recording tag. This is the first time a drone was used to tag free-swimming large whales in U.S. waters. Photo credit: Laura Howes under NOAA Fisheries Permit No. 18786-06. [📄](#)

Watch the video in [Example 2.2.1](#) about how NOAA scientists collect data about white sharks, and participate in [Data Debates 2.2.1](#) and [2.2.2](#).

Example 2.2.1. [📄](#) *White Shark Research (NOAA Greater Farallones National Marine Sanctuary): NOAA scientists collect data about white sharks to study their coloration, feeding strategies and social behavior.* [🔗](#)

Data Debate 2.2.1 [📄](#) *What are the means by which researchers collect data in this video? What kind of data do they collect? How will their*

research help white sharks? 

Data Debate 2.2.2  *What should marine scientists be concerned about while observing animals in their natural environment? Are there any ethical concerns as well as impact on the environment during the course of observational studies for scientific research?*

ReferenceBox 2.2.1: Methods of data collection

- *Manual observations and measurements* are a common method of data collection, particularly in specialized scientific experiments and studies. These data can be quantitative, or qualitative, and tend to be limited in size due to human capacity.
- *Automated data collection* is done by software and AI applications, e.g., through sensors, cameras and network probes, by monitoring online activity on social platforms, or by extracting it from online content like news websites.
- *Time-related measurements* are special types of data that are measured repeatedly over a period of time. *Time series and data streams* are examples of such time-dependent measurements.

Over the decades, data collection has evolved with highly efficient technology that can collect data automatically, economically and without endangering humans. Data can be collected by satellites to capture images and radio signals, by automatic sensors such as rain gauges, ocean buoys, temperature and pressure sensors, and by air-quality monitors. Such automated data-collection methods have led to amazing AI applications, some of which are described below.

- [Figure 2.6](#), left, shows a NASA satellite image of Manhattan at night used for studying the effect of light on climate and bird migrations. Image data collected by satellites is used in many AI applications including early detection of natural disasters.
- Computerized machines like cars, rockets and robots generate petabytes of data from millions of sensors that continuously collect data to monitor and guide the functioning of these machines.
- Drones collect images, audio and video of traffic for urban planning, accident monitoring, and for assessing damage caused by tornadoes, floods and wildfires.
- Genetics, which has come a long way since Mendel, is another important area where large amounts of data are used to study how diseases are passed from generation to generation. [Figure 2.6](#), right, shows a National Cancer Institute's image of cancer cells and how they move and spread through organs.
- Social media platforms automatically capture data about online behavior such as purchases, viewing behavior, likes and dislikes, friend networks, and other preferences. They use such data to personalize online experience and recommend shows, services and products that are most likely to appeal to a user.

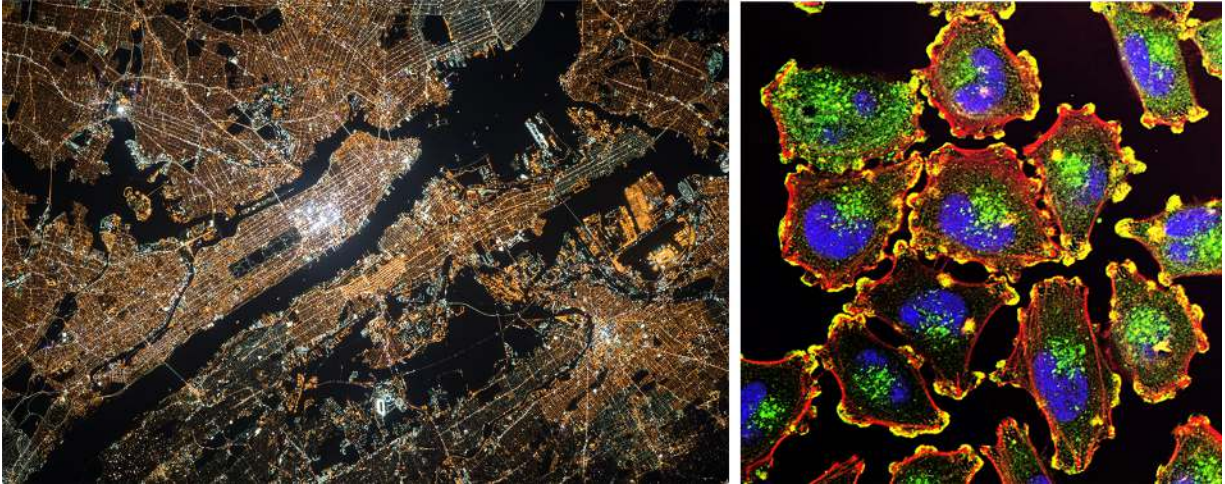


FIGURE 2.6: Left: NASA satellite captures Manhattan at night. Such images are used to study the effect of light on migratory birds and climate. Right: DNA and genomic data are used for gene sequencing and identifying genes that trigger diseases like cancer. National Cancer Institute's image shows how cancer cells move and spread. [↗](#)

Automated data collection by software probes, sensors, transaction logging systems and other means results in millions of observations per second leading to massive amounts of data, referred to as *Big Data*.

ReferenceBox 2.2.2: Big Data [↗](#)

Automated data collection has led to the creation of vast amounts of data, often referred to as Big Data. It is characterized by:

- *Volume*, massive quantities
- *Velocity*, collected with great speed
- *Variety*, from different sources, and of many types, e.g., text, images, video, audio, unstructured data

In addition, it should have:

- *Validity*, of relevance to the application and project
- *Veracity*, of high quality

Specialized tools and software are required to manage Big Data. Data streams that are collected and updated in real-time are examples of Big Data, e.g., a live video feed of the Super Bowl that sends visuals as the game unfolds in real life.

GeminiBox 2.2.1: Big Data

“Give me three examples of Big Data.” 

Data Debate 2.2.3 *Choose one example from the three provided by Gemini AI in response to prompts in GeminiBox [2.2](#). Discuss with reference to the five characteristics, or V's, of Big Data in [ReferenceBox 2.2.2](#), and determine if they meet the criteria.*

2.2.1 Manual data collection methods

Manual data collection happens through observation and measurement, and through posing questions with different objectives. We have discussed examples of observational measurements such as those gathered by Jane Goodall, Gregor Mendel, and researchers like the one in [Figure 2.4](#), left. Other methods of manual data gathering consist of asking questions and seeking information. Common examples are surveys, focus groups and interviews. These methods are popular in social sciences where people's opinions, attitudes and other behavioral characteristics of interest are measured.

For example, political polls are common around election season. They are conducted by academic institutions, media companies and social research organizations by sending out surveys or questionnaires to carefully selected sets of people. Such surveys were frequently used during the pandemic to gauge the health and well-being of the population, impact of the lockdown on school-aged children, and attitudes toward vaccines. The Pew Research Center, which describes itself as a “fact tank”, uses surveys extensively for opinion polling and social science research. It provides guidance on survey design in its article *How to Design a Survey* in [Section 2.5. Designing Surveys: A Guide to Decisions and Procedures](#) by Blair *et al.* [5] provides a thorough introduction to survey design.

ReferenceBox 2.2.3: Methods of manual data collection

- *Questionnaires, surveys and forms:* A questionnaire consists of a set of questions created with a specific purpose to collect data from individuals. Those who answer a questionnaire are called *respondents*. Questions can be open-ended where the respondent can answer in free text as they choose, or closed questions where they will have to choose from a given set of options.

A survey is a method of collecting data using a questionnaire.

Forms are used to collect specific pieces of personal information. Unlike surveys which are intended for learning about the whole group and not any particular individual, forms are intended for personal information in the context of a specific task related to the individual, e.g., evaluation for school admission or eligibility for participating in a sport.

- *Focus groups* consist of hand-picked individuals whose opinions are collected on a specific topic, product or service.
- *Interviews* are one-on-one data collection methods used to gather detailed, in-depth information, either in person or via a teleconference.

The U.S. Census Bureau conducts a special survey every ten years, called a *census*, to collect demographic data about the population of the United States by state and county. A census requires that every single individual in a population be included in the survey. The original purpose of the U.S. Census (according to Article 1, Section 2 of the U.S. Constitution) was to apportion representatives and to direct taxes among the states. Census data is now commonly used for research, marketing, and planning. For example, population and housing data might determine where schools are constructed or where new bus routes are placed. The articles *U.S. Census Bureau at a Glance* and *What Is a Census and What Kind of Data Is Collected?* in [Section 2.5](#) provide additional information.

Surveys, like the one in [Figure 2.7](#), right, often use a Likert scale that uses pre-defined categories like “Mostly positive”, “Neither positive nor negative”, and “Mostly negative” to gauge the sentiment of participants. Limiting responses to a handful of options makes it easier to process and analyze the data from the survey respondents. For a hands-on experience of surveys, explore the prompt in [GeminiBox 2.2.1](#).



More teens say social media has had a negative effect on people their age than on them, personally

% of U.S. teens who say social media has had a ___ effect on ...

	Mostly positive	Neither positive nor negative	Mostly negative
People their age	24	45	32
Them, personally	32	59	9

Note: Teens are those ages 13 to 17. Those who did not give an answer are not shown.

Source: Survey conducted April 14-May 4, 2022.

"Connection, Creativity and Drama: Teen Life on Social Media in 2022"

PEW RESEARCH CENTER

FIGURE 2.7: Left: Weather station in Liberty Park, Manhattan, gathers weather data in real-time using sensors. (Photo credit: G. Murthy). Right: The Likert scale used in this Pew Research Center's survey is based on limiting the answers on a survey to a small number of ranked categories, e.g., "Mostly positive", "Neither positive nor negative", and "Mostly negative". [📄](#)

Data Debate 2.2.4 *What type of data is a Likert scale measurement?*

GeminiBox 2.2.2: Survey design

"How do I design a survey to find the most popular starburst flavor?" [🔗](#)

Data Debate 2.2.5 *What do you think of the output generated by Gemini in response to the prompt in [GeminiBox 2.2.1](#)? Create and submit your own Gemini prompts to learn more about survey design, e.g., beauty products. Conduct the survey in your class.*

Focus groups are used in marketing, pharmaceuticals, television and politics as "trial balloons" for new ideas and products. For example, when a

new TV show is being produced by a network such as NBC, they select a focus group of a specific age, e.g., 35-59, and compensate the participants for sharing their TV viewing preferences. *An ongoing list of focus groups actively seeking participants at the website focus-groups.org in [Section 2.5](#)* features a list of focus groups actively seeking participants.

Interviews are perhaps one of the most labor-intensive data collection methods, either involving face-to-face meetings between individuals or sending a set of questions that require long-form answers. Ethical concerns such as privacy, keeping the content confidential, e.g., psychiatric interviews, and limiting access to the content become critical. Interviews play a key role in medical research.

Data Debate 2.2.6 *What type of data would you collect if you were:*

- *An epidemiologist studying vaccine effectiveness? What could be the ethical, privacy and data quality issues?*
- *A scientist studying endangered species?*
- *An online company trying to research the type of movies popular with the age group 11-14?*

How would you collect this data?

Data Debate 2.2.7 *Compare surveys, forms, focus groups and interviews. Which are the most likely to be biased and why? Is it good or bad to have such bias?*

2.2.2 Automated data collection methods

The explosive growth of AI has been fueled by automated methods of data collection. They are the norm in many commercial and scientific endeavors,

and they result in a richer, more complete type of data.

Sensors, like the ones on the weather station in [Figure 2.7](#), left, can be attached to a variety of objects such as drones, balloons, ocean bouys, or spacecrafts, and they can record measurements, audio and video. Any method of deploying a sensor can be used for data collection. Here are some examples.

- *Drones for collecting images:* Drones have become important data-collection tools across many fields of study and business. They can carry different types of equipment including sensors, cameras and communication devices. Drones can be used for studying changes in oceans and snow caps, assessing impact of natural disasters like floods and wildfires, migration patterns of animals, traffic flow on major roadways, military movements, monitoring hard-to-reach equipment like cell towers, and studying the health of herds and crops.
- *Satellites for monitoring air quality:* NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) and the Environmental Protection Agency collect hourly data via satellites to provide continuous air quality data to the public at their portal [AirNow](#) . [Figure 2.4](#), right, shows the information for Joshua Tree National Park in California.
- *Balloons for gathering climate data:* Weather instruments like radar trucks, weather stations with sensors and gauges, ocean buoys, and even balloons collect huge amounts of data continuously.

Feedback is another type of automatically collected data. It is an essential part of training robots, building smart homes and predictive AI models. Feedback actively changes the behavior of a system. For example, a smart home's thermostat, like the one shown in [Figure 2.8](#), left, keeps adjusting

the settings depending on time of day, whether the house is empty or occupied, which floor the occupants are likely to be on, and other such criteria. It can result in considerable energy savings. Another example of a feedback loop, known as closed-loop monitoring, is used to monitor wind turbines that harness wind power, shown in [Figure 2.8](#), right. The angle of the blades, speed of rotation and anomalies are continuously controlled through a feedback loop to minimize wear and tear while maximizing energy capture. Feedback loops are used by social media companies to constantly tweak the look and feel of their apps and websites to increase user engagement.



FIGURE 2.8: Feedback is a form of data collection where the output of a system is used to monitor and influence the system's behavior. Left: A smart thermostat adjusts settings depending on occupancy, external temperatures and time of day to optimize energy use. (Photo: Raysonho). Right: The angle of blades and speed of rotation of wind turbines are monitored using feedback to minimize damage and maximize functionality. (Photo: Erik Wilde) [↗](#)

ReferenceBox 2.2.4: Methods of automated data collection

- *Sensors* attached to surfaces of drones, ocean buoys, spacecraft, for collecting measurements about light, temperature, atmospheric pressure.
- *Cameras and microphones* for capturing images, audio and video mounted on buildings or moving objects.
- *Feedback loop* data to capture the output of a system continuously as input to affect its future output to meet specified objectives, e.g., minimize energy consumption in a smart home.

GeminiBox 2.2.3: Feedback Loop

“Give me an example of a feedback loop.” 

2.3 Storing, Managing and Accessing Data

Data storage has a long history. Humans have always wanted to communicate and store data for future use. Consider the collage in [Figure 2.9](#), starting at the top-left corner and examine the images in a clockwise direction. One of the earliest examples of data stored for future generations is in the form of cave paintings. Later, people developed language and passed on data by memorizing it and teaching the next generation, often in the form of stories, songs and lectures. Then came written scripts, where information was chiseled on stone tablets or scratched out on papyrus or written with ink on parchment. Manuscripts were written and copied painstakingly for safety and distribution, which also led to beautiful illustrations and art. [The Book of Kells](#), displayed at Trinity College in

Dublin, Ireland, is an exceptional example. Subsequently, paper and notebooks became common ways of storing data. The invention of the printing press sped up the process of storing and distributing large amounts of data, whether numbers, text or images.

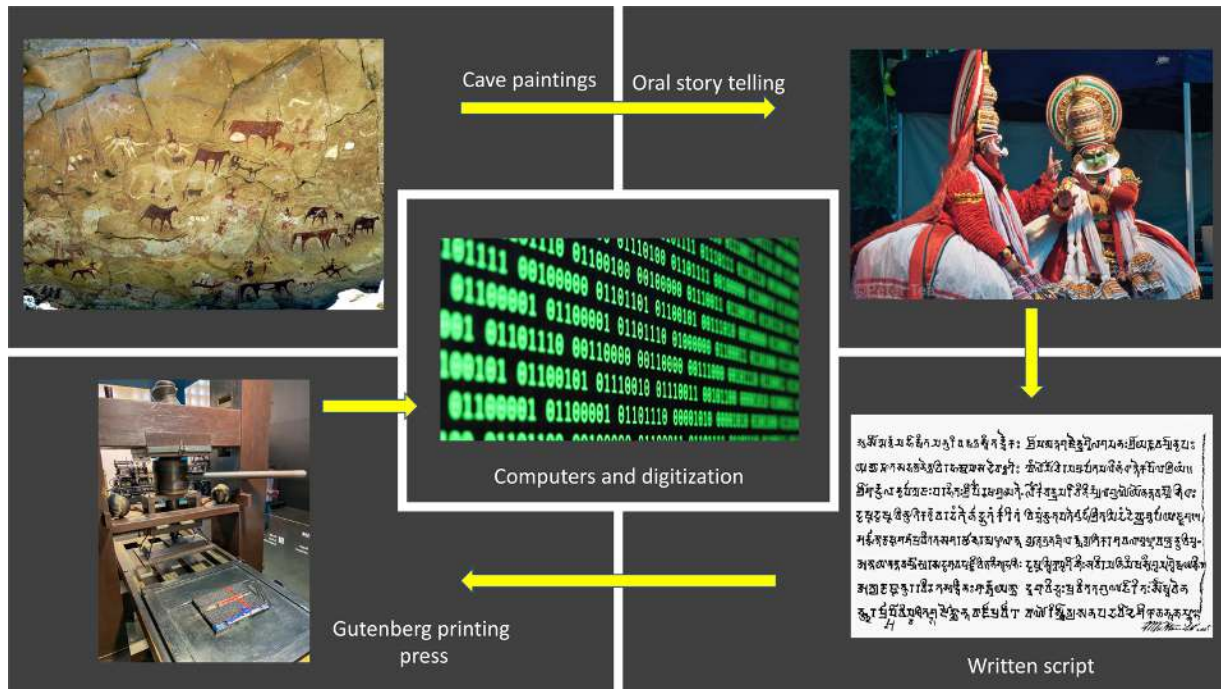


FIGURE 2.9: Data storage methods over centuries, some milestones, clockwise from top left: (1) Cave paintings. (2) Oral traditions of passing on data through storytelling. (3) The development of written script to store data on stone tablets, papyrus and other writable surfaces. (4) The printing press sped up the process of storing large amounts data in newspapers and books. (5) Computers and digitization resulted in Big Data, and cloud storage resulted in the ability to access any type of data, any size, any shape, even movies, from any place, anytime. [↩](#)

Now, thanks to *digitization*, the ability to represent any information as a combination of 0's and 1's (called “bits”), unimaginable amounts and varieties of data can be stored on computers and in the cloud, accessible

from any place, anytime. Without digitization, there would be no smartphones or computers, no streaming movies, no addictive social media apps; in fact, there would be no internet! Isn't that a remarkable journey, from etchings on a cave wall, to any and all data, accessible all the time from anywhere?

Management of data is highly formalized and studied systematically by the field of database management systems. It is a vast topic that requires an entire sequence of courses to understand the fundamentals. It is not the focus of this book. However, *Database Management Systems* by Ramakrishnan and Gehrke [34] provides an excellent introduction.

Example 2.3.1 *Digitization of medical records: Digitization has revolutionized the health care industry where medical professionals increasingly use Electronic Medical Records (EMR) of patients at hospitals or clinics. EMRs allow instantaneous sharing of information leading to better diagnoses. It is believed that EMRs can improve the quality and efficiency of health care, as well as make the healthcare process more transparent to patients. To learn more, read the article Electronic Health Records in [Section 2.5](#).*

Data Debate 2.3.1 *Discuss examples of data storage. How do computers and automation help? What are some challenges?*

Data storage is a part of daily life. We take notes, store contacts and photographs on our phone, create spreadsheets to track expenses, and download interesting news articles. Some examples of common data storage methods and tools include:

- *Manual handwritten records* are rare now, but for centuries data was recorded manually in notebooks and logbooks.

- *Spreadsheets* are used to store *tabular data*, i.e., data organized in rows and columns. Examples include Excel spreadsheets and Google Sheets. [Figure 2.10](#), left, is an example of a spreadsheet used to record the medal tally at the 2022 Winter Olympics. Both qualitative and quantitative data can be stored in spreadsheets. Some spreadsheets allow embedding images and audio/video objects as well. Companies like Microsoft and Google incorporate AI into their data toolsets. [Figure 2.10](#), right, shows an old-fashioned card catalog used to record, store and retrieve information about books in libraries before the digital age.
- *Image libraries* include iPhotos, GooglePhotos and Flickr where the images are *tagged* with labels like “cat” or “flower” or “baby”, to make it easy to search. Facebook, iPhotos, GooglePhotos automatically create personalized memory albums from image libraries.
- *Multimedia libraries* store music and movies, and they are used by streaming companies such as Spotify and Netflix. Such databases can store information about genre, artists, duration and popularity along with the actual movie or music. They can also collect information about who accessed any given title and for how long, and they make recommendations to friend-networks based on an individual's viewing patterns.
- *Knowledge bases* typically store rules used by AI applications such as Siri. The rules can be complex, involving logic and scenarios. For example, self-driving cars function using training rules such as “if traffic light is red, come to a total stop.” AI applications also learn strategy games by ingesting vast amounts of data, e.g., Google's *AlphaGo* that “trained” itself on millions of Go games and eventually beat the world's

best player, Lee Sedol. Watch the fascinating documentary *Alpha Go* referenced in [Section 2.5](#).

- *Databases* or database management systems (DBMS) are special software applications that permit storing and efficiently accessing large volumes of different types of data. They come in different types and are used by companies and scientists to quickly store, access and process data for analysis. For example, *Oracle*, *Hadoop* and *Teradata* are databases that are used for everything from marketing sneakers, to observing the oceans and predicting tidal patterns. There are specialized databases for spatial and temporal data, images, connection graphs like friend networks, a mix of different types of data (unstructured), and many other types. Database management is an entire course of study by itself. Explore [GeminiBox 2.3.1](#) for a brief overview.
- *Specialized database systems* are the underpinnings of AI. Machine Learning and AI applications use special databases that provide speed, flexibility, analytics and natural language processing abilities to interact with humans through human languages. Examples include Google's *Tensor Flow*. Generative AI applications use vector databases that allow both structured and unstructured data to handle alphanumeric data, text, images, and audio and video.

Winter Olympics 2022 Beijing Medals Table

Rank	NOC	Gold	Silver	Bronze	Total
1	Norway	16	8	13	37
2	Germany	12	10	5	27
3	China*	9	4	2	15
4	United States	8	10	7	25
5	Sweden	8	5	5	18
6	Netherlands	8	5	4	17
7	Austria	7	7	4	18
8	Switzerland	7	2	6	15
9	ROC	6	12	14	32
10	France	5	7	2	14
11	Canada	4	8	14	26
12	Japan	3	6	9	18
13	Italy	2	7	8	17
14	South Korea	2	5	2	9
15	Slovenia	2	3	2	7
16	Finland	2	2	4	8
17	New Zealand	2	1	0	3
18	Australia	1	2	1	4
19	Great Britain	1	1	0	2
20	Hungary	1	0	2	3
	Belgium	1	0	1	2
	Czech Republic	1	0	1	2
21	Slovakia	1	0	1	2
24	Belarus	0	2	0	2
	Spain	0	1	0	1
25	Ukraine	0	1	0	1
	Estonia	0	0	1	1
	Latvia	0	0	1	1
27	Poland	0	0	1	1
Totals (29 entries)		109	109	110	328



FIGURE 2.10: Left: Spreadsheet of medal tally at the Winter Olympics in Beijing, 2022, based on information provided by the International Olympic Committee (IOC). Right: An old-fashioned database, a card catalog of library books. Photo credit: “Original Database” by shindoverse. [🔗](#)

GeminiBox 2.3.1: Overview of Database Management Systems (DBMS) [🔗](#)

“Give a brief overview of database management systems.” [🗨️](#)

In order to retrieve stored data efficiently and quickly, data are *indexed*. For example, the index at the back of a book has a mapping between keywords and page numbers where the content related to the keywords can be found. The index, which is much more compact than the original data, is searched to locate the appropriate page rather than the entire book. Databases have an analogous way of indexing stored data to make locating the relevant data quicker. Efficient indexing is the key to providing quick search results.

GeminiBox 2.3.2: Indexing

“What is indexing? Give me an example. What is the Dewey Decimal System (DDS)?” 

Data from different applications, e.g., geospatial data versus videos, have different structures, internal relationships and pose different types of technological challenges in terms of storage and other processing. It might not be possible to have a single, general purpose database management system that can handle all different types efficiently, hence the need for customized database applications to handle each type. Database management systems use specialized software and languages to query the database and retrieve relevant data of interest. SQL (Structured Query Language) is a common database query language. Use the prompts in GeminiBox [2.3](#) to learn more.

GeminiBox 2.3.3: Query Language

“How is data retrieved from a database? Give me examples.” 

Database Management Systems by Ramakrishnan and Gehrke [[34](#)] is an ideal introduction for those interested in studying this topic further.

2.4 Conclusion

The science of collecting, managing and analyzing data is continuously evolving due to the dynamic nature of AI and its applications. Such technology is also an effective means of cultivating data intuition which helps unearth useful insights and creates “aha” moments. For example, the turkey products company *Butterball* needed to modernize their data infrastructure in order to integrate customer data with systems that housed financial, operations and marketing data. During the process, they combed through transcripts of decades worth of customer calls to Butterball

helplines, especially near Thanksgiving, and made an important discovery. The biggest pain point for customers? Thawing the turkey. This data-driven insight led their food scientists and researchers to create the “cook from frozen” turkey that did not require thawing. Read *Thank Big Data for Killing Off the Worst Part of Thanksgiving Meal Prep* in [Section 2.5](#) for details.

In the next chapter, we will focus on exploratory data analysis, another important way of developing data intuition. Given that AI results are famously opaque and not very human-friendly, developing a strong data intuition helps us move closer to understanding and interpreting the results of AI algorithms.

2.5 Read & Reflect

- *U.S. Census Bureau at a Glance.* 
- *What is a census and what kind of data is collected?* .
- *How to design a survey.* 
- *An ongoing list of focus groups actively seeking participants at the website [focusgroups.org](https://www.focusgroups.org).* 
- *Electronic Health Records.* 
- *A fascinating documentary about Alpha Go about how Google's AI beat Lee Sedol, the world's top Go player.* 
- *Thank Big Data for Killing Off the Worst Part of Thanksgiving Meal Prep (Wall Street Journal).* 

2.6 Exercises

1. **Data relevance:** Suppose that you want to find the best purchase price for a pair of jeans.
 - (a) What data would you collect? What are the different ways you can gather the data?
 - (b) Would AI be of help? If yes, how?
2. **Qualitative data:** In [Figure 2.2](#), identify and write down as many qualitative variables as you can from the description generated by Gemini.
3. **Metadata:** Present to the class a spreadsheet with three pieces of information that you have created without column headers or description of the data. Ask them to guess what the data is about. How many can guess correctly?
4. **Data collection with drones:** List three data collection activities that drones can be used for.
5. **AI for creating a survey:** Use Google's Gemini AI Chatbot to come up with questions for a survey about a topic of interest. Potential topics include: healthy snacks, workouts, testing a new video game, pizza, makeup, etc.
 - (a) What is your goal in designing this survey?
 - (b) Try different prompts to find the right questions to meet your goals.
 - (c) Are you satisfied with the questions that were generated by Gemini for your survey?

(d) SurveyMonkey is a popular application that can create online surveys using AI and collect and organize data. Try using this application to create a survey about a topic that interests you. 

(e) Compare the surveys created by Gemini and Survey Monkey.

(An example of a survey for determining the most popular Starburst flavor using Gemini is available in the Resources section .)

6. **Census survey:** Questionnaires can contain questions of different kinds: open-ended questions (write a response), closed questions (yes/no), Likert scale (strongly disagree, disagree, neutral, agree, strongly agree), multiple-choice and check-boxes. Look at the census questionnaire in this link: .

(a) What are the different kinds of questions on this census survey?

(b) What type of data (qualitative, quantitative) is being collected in this survey?

7. **Focus group:** You and two friends have created a new video game for children aged 10-13. You would like to test this video game with a focus group and get feedback to improve the game before it is released online.

(a) How would you form this focus group?

(b) What are examples of questions that would you pose to this focus group?

8. **Big Data:** Big Data are measured, for example, in Megabytes, Gigabytes, Terabytes. What comes next? What is the largest known data quantity measurement?

9. **Data streams:** Write down three examples of data streams from different settings. Think of social media, NASA telescopes, stock tickers and so on.

10. **Digitization of data:**

- (a) Write a short paragraph on how digitization, or the binary system, is used to transmit data.
- (b) What is the impact of digitization on the diversity, equity, access, participation, inclusion, or fairness of music cultures? The article “Music, Digitalization, and Democracy” discusses some of these issues and can be found at this link [🔗](#). Use Gemini to summarize this article with an appropriate prompt. Do you think this is an accurate summary of the article? Do you agree or disagree? Why or why not?

11. **Databases:**

- (a) Use Gemini to research different types of databases. Which one would you use if you wanted to build a database to store ancestry information which might include family relationships, property ownership documents and photos?
- (b) What kind of database would you use to store contact information of all your family and friends?

2.7 Summative Project

Exploring Data: Wind Power

Global Wind Atlas, shown in [Figure 2.11](#), is an application for exploration of wind speed around the globe using an interactive map and tools. The website [🌐](#) has information and data sets available for researchers and organizations to use.

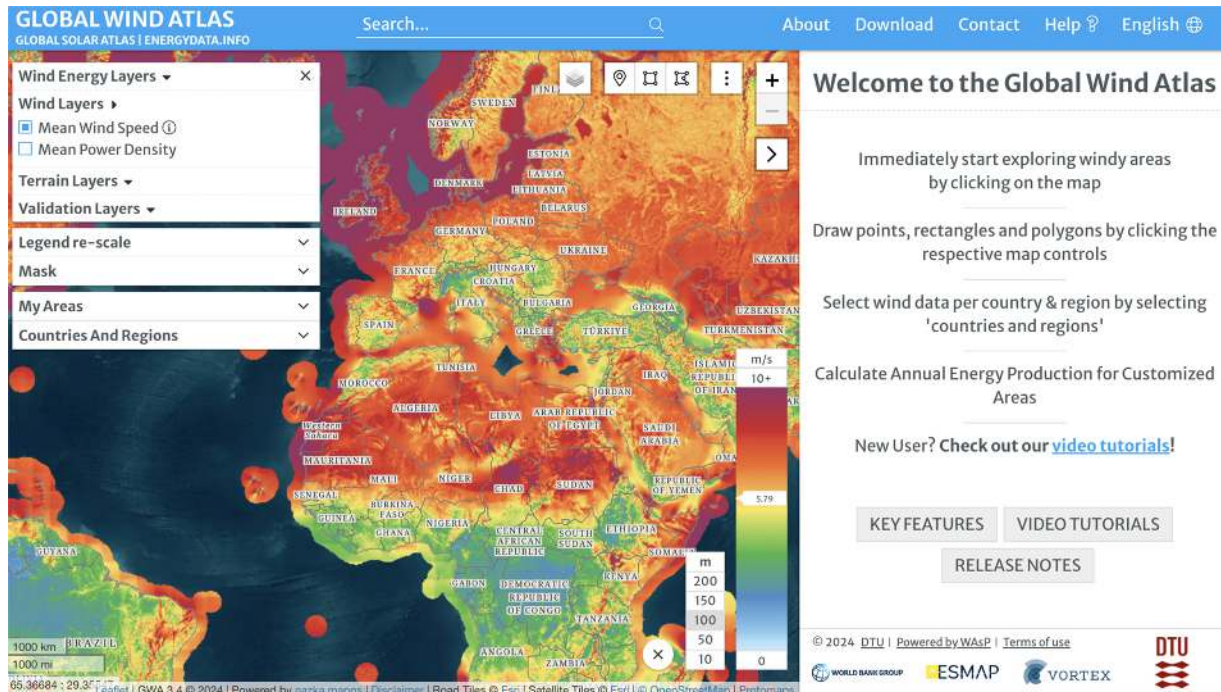


FIGURE 2.11: Global Wind Atlas (GWA) is a web-based interactive application that allows you to explore wind speed observed in different regions around the world. The GWA App is owned and maintained by the Technical University of Denmark (DTU), a university registered in Denmark. [🔗](#)

Suppose you are part of a research team for an organization/corporation that is looking for potential sites for wind power production. Your team is exploring the web-based Global Wind Atlas application to gather information. The Map and the tools in the application are interactive and allow you to explore different regions from around the world. The video

tutorials are helpful for navigation of the map and learning how to use different tools. Go to the web application and start exploring!

Part 1: Types of Data Answer the questions as you explore the map using the definitions and terms we discussed in [Chapter 2](#).

1. What are your general observations when you first see the map?
2. What do the different colors in the map signify?
3. What types of data are available? Explain and classify three different data you observe. (For example, country name is qualitative data; wind speed is quantitative)
4. Using the tools provided, identify a country and region that you think is suitable for wind power.
5. What tools do you think were used to gather data for this application?

Part 2: Collecting and Understanding Data Select three regions in a country where you would like to develop a wind farm by using the map controls.

1. What regions did you select? Why?
2. List the relevant data for these three regions.
3. What can you conclude from this data?

Attribution: “[Data/information/map] obtained from the Global Wind Atlas version 4.0, a free, web-based application developed, owned and operated by the Technical University of Denmark (DTU). The Global Wind Atlas version 4.0 is released in partnership with the World Bank Group, utilizing data provided by Vortex, using funding provided by the Energy

Sector Management Assistance Program (ESMAP). For additional information: <https://globalwindatlas.info>”

Exploratory Data Analysis

Exploring data from different perspectives is essential to establish its relevance to the problem, to unearth data quality issues, and to develop an intuition for that data in its real-world context. This process is called *Exploratory Data Analysis (EDA)*, where data is analyzed outside the framework of any specific task such as prediction. EDA is an essential aspect of preparing data for AI. This chapter introduces the basics of EDA such as descriptive statistics and visualization techniques. AI tools have made it easy to perform EDA on a large scale, and AI applications rely extensively on EDA to calibrate and make real-time decisions.

The United States Census is carried out every ten years and aims to count every person living in the United States. That is, more than 300 million people, resulting in 300 million observations. Studying each individual that lives in the United States would not be possible or useful. However, simple summaries such as counts and averages, and visualizations based on these summaries convey the big picture about what is happening in the country, and identify general trends and characteristics. For example, [Figure 3.1](#) created by the United States Census Bureau is packed with information based on statistics that describe residents of the United States. According to the figure, the number of households with three or more people dropped from 50% in 1973 to 36% in 2023. However, the percentage of homeowners older than 75 doubled from 7% in 1973 to 15% in 2023. Such information is useful for

planning housing, e.g., build smaller homes for smaller households but with more amenities for the elderly.



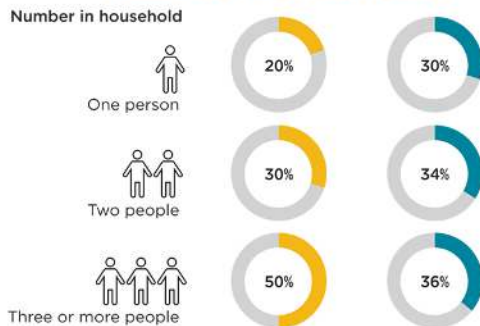
50 Years of the American Housing Survey: Households Then (1973) and Now (2023)



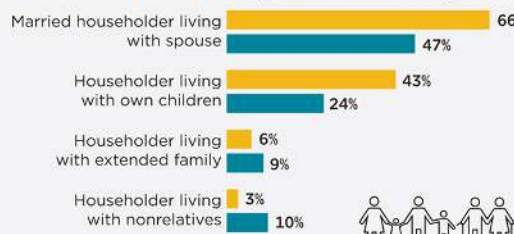
Changes in Householder Race and Ethnicity



Household Size Decreased

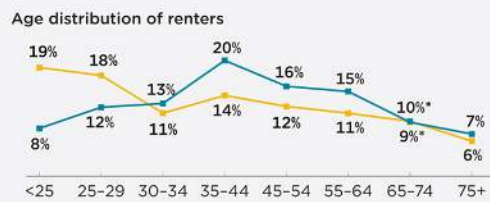
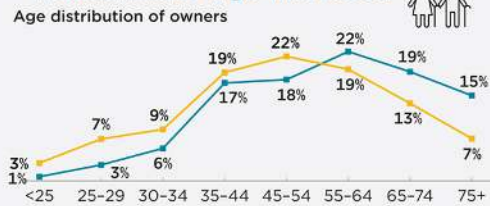


Household Composition Changed



Note: Data on race/ethnicity, marital status, household social characteristics, age, veteran or active military status, and move-in pertain to the householder (the primary owner, renter, or occupant). Household composition categories are not mutually exclusive. More than one category may apply to a unit. "Extended family" means a relative of the householder who is not their spouse or child. When necessary, variables from 2023 and 1973 were recoded for compatibility. Note that in 1973, the military variable reflects householder's service status as of 1970 (three years prior to the survey), whereas in 2023, the military variable reflects householder's current service status. All percentages are rounded to the nearest whole number. All comparisons between 1973 and 2023 are significantly different at the 90 percent level, except for one age comparison. Definitions and more information on confidentiality protection, methodology, and sampling and nonsampling error are available at www.census.gov/ahs. The U.S. Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System [DMS] number: P-7533599; Disclosure Review Board [DRB] approval number: CBDRB-FY25-SEHSD003-066).

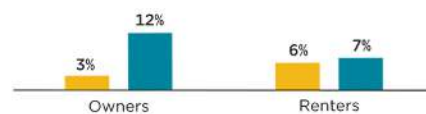
Householder Age Increased



*Not statistically different between 1973 and 2023.

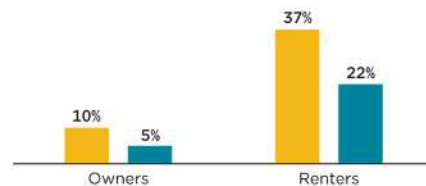
More Veterans and Service Members Became Homeowners

Percent of householders with veteran or active military status



Fewer Owners and Renters Recently Moved

Percent of householders who moved in within last 12 months



United States[®]
Census
Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov



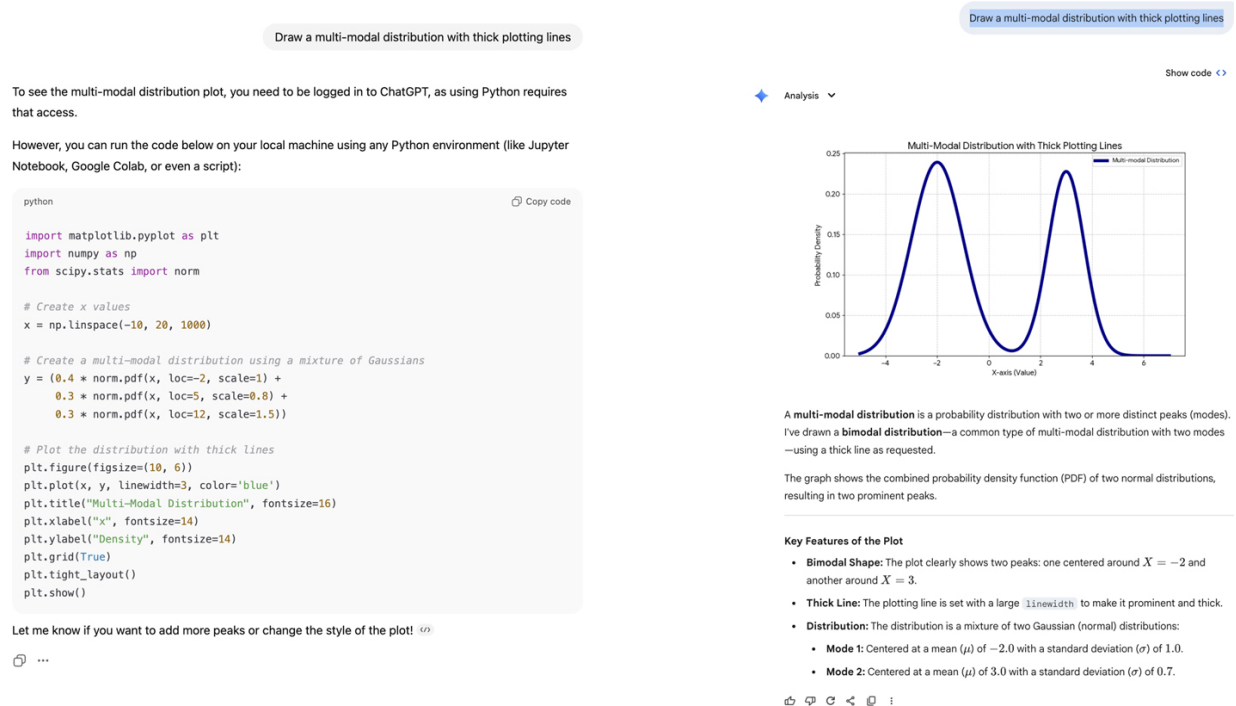
Source: U.S. Census Bureau and U.S. Department of Housing and Urban Development, 1973 and 2023 American Housing Surveys.

[Extended Description](#)

FIGURE 3.1: Changes in American households between 1973 and 2023, EDA based on visualization and descriptive statistics. Source: United States Census Bureau. [↵](#)

The process of creating data summaries and visualizations to understand general characteristics of data is called *exploratory data analysis* (EDA). John Tukey, a famous statistician at AT & T Bell Laboratories, coined the term EDA in 1970, and authored the book *Exploratory Data Analysis* [42]. He would manually compute statistics and draw plots, sometimes with just a pencil and a yellow notepad. That is no longer feasible, given the volume and complexity of data. Specialized software like *R* and *Python* offer packages that perform large-scale EDA and automate repetitive EDA tasks.

AI tools have completely changed the way data scientists approach EDA. For instance, we gave the prompt in [GeminiBox 3.0.1](#) to two AI bots, Google's Gemini and OpenAI's ChatGPT, to plot a data distribution with more than one peak, called a *multi-modal distribution*. It took the AI tools just a few seconds to create the code and generate the plots. The output is shown in [Figure 3.2](#).



Data typically consists of *objects* and *attributes* that either describe or measure characteristics of the objects. Such information is often represented as a table where the rows are objects, and the columns are attributes. For example, in the US Census data, each individual living in America would be an object and correspond to a row, and attributes such as age, gender, ethnicity, homeownership status, etc. that describe characteristics of the person would constitute columns.

ReferenceBox 3.0.1: Objects and Attributes

- An *object* or *entity* or sample is a unit about which data is gathered.
- A *variable* or *attribute* is an observation or measurement related to the object that describes some characteristic of the object.

3.1 Data Sets

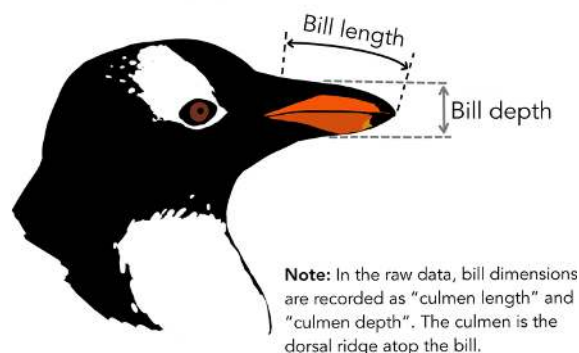
In order to illustrate EDA concepts through practice, we present two data sets, one real and the other hypothetical. The Palmer penguin data was collected by research scientists Katherine Gorman *et al.* [18] that studied the foraging habits of penguins that inhabit the Palmer Archipelago in Antarctica, and whether that behavior differed by gender. The data set was prepared and made available to the *R* community by Allison Horst *et al.* [21]. The data is a mix of different types of attributes, and has the richness and imperfections typical of real-world data that are essential for developing data intuition. The second data set, of the heights of middle school students, is hypothetical and is very small with just 11 observations and a single quantitative attribute. The

purpose of such a simple data set is to explore the data manually, without even a calculator, to understand the concepts and logic behind EDA.

3.1.1 Real-world example: Palmer penguins data

Penguins are unique birds that live in the Southern Hemisphere, on the islands of Antarctica. Scientists at the Palmer Station Ecological Research Program in Antarctica collected data on penguins that inhabit the Palmer Archipelago to study the impact of climate change on these birds, specifically changes to gender-related foraging behavior.

Example 3.1.1.  *The Palmer penguins data contains size measurements for three penguin species, Chinstrap, Gentoo, Adelie, found on the Palmer Archipelago, Antarctica, over a study period of three years from 2007 to 2009. The data set contains 8 variables: species, island, bill length, bill depth, flipper length, body mass and sex. [Figure 3.3](#), right, shows how bill length and bill depth are measured. The study includes 344 penguins.*



[Extended Description](#)

FIGURE 3.3: Left: Gentoo Penguins in Antarctica, walking along a “penguin highway”, a path that joins the sea and their nesting area on a rocky outcrop. (Photo credit: Arturo de Frias Marques. January 2014, [wikiimedia.org](https://commons.wikimedia.org/wiki/File:Gentoo_Penguins_in_Antarctica.jpg).) Right: Measuring bill length and bill depth. (Artwork by Allison Horst.) 

The Palmer penguins data was collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network; Allison Horst et al made it accessible as a part of their R Education project found at [🔗](#).

We will use the Palmer penguin data throughout this textbook. Therefore, it is useful to become familiar with it through EDA. [Table 3.1](#) shows a sample of the Palmer penguin data. Each row in the table pertains to one penguin (and one penguin only). Each column pertains to a variable such as species or bill length. This is a classic tabular representation and spreadsheet storage of qualitative and quantitative data discussed in [Chapter 2](#).

TABLE 3.1: Sample data from the Palmer penguins data set. [📄](#)

species	island	bill_length	bill_depth	flipper_length	body_mass
Adelie	Torgersen	39.1	18.7	181	3750
Adelie	Torgersen	39.5	17.4	186	3800
Adelie	Torgersen	40.3	18.0	195	3250
Adelie	Torgersen	NA	NA	NA	NA
Adelie	Torgersen	36.7	19.3	193	3450
Adelie	Torgersen	39.3	20.6	190	3650
Adelie	Torgersen	38.9	17.8	181	3625
Adelie	Torgersen	39.2	19.6	195	4675
Adelie	Torgersen	34.1	18.1	193	3475



Data Debate 3.1.1 *What are the objects in the Penguin data set? What are the names and data types of the variables? Do you think it is possible to have multiple classes of objects in the same data set?*

3.1.2 A simple example: Middle school student heights

To build data intuition for EDA, consider a simple, hypothetical data set of the heights of 11 students in middle school, measured in inches presented in [Example 3.1.2](#). The small size of the data set makes it ideal for manual exploration.

Data Debate 3.1.2 *What data type is “height”?*

Example 3.1.2. [↗](#) *Heights of 11 middle school students measured in inches: 48, 51, 51, 51, 50, 55, 59, 60, 70, 55, 58.*

3.2 Exploratory Data Analysis

Statistical summaries of attributes such as counts, averages, minimum and maximum values, and visualizations that can highlight trends and patterns are useful in understanding general characteristics of the sample of objects, and to develop a strong intuition for the population of objects as a whole. Exploring data also helps uncover data quality issues such as missing values, assess its relevance and sufficiency for AI tasks, and identify unusual and abnormal values of objects that could either merit special treatment or potentially derail AI models.

ReferenceBox 3.2.1: EDA and Visualization

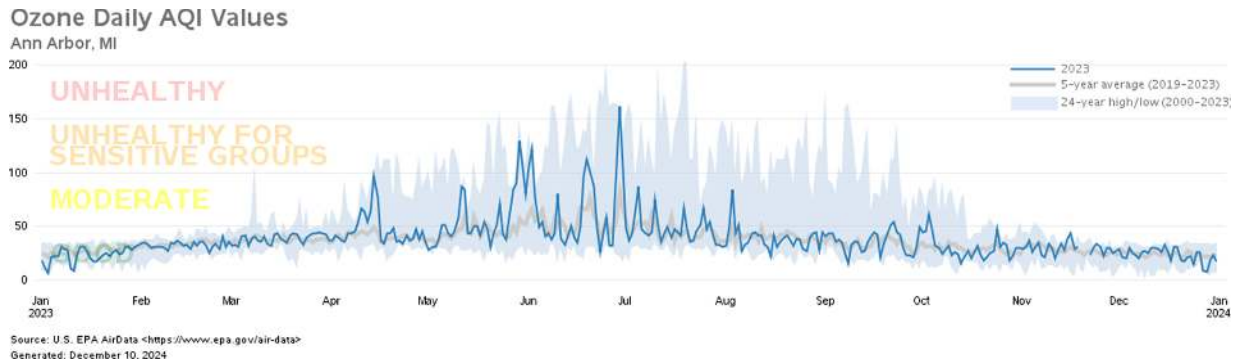
- *Exploratory Data Analysis (EDA)* is the process of summarizing general characteristics of a collection of objects, through statistical summaries and visualizations of the attributes.
- *Descriptive statistics* are computations based on attributes that capture important characteristics of a given collection of objects. The

minimum or the smallest value, maximum or largest value, average and median are examples of descriptive statistics. No matter the size of the data, descriptive statistics provide a compact set of statistical summaries that capture important aspects of the collection of objects.

- *Visualizations* are graphical and pictorial representations of statistical characteristics. Bar charts, dotplots, line graphs, scatterplots and maps are examples of visualization of data.

To understand the power of EDA, consider [Figure 3.4](#) created by the United States Environmental Protection Agency ([EPA](#)) through their interactive air quality index dashboard. Ozone data is used to monitor and forecast ozone levels and warn the public about hazardous air quality conditions which could be life-threatening to vulnerable groups such as the elderly. The data (ozone variations in Ann Arbor, MI during the year 2023) consists of thousands of observations and would be incomprehensible as a collection of numbers, but when summarized and visualized, immediately reveals key insights:

- The ozone levels (dark blue line) in 2023 spiked to dangerous highs during the month of July making the air unhealthy for humans.
- However, the 2023 ozone levels were not the worst (or best) observed in the past 24 years based on the high/low values (light-blue shaded region).
- The ozone levels in 2023 were worse than the five-year average of 2019-2023 ozone levels (thick gray line).



[Extended Description](#)

FIGURE 3.4: The United States Environmental Protection Agency collects air quality data and makes it available through its interactive air quality index dashboard. The plot shows ozone variations with respect to time during the year 2023 in Ann Arbor, MI. [↗](#)

This type of visualization is a time series plot, explained in [Section 3.8.6](#). To explore other ozone data on the EPA website, click [🔗](#).

Data Debate 3.2.1 *Discuss the plot of ozone variations in [Figure 3.4](#). Were there months in 2023 when the air was unhealthy for sensitive groups? What type of statistical summaries were used to create the graph?*

As another instance of the application of EDA, consider the data in [Example 3.1.2](#) in [Section 3.1](#). We can make the following observations:

- The shortest student is 48 inches tall.
- The tallest student is 70 inches tall.
- The most frequent height is 51 inches.
- The height range is 48 to 70, with a spread of $70 - 48 = 22$ inches.
- The middlemost observation when the heights are sorted in order is 55.

Computing quantities such as “shortest”, “tallest”, “most frequent” and “range of values” are common ways of capturing data characteristics and patterns, and good examples of descriptive statistics.

Statistical summaries and descriptive statistics depend on the type of attributes. In this chapter we will discuss summaries suitable for quantitative variables in [Sections 3.3](#) (Typical values), [3.4](#) (Data spread) and [3.5](#) (Shape of the data spread), and summaries appropriate for qualitative variables in [Section 3.6](#) (Counts and proportions).

Important note: There is a distinction between data that we observe, such as the 344 penguins in [Example 3.1.1](#), and the data of all possible penguins, possibly over 40 million in Antarctica. The subset that we observe is called a *sample*, while the set of all possible penguins is called a *population*. We will learn more about samples and populations in [Chapter 4](#). Descriptive statistics based on samples are used to learn general properties of the entire population. In this chapter, we will *define descriptive statistics in terms of computations using the observed data* (as opposed to the probabilistic notion of expected values) since the focus is on data intuition.

3.3 Descriptive Statistics: Typical values

Descriptive statistics are used to convey important characteristics of a group of objects, and to compare different groups. For example, the average height of a middle-schooler is 54 inches, and the average height of a basketball player is 80 inches. Irrespective of the size of the group, whether 11 students or 4 million students, a single number, the average height, represents the typical height of middle schoolers. Averages can also be used to compare typical heights of middle school students (54 inches) with typical heights of

basketball players (80 inches) and infer that on average, middle school students tend to be considerably shorter than basketball players.

The average, or mean, represents the characteristics of a typical or representative object of the group. Intuitively, many objects in the group have attribute values close to the typical value, while a smaller proportion of objects would have attribute values farther away. The notion of *close* is based on a numerical distance, e.g., 4 is closer to 3 than 10.

There are several good choices for describing typical values of quantitative attributes. In this section, we discuss three popular statistics: the mean (also known as the average), median and mode.

ReferenceBox 3.3.1: Mean, Median, and Mode [↗](#)

- The *average* or *mean* of a set of attribute measurements $X = \{x_1, x_2, \dots, x_n\}$ is obtained by adding all the values in X and dividing the sum by the number of values, n , also known as the *size* of the data set X .

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

- A *median* is the “middlemost” observation of a set of ordered values. After ranking the values in X in ascending or descending order, the median is a value in X such that half the values in X are less than the median value, and half the values are greater. Try the prompt in [GeminiBox 3.3.1](#).
- A *mode* of a set of numbers X is the value that occurs most often.

GeminiBox 3.3.1: Median computation [↗](#)

“How is the median computed?” What happens when there is a tie for middle values? 

Review the definitions in [ReferenceBox 3.3.1](#) and use the data in [Example 3.1.2](#) to manually compute the three descriptive statistics to get an intuition for their meaning. Compare your results with those below.

- The mean or average height of the students is sum of the heights of all 11 students divided by the number of students:

$$608/11 = \mathbf{55.27}.$$

Note that while the average height is representative of the whole group, there is no student in the group who is actually of average height.

- The median height is 55, because it is exactly in the middle of the rank ordered data, with 5 values less than 55 and 5 values greater than 55.

48, 50, 51, 51, 51, **55**, 55, 58, 59, 60, 70.

It is 6th largest value among 11 numbers.

- The *mode* of the 11 student heights is 51.

48, 50, **51, 51, 51**, 55, 55, 58, 59, 60, 70.

There are three students that are 51 inches tall, making it the most frequently occurring value of height.

Data Debate 3.3.1 *Which of the three descriptive statistics, mean, median, mode, is the easiest to compute? Which is the hardest to compute? Why? What happens when there are a thousand data points, not just 11? A million?*

The median requires ranking the data in ascending or descending order, called *sorting*. The mean requires adding numbers together; the order of the data

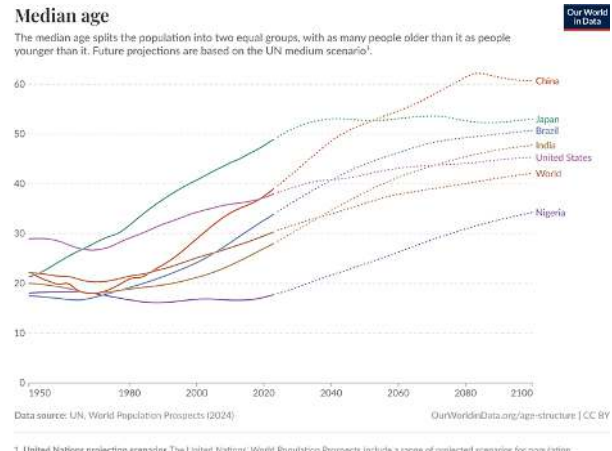
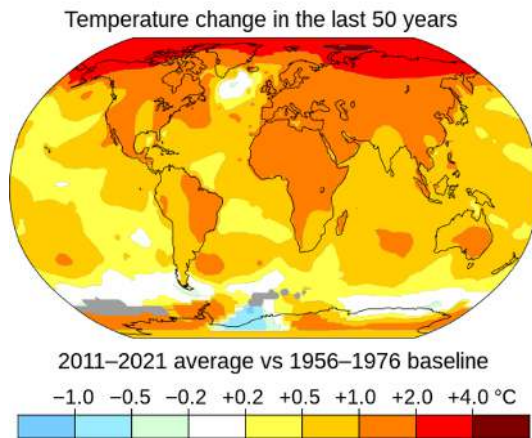
does not matter. Use the prompt in [GeminiBox 3.3.2](#) to learn more about the computing process, and how it varies with the size of the data.

GeminiBox 3.3.2: Mean vs. median

“Which is harder to compute, mean or median, why?” 

Averages and medians are used extensively in many fields because they are easy to understand and useful for comparing different groups. In [Figure 3.5](#), left, changes in average temperatures are used by NASA to study changes in climate. This type of figure where shades of color vary in proportion to the values of the variable is called a *heatmap*, described in [Section 3.8.7](#). The average temperature over the period 2011-2021 is compared with the average temperature during the 20-year baseline period 1956-1976. The difference in the averages is plotted on the world map. Cooler blue shades represent regions where average temperatures have dropped since the baseline period, and warmer red shades represent regions where the average temperatures have risen. The greater the change, the deeper the shades. We can make the following observations:

- The North Pole and regions around it have seen a dramatic increase in average temperature from the period 1956-1976 to the period 2011-2021, an increase between 3.6 ° F to 7.2 ° F.
- A small region near the South pole has seen a cooling of average temperatures, around -1.8 ° F to -0.9 ° F drop.
- Most of the world has seen a general increase in average temperatures as reflected by the preponderance of warmer hues.



[Extended Description](#)

FIGURE 3.5: Left: Daily averages are used in climate studies to track changes (NASA). Right: Median ages are used in demographic studies (Our World in Data). [↗](#)

Data Debate 3.3.2 Based on [Figure 3.5](#), what happened to average ocean temperatures between 1956-1976 and 2011-2021? Where are the regions in the world that have seen the greatest increase, more than 7.2°F denoted by a deep red? Where are the regions that have seen the greatest drop in average temperatures, more than -1.8°F , denoted by blue? What are the two scales C and F? Is there an issue in comparing a 20-year average with a 10-year average?

[Figure 3.5](#), right, shows median ages of the populations of six countries, plus the world, over the period 1950 to 2100. This is a *time series* plot where the values of a quantitative variable are plotted with respect to time, discussed in [Section 3.8.6](#). In [Figure 3.5](#), right, future values of median age based on historical values, called *forecasts*, are shown as dotted lines. We notice that:

- The median age of the world population has been rising rapidly since the 1950s and 1960s, with Japan seeing the steepest increase.

- Based on median age, Japan and the United States have the oldest population at the moment, and India and Nigeria the youngest.
- China will overtake both Japan and the USA by the 2050s in terms of an aging population.
- Nigeria will continue to have the youngest population based on median age.

Data Debate 3.3.3 *What does an increase in median age over time mean? (Hint: longevity) Why do you think that is? What are the consequences of having an older population? Is it an advantage for a country to have a predominantly young population? Would it have been better to use the mean age to study the changes over time? Why or why not?*

3.3.1 Robots and robust statistics

Companies like Amazon use robots and AI in their warehouses to perform physical tasks such as lifting objects that could potentially cause repetitive stress injuries to humans. AI also helps with the optimal placement in the warehouse of goods, to sort them quickly, and to find the most efficient way to deliver them. Read the article *How Amazon Uses Robots in Its Warehouses to Improve Shopping, Optimize Stocking and Delivering Products* in [Section 3.11](#). In order to monitor the performance of robots and quantify the benefits of their use, simple descriptive statistics are computed and updated continuously as new data and measurements become available. What happens when a descriptive statistic needs updating?

As a hypothetical example, suppose that the warehouse robots are assigned numerical scores that reflect their ability to lift objects. The greater the score, the greater the weight they can lift. [Figure 3.6](#) shows a schematic depiction of robots with their numerical strength score. The black scores represent the

original scores. Then two of the robots are upgraded and the new strength scores are shown in orange beneath the original scores. Using the definitions in [ReferenceBox 3.3.1](#), the mean strength score before the upgrade is, **5**, the median is **5**, and the mode is also **5**. After the upgrade, the mean is **5.67**, the median and mode remain unchanged at **5**.

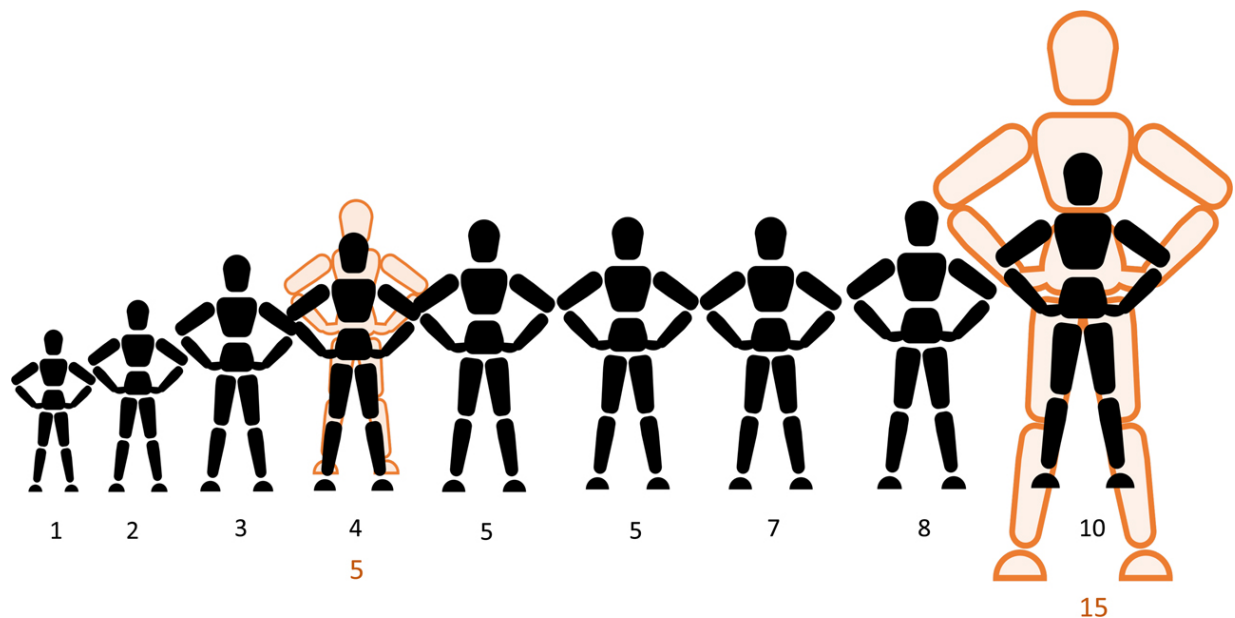


FIGURE 3.6: Example of warehouse robots assigned scores based on their strength, shown in black. Two of the robots are upgraded to newer models (orange). The strength scores of the new robots are shown in orange. Descriptive statistics such as mean and median robot strength will need to be updated after the robot upgrade. Would that be easy? [🔗](#)

What if the new strength score of 15 was mistyped as 150? The mean becomes **20.67**, while the median and mode remain the same. And if the mistyped value is 15000, the mean shoots up to **1670.67** while the median and mode remain unchanged. That is, the mean is highly sensitive to unusually large or small values and can change drastically even if one data value is abnormal or mistyped.

What would it take for the value of the median to change from 5 to 8? That means exactly half the robots should have strength scores of 8 or more. This would require at least 3 robots to be upgraded to have strength scores of 8 or more. Therefore, in order for the median to change, almost half the data values need to change, making it insensitive or *robust* to a few errors or abnormal values.

Both mean and median are easy to understand and interpret. A single unusual value can significantly change the value of the mean, while the median changes only when a significant proportion of data values change. In other words, a mean is more sensitive but less robust to abnormal values as compared to the median. Which one to choose depends on the application. A rigorous introduction to robust statistics can be found in Peter Huber's *Robust Statistical Procedures* [[22](#)].

Data Debate 3.3.4 *Why do we need both the mean and the median? What are the advantages and disadvantages of each? When would you use the mean, and when would you use the median? Give examples.*

3.4 Descriptive Statistics: Data spread

In addition to typical values, the spread or range of data values is an important characteristic. Data spread can be used to identify abnormal values, establish acceptable ranges, and for planning purposes. For example, to prepare for a family reunion, it would be useful to know the age ranges of attendees, e.g., 2 months to 99 years, in order to provide bassinets for babies, high chairs and booster seats for young children, and wheelchairs for the elderly, if necessary. Data spread is used in medical tests such as those that monitor blood pressure or cholesterol, by specifying a range of normal values. Any values outside that

range are flagged as problematic. Such tests compute normal ranges of blood pressure and cholesterol based on observation of millions of patients.

Dispersion is the statistical term for the spread of attribute values of a quantitative variable. The greater the dispersion, the greater the span or spread of values across the number line. We describe below common descriptive statistics used to quantify dispersion of the values of a quantitative variable.

ReferenceBox 3.4.1: Min, Max, and Range

- *Dispersion* refers to the spread of the values of a quantitative variable along the number line, or in attribute space. Attributes whose values are bunched together have small dispersion while attributes whose values span larger parts of the number line have a high dispersion.
- The *maximum* of a quantitative attribute X is its largest observed value.
- The *minimum* of a quantitative attribute X is its smallest observed value.
- The minimum and the maximum together are called *extreme values* since they represent the two extremes of the observed spread of attribute values.
- The *range* of an attribute X is the difference between the maximum and minimum values.

$$\text{Range} = \text{Maximum} - \text{Minimum}.$$

After reviewing the definitions in [ReferenceBox 3.4.1](#), consider the data in [Example 3.1.2](#). The smallest height, known as *minimum* height is **48** inches, the largest height known as the *maximum* height is **70**, and the range which is the difference between the maximum and the minimum is $70-48=\mathbf{22}$.

Data Debate 3.4.1 *Do you think these statistics will change if you choose a different set of 11 middle school students?*

Data Debate 3.4.2 *Do you think the minimum, maximum and range are robust? Are they easy to compute? When should they be used, and when should they be avoided?*

3.4.1 Rank based statistics

Percentiles are popular descriptive statistics that measure dispersion. They are used for ranking an individual relative to their peers, e.g., in pediatric growth charts, educational testing such as SAT and ACT, athletic performances, and measuring health of individuals. Percentiles are particularly useful for data sets that contain millions of data points, such as the set of students who take standardized tests, roughly 1.4 million each year. Students in the same percentile have similar scores, while students in higher percentiles have higher scores than those below. For example, a student with a composite ACT score of 24 would be in the 80th percentile of test takers, i.e., the student's ACT score is higher than the scores of 80% of students who take the ACT.

Percentiles are computed by arranging the data in rank order, e.g., smallest to largest, or vice versa, and dividing the ordered data into 100 intervals, *each* of which contains 1% of the data values. The values at the boundaries of these intervals are the percentiles. That is, the interval between two consecutive percentiles will contain one percent $(1/100)^{th}$ of the ordered data.

The notion of percentiles can be generalized to any fraction, e.g., deciles that divide the ranked data into tenths $(1/10)^{th}$, quartiles that divide the ranked data into quarters $(1/4)^{th}$, and so on. The term *quantile* refers to values that divide the ranked data into specified equal portions.

ReferenceBox 3.4.2: Quantiles: Percentiles and Quartiles [!\[\]\(cbe2492b119e39e02a1dab2af4a4b296_img.jpg\)](#)

- A *percentile* of a set of quantitative attribute values is a number which indicates the percentage of values that are less than that value. For example, if you are in the 61st percentile of a group of runners, then 61% of the runners are slower than you. Roughly 1% of the attribute values lie in the interval between two adjacent percentiles.
- *Quartiles* divide ordered attribute values into quarters. For example, the first quartile, denoted by Q_1 , is greater than 25% of the ordered values but less than 75% of values. The third quartile, denoted by Q_3 , is greater than 75% of the data but less than 25% of data. The median is the second quartile, denoted by Q_2 , and denotes the attribute value that divides the ordered data in two such that each contains 50% of the ordered values.
- In general, a set of *q*-quantiles divide the set of rank ordered attribute values into q intervals that contain equal $(1/q)$ proportion of the data.
- A common measure of dispersion based on quantiles is the *Interquartile Range (IQR)*.

$$IQR = Q_3 - Q_1.$$

Review the definitions in [ReferenceBox 3.4.2](#) and verify that for [Example 3.1.2](#), the first quartile of student heights is $Q_1=51$, the second quartile (median) is $Q_2=55$, and the third quartile $Q_3=58.5$. The interquartile range is $IQR = Q_3 - Q_1 = 58.5 - 51 = 7.5$.

Dispersion can be measured based on the distances between successive quantiles. If the data is contained in a compact interval, the quantiles will be close together on the number line and the distances between adjacent quantiles will be small. Conversely, if the data is dispersed widely across the number line, the intervals between consecutive quantiles will be larger.

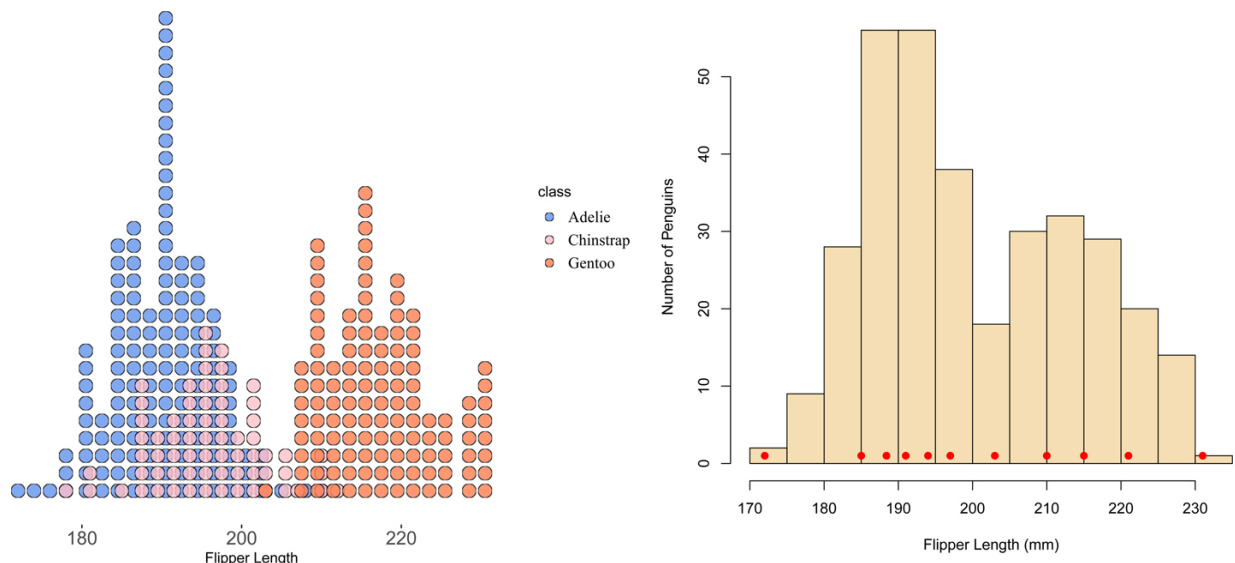
Furthermore, if the distances between successive quantiles are roughly the same, then the data is evenly spread. If not, the data is concentrated more in regions where the quantiles are close together, and it is spread out where the adjacent quantiles are farther apart. This intuition underlies the concept of probability distributions discussed in [Chapter 5](#).

Data Debate 3.4.3 *What type of a percentile is the median? What type of percentiles are the minimum and maximum values? Would it make sense to compute percentiles when you have 30 data values?*

Consider the dispersion of data in [Figure 3.7](#). The charts are discussed in detail in [Section 3.8](#) with respect to visualization charts *dotplots* and *histograms*. [Figure 3.7](#), left, depicts the dispersion of the quantitative attribute flipper length from the Penguin [Example 3.1.1](#). In [Figure 3.7](#), left, each dot represents a penguin, and each species is represented by a different color. Penguins with similar values are shown as a stack of dots. [Figure 3.7](#), right, shows the dispersion of flipper length as a histogram, where the height of the bar measured along the Y -axis corresponds to the number of penguins whose flipper length falls in the interval on the X -axis corresponding to the bar. Deciles are represented as red dots above the x -axis. We notice the following:

- Adelie and Gentoo penguins have similar dispersion of values of flipper length, but in general the flipper lengths of Gentoo penguins are much higher. Adelie and Chinstrap penguins have similar flipper lengths, but the dispersion in flipper lengths of Adelie penguins is higher.
- The collective shape of the dotplots on the left, i.e., the way the values are stacked, is similar to the histogram on the right, with two peaks. The two plots represent the same information but the dotplot breaks down the distribution of flipper lengths by species.

- The distances between adjacent deciles (red dots) and the height of the bars of the histogram are inversely related, i.e., the closer the decile dots, the higher the bars. Why?



[Extended Description](#)

FIGURE 3.7: Data spread for the quantitative variable flipper length, Palmer Penguin data. Left: A dotplot where each dot represents a penguin's flipper length measured on the X -axis. Different colors represent different species. Stacked dots represent penguins with the same or similar flipper length. Dotplots are discussed in [Section 3.8.5](#). Right: A histogram, discussed in [Section 3.8.3](#), of flipper length, where the red dots are deciles. The figures were drawn using the programming language *R*. [↗](#)

Data Debate 3.4.4 *What does the uneven spacing of the deciles mean in the context of dispersion of values of flipper length?*

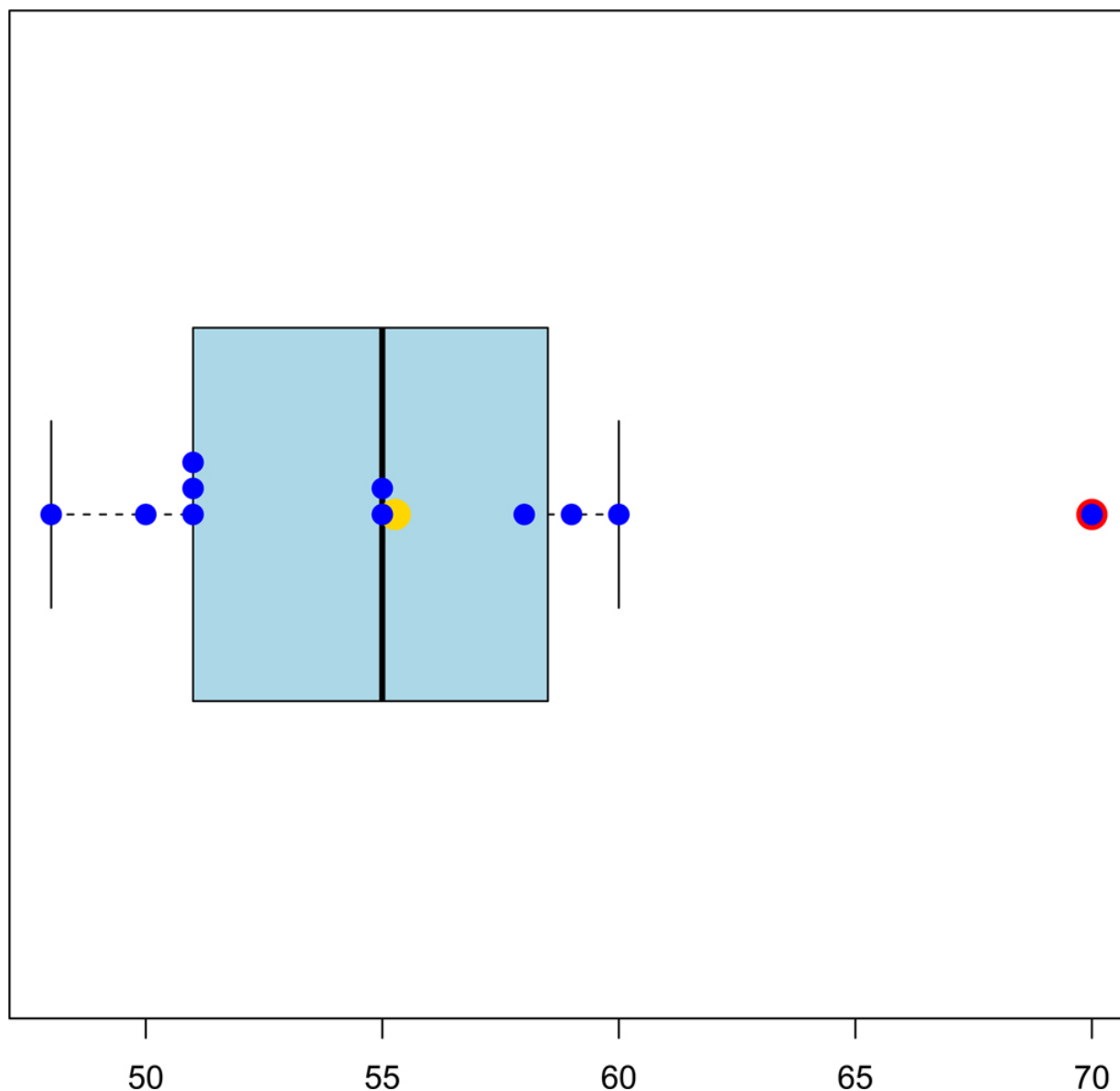
The plots in [Figure 3.7](#) were created in *R*. The code is available [🔗](#) in Supplementary Resources, [Chapter 3](#).

3.4.2 Deviation based statistics

Ranking attribute values in order to quantify dispersion is difficult when the number of values is large. Instead, it is easier to measure dispersion with respect to a reference point. The distance of a value of a quantitative variable from a reference point is called a *deviation*, e.g., how much does a value deviate from a typical value? The greater the deviation of values, the greater dispersion in the data.

[Figure 3.8](#) shows a bar plot (discussed in [Section 3.8.4](#)) of the student heights from [Example 3.1.2](#). Blue dots represent the individual data points. The yellow dot represents the mean. The distances of the blue dots from the yellow dot (mean) along the X -axis represent deviations from the mean.

Boxplot of Student Heights with Dotplot Overlay



[Extended Description](#)

FIGURE 3.8: A horizontal bar plot of student heights where the dark line inside the box is the median, and the left and right edges of the blue box are $Q1$ and $Q3$, respectively. The two small vertical lines connected to the blue box with dotted lines represent the whiskers. Any values outside the whiskers, such as the dot circled in red at the extreme right, represent outliers, or extreme values. bar plots are discussed in [Section 3.8.4](#). An overlaid dotplot

shows the 11 heights of individual students. The yellow dot is the mean. The distances of blue dots from the yellow dot (mean) represent deviations from the mean. The figure was drawn using the *R* programming language. [↩](#)

Variance, σ^2 , is a common measure of dispersion computed with respect to deviations from the mean. It is the average of the squared deviations from the mean. The formula is shown in [ReferenceBox 3.4.3](#). The square root of the variance is the *standard deviation*, σ .

In [Example 3.1.2](#), the 11 deviations from the mean value of 55.27, corresponding to the 11 values of height are given by:

$$\{-7.27, -4.27, -4.27, -4.27, -5.27, -0.27, 3.72, 4.72, 14.72, -0.27, 2.72.\}$$

The variance, computed using the formula in [ReferenceBox 3.4.3](#) is $\sigma^2 = 39.62$, and its square root, the standard deviation $\sigma=6.29$.

The *median absolute deviation (MAD)* uses the median as a reference point. It defines dispersion as the median of the absolute values of the deviations from the median. For the data in [Example 3.1.2](#), the deviations from the median value of 55 are given by:

$$\{-7, -4, -4, -4, -5, 0, 4, 5, 15, 0, 3\}$$

and the absolute deviations from the median are:

$$\{7, 4, 4, 4, 5, 0, 4, 5, 15, 0, 3\},$$

and the ordered absolute deviations from the median are:

$$\{0, 0, 3, 4, 4, 4, 4, 5, 5, 7, 15\}.$$

The median of the absolute values of the deviations from the median is **4**.

ReferenceBox 3.4.3: Deviations: Variance and Median Absolute Deviation

- The *deviation* of a point from the mean is the difference between the mean and that value.
- The *variance* is given by the mean of the squares of the deviation from the mean. For a set $X = x_1, x_2, \dots, x_n$ of n values, the variance σ^2 is given by:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

The *standard deviation* σ is the square root of the variance.

- The *median absolute deviation (MAD)* is the median of the absolute values of the deviation of individual values from the median.

$$MAD = \underset{1, \dots, n}{\text{median}}(|x_i - \text{median}(X)|)$$

Review the definitions in [ReferenceBox 3.4.3](#), do the computations, and verify that your answers match the ones shown above.

Data Debate 3.4.5 *Compare the values of the standard deviation and MAD in the above example. Are they the same? Why or why not?*

Most of the statistics discussed in this section that measure dispersion are easy to compute and understand. Variance is less robust than MAD, analogous to the mean and the median. Percentiles and other quantiles are useful when there is a sufficient amount of data, e.g., more than 500 values. However, it is harder to rank order large amounts of data. The correct choice of a statistic

depends on the application, and the nature of data. The first five chapters of Ron Pearson's *Mining Imperfect Data* [32] provide an intuitive and hands-on introduction to descriptive statistics and EDA.

Data Debate 3.4.6 *Can you think of additional ways of quantifying typical values and data dispersion? Are they robust? Are they easy to compute? Are they easy to understand?*

3.5 Descriptive Statistics: Shape of the data spread

In addition to the amount of spread along the X -axis, the manner in which the data is distributed between the extremes of the minimum and maximum values is important as well. Is it all piled up in one place or is it spread out evenly? Consider [Figure 3.9](#), generated by Gemini AI. The horizontal X -axis represents the values of a variable and the vertical Y -axis represents the proportion of data that value accounts for. This is a high level, approximate interpretation of the figure, we will learn more about probability distributions in [Chapter 5](#).

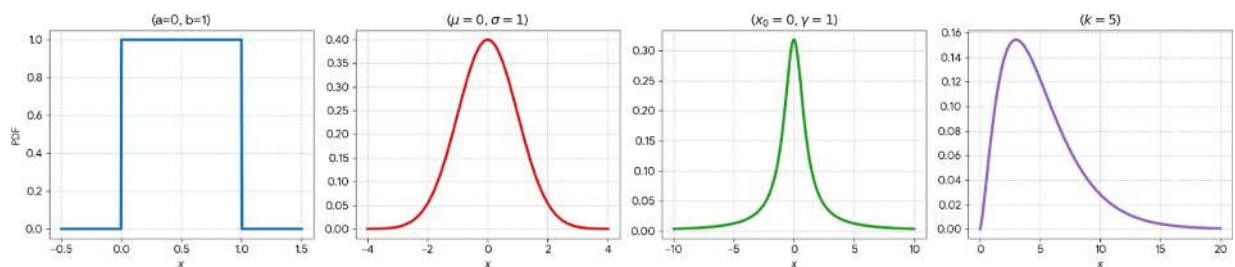


FIGURE 3.9: This figure was generated using Gemini AI. It shows four data spreads, each with a different shape depending on how the data is distributed along the X -axis. Left to right: Evenly spread (Uniform), bell-shaped and symmetric (Normal) that follows the 68-95-99.7% rule, more data away from the typical value but still symmetric (Cauchy), asymmetric (χ^2). [📄](#)

The first graph on the left shows an even distribution of values. For example, if you hit a buzzer that pops up a letter of the alphabet, A through Z, 26,000 times, then each of the 26 values, {A, B, ..., Z } will occur roughly 1000 times. A plot of counts of each of the alphabet values will look like a table top, called a *Uniform* distribution.

The second graph from the left is a *Normal* distribution, where typical values occur more frequently than others but follow a particular set of rules so that 68% of the data falls in an interval on the X -axis within one standard deviation of the mean, 95% within two standard deviations of the mean, and 99.7% within three standard deviations of the mean. This is often referred to as a bell-curve. The particular shape and spread of the data that follow the 68-95-99.7% rule for dispersion of data are special because many natural phenomena based on averages mimic this shape. We will learn about the Normal distribution, also known as Gaussian, in [Chapter 5](#).

The third graph is also symmetric but has more data concentrated away from the center of the distribution relative to the Normal distribution. It is said to be “heavy-tailed” since it has more data piled away from the center. Such distributions do not follow the rule that a Normal distribution does. In fact, some such distributions do not even have a well-defined standard deviation.

The last graph is an asymmetric dispersion of data, with a large proportion of data piled up at the smaller values on the X -axis and tapering away to a lesser proportion of data as the data values get larger. Many income distributions demonstrate this shape.

ReferenceBox 3.5.1: Symmetry and Shape: Skewness and Kurtosis [↗](#)

- *Skewness* of a set of attribute values measures symmetry in the data spread relative to the Normal distribution. There are multiple ways of measuring skewness as discussed in this Wikipedia [article](#), but the most intuitive one is Pearson 2 skewness, given by:

$$\gamma = \frac{3 * (\bar{X} - Median)}{s},$$

where s is the data standard deviation. The Normal distribution has 0 skewness since the mean and median are the same.

- *Kurtosis* refers to the presence of outliers in the data relative to the Normal distribution. It is given by:

$$\kappa = \left\{ \frac{\sum_{i=1}^n (x_i - \bar{X})^4}{n} \right\} / s^4$$

where s is the data standard deviation. In general, larger the value of κ , greater the proportion of non-typical values in the data.

Skewness, defined in [ReferenceBox 3.5.1](#), denoted by γ , represents *the asymmetry in the spread of data values*. It measures the deviation of the median from the mean relative to the dispersion measured by the standard deviation. The Normal distribution has 0 skewness, since it is symmetric and its mean is always equal to its median. A positive value of skewness means that the mean of the data values is greater than the median, implying that the data values are spread out farther on the higher side, while a negative skewness implies that the data is spread out farther on the lower side. It takes just one value to change the skewness, since it is based on distance rather than the proportion of values.

Kurtosis, defined in [ReferenceBox 3.5.1](#) and denoted by κ , quantifies the proportion and extent of outliers in the data. The value of κ is 3 for the Normal distribution. If the kurtosis of a set of data values is less than 3, then the data has fewer non-typical values, and a value of kurtosis greater than 3

implies a greater preponderance of non-typical values relative to the Normal. Many software tools report the *excess kurtosis*, i.e., $\kappa - 3$, to make it easier to compare to the Normal distribution whose excess kurtosis is 0.

Computing skewness and kurtosis from data, especially small data sets, is tricky. The formulas need adjustments to keep the statistics representative of the population values of skewness and kurtosis. The adjustment is based on the *asymptotic* behavior of the sample statistics as the sample size gets larger. It is best to compute skewness and kurtosis using software tools. For a detailed explanation, read the description in the *Engineering Statistics Handbook* provided by the *National Institute of Standards and Technology (NIST)*. 

Data Debate 3.5.1 *We have used the package “moments” in R to compute the following values of skewness and kurtosis.*

The sample skewness for the data in [Example 3.1.2](#) is $\gamma=1.096311$. Is the data in [Example 3.1.2](#) skewed? If so, in which direction?

The kurtosis is $\kappa=3.669507$, with an excess kurtosis of 0.669. Does it have a propensity to outliers relative to the Normal distribution? Use Gemini or any other software tool to compute γ and κ for the 11 student heights and compare with the results above.

3.6 Descriptive Statistics: Qualitative variables

Descriptive summaries used to represent quantitative variables cannot be used to describe qualitative variables. For example, what is the mean of the attribute “color” that can take values “red”, “blue” or “green”? Typically, qualitative variables are described using *counts* and *proportions*.

For instance, consider the penguin data set in [Example 3.1.1](#). There are three qualitative variables: species, island and sex. The variable sex can either

be “male” or “female”. Of the 344 data records, only 333 have legitimate values for the variable sex. Of these 333, there are 165 females and 168 males. This information is summarized in [Table 3.2](#).

TABLE 3.2: Descriptive statistics for a qualitative variable sex (Palmer penguins data): counts, proportions and percentages. [↗](#)

Sex	Count	Proportion	Percentage
Female	165	0.495496	49.55%
Male	168	0.504504	50.45%
Total	333	1.0	100%

Counts and proportions are necessary for visualizing qualitative variables, e.g., bar charts and pie charts, as discussed in [Section 3.8](#).

ReferenceBox 3.6.1: Counts, Proportions and Percentages

- *Counts* represent the frequency of each value of a qualitative variable in the data. For example, according to [Table 3.2](#), the qualitative variable “sex” was recorded as “female” for 165 penguins and “male” for 168 penguins in the Palmer penguin data described in [Example 3.1.1](#).
- *Proportions* represent the fraction of objects that have a particular value of a qualitative variable. For example, the proportion of female penguins is 0.495496.

$$Proportion = \frac{Count}{Total}.$$

- *Percentages* are derived from proportions by multiplying them by 100.

$$\text{Percentage} = \text{Proportion} * 100.$$

3.7 EDA of Palmer Penguins Data with Descriptive Statistics

Consider the penguins data set discussed in [Example 3.1.1](#) to practice EDA and visualization skills. Here is a link to a downloadable version of the Penguin data set: [🔗](#). AI tools can provide analytics in an instant, but the best way to develop data intuition is to conduct EDA in a step-by-step manner.

While any software package like *R* or Python can be used, Excel or Google Sheets do not require a knowledge of computer programming languages. Data sets in CSV format can be downloaded directly to these spreadsheets. Google Sheets has an extension called *XLMiner Analysis ToolPak* for EDA. Here is a link to the instructions on how to perform EDA using Google Sheets: [🔗](#)

Microsoft Excel provides a *Data Analysis Tool Kit* that can be downloaded to an Excel spreadsheet to find descriptive statistics for a variable. Here is the link to the step-by-step instructions: [🔗](#)

Example 3.7.1 *Here is a step-by-step guide for computing descriptive statistics for the body mass of penguins on Biscoe Island using Google Sheets.*

- *Select column A and choose “Data” from the menu bar, sort A-Z.*
- *Copy and paste the data for each island on a separate sheet.*
- *For Biscoe Island (C1:C168 rows), compute the descriptive statistics for body mass by ticking the “Summary Statistics” box.*

The screen shot of the dialog box for computing the descriptive statistics and the resulting output are shown in [Figure 3.10](#).

▼ Descriptive Statistics

Input Range:

Grouped By: ☒ Columns ☐ Rows

☒ Labels in First Row

Output Range:

☒ Summary Statistics

☐ Confidence Level for Mean: %

☐ Kth Largest:

☐ Kth Smallest:

Bodymass in gms	
Mean	4716.017964
Standard Error	60.57919615
Median	4775
Mode	5550
Standard Deviation	782.8557429
Sample Variance	612863.1141
Kurtosis	-0.4589561965
Skewness	-0.4419526446
Range	3450
Minimum	2850
Maximum	6300
Sum	787575
Count	167

FIGURE 3.10: Left: Screen shot of dialog box for computing statistics. Right: Screen shot of output: descriptive statistics for body mass of penguins from Biscoe Island. [↗](#)

Data Debate 3.7.1 Replicate the computations in [Figure 3.10](#). Do your results match those in the figure? What can you say about the shape of the data distribution of the body mass of penguins on Biscoe Island?

3.7.1 Data quality issues

Data collection, storage and retrieval processes typically involve many stages, and problems can arise at any of these stages, impacting the quality of data. Data may be missing, mislabeled, or incorrectly recorded due to human error or software glitches. For example, a response of “30 hours” to an online survey regarding *daily* hours spent on social media is obviously incorrect. The first task in EDA is to check for quality issues. Do you see any missing values in the Penguin data set?

Data Debate 3.7.2 *What could happen if we try to create charts, or perform data analysis with missing or incomplete data? Should we discard the missing data? What if there is too much missing data? What if the data is missing at random? What if the data is missing only for a particular group?*

For the purpose of this analysis, we have dropped missing values from computations. We could have potentially guessed the missing values, using a process called *imputation*. However, changing data values has consequences. In [Chapter 8](#), we will learn more about these topics and the potential for distorting data characteristics while cleaning data.

Next, let us use an Excel spreadsheet and its built-in functions (formulas) that calculate descriptive statistics such as mean, median, mode, range, and standard deviation. The descriptive statistics for bill length of the Palmer penguins are shown in [Figure 3.11](#), from which we can infer that:

- There are more Adelie penguins than other species, e.g., more than twice as many as Chinstrap. This means that the characteristics of the Adelie penguins will dominate any general statistics computed for all the penguins collectively.
- Adelie penguins have smaller bill lengths compared to Chinstrap and Gentoo based on the mean and median values. There is less dispersion in the values of Adelie penguin bill lengths based on range and variance of the bill lengths.
- Gentoo bill length values have a small positive skewness implying that the higher values are farther away on the number line, while the Chinstrap bill length values have a small negative skewness implying that the lower values are farther away making the mean value less than the median.

Descriptive statistics for the three penguin species					
<i>Adelie</i>		<i>Chinstrap</i>		<i>Gentoo</i>	
Mean	38.7913907	Mean	48.8338235	Mean	47.5163934
Median	38.8	Median	49.55	Median	47.3
Mode	41.1	Mode	51.3	Mode	50
Standard Deviation	2.66340485	Standard Deviation	3.3392559	Standard Deviation	3.09190788
Sample Variance	7.09372539	Sample Variance	11.1506299	Sample Variance	9.55989432
Skewness	0.16167431	Skewness	-0.0905746	Skewness	0.63964782
Range	13.9	Range	17.1	Range	18.7
Minimum	32.1	Minimum	40.9	Minimum	40.9
Maximum	46	Maximum	58	Maximum	59.6
Sum	5857.5	Sum	3320.7	Sum	5797
Count	151	Count	68	Count	123

FIGURE 3.11: Descriptive statistics for the bill lengths of three species of penguins in the Palmer Penguins data set, calculated using the data analysis feature of an Excel spreadsheet. [📄](#)

Data Debate 3.7.3 *Are the different descriptive statistics used to represent typical values consistent, i.e., if you were to use mean, median or mode to compare the relative bill lengths of the three species, would you come to the same conclusion in all three cases?*

3.8 Exploring Data through Visualization

As computer hardware and software become more sophisticated, it has become possible to draw brilliant and complex graphs, including with animation and three-dimensional effects like those seen in AI-generated holograms. Every field, from genetics to marketing to astronomy, has become increasingly reliant on graphics to quickly and easily communicate findings. This is because the human brain excels at picking up visual cues and patterns.

The power of visuals is such that AI has recreated the images of dead actors to star in entire films to capture their magic even after their passing. See [Section 3.11](#) for BBC's article *How AI is bringing film stars back from the dead*. While the magic of AI can create holograms, there is a small set of canonical chart types that is essential for EDA. These are commonly used in scientific journals, in mass media, in marketing and financial analytics, and in many disciplines. Such charts are intuitive and easy to understand, some of which we have already explored in [Figures 3.4](#), [3.5](#), [3.7](#) and [3.8](#). We detail below how to create and interpret some popular visualization charts.

3.8.1 Circle (pie) charts

A *pie chart* is used to visualize qualitative variables, and to show the share of each of its category values. The proportion of each category is represented by the percentage of the area of the circular pie. This is one of the simplest graphical representations of data. Here is the link to the step-by-step instructions on how to use the Excel tools to create a pie chart. 

[Table 3.3](#) shows the count of penguins on each of the three islands, and the counts of penguins that belong to each of the three species. [Figure 3.12](#) is a pictorial representation of these counts where the proportion of the area of each slice of the pie to the area of the entire pie is the same as that of the count of that category to the total. We can infer that:

- Almost half the penguins live on Biscoe Island
- Almost half the penguins belong to the Adelie species
- The pie charts do not have any information about which species live on which island

TABLE 3.3: Counts of penguins by species and by island. 

Species	No. of Penguins
Adelie	151
Gentoo	123
Chinstrap	68

Island	No. of Penguins
Biscoe	167
Dream	124
Torgersen	51

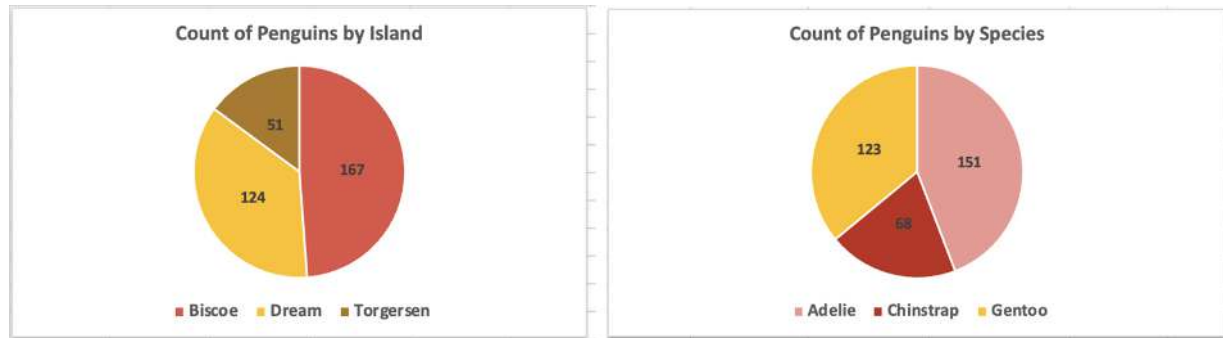


FIGURE 3.12: Pie charts were created using Excel spreadsheets. Left: The distribution of penguins in the three islands in the study. Right: The distribution (share) of each of the three species of penguins in the study. [📄](#)

Data Debate 3.8.1 *What are the limitations of pie charts? What if there were a hundred species of penguins in the Palmer penguin data set rather than just three?*

3.8.2 Bar charts

A *bar graph* or *bar chart* represents the share of each category of a qualitative variable with rectangular bars whose heights or lengths are proportional to the counts of values that they represent. The bars can be plotted vertically or horizontally. The data represented by the bars can be counts, percentages or proportions, or averages of a quantitative variable corresponding to a category. Each data value should belong to only one category, and every

category should be represented by exactly one of the bars. Bar graphs are intended for qualitative or discrete variables.

[Figure 3.13](#), left, shows the same information as the pie chart in [Figure 3.12](#), right, but not as a share of a pie. However, the relative proportion of each category is easier to perceive in a pie chart. Bar charts are arguably better than pie charts when the number of categories of a qualitative variable is larger.

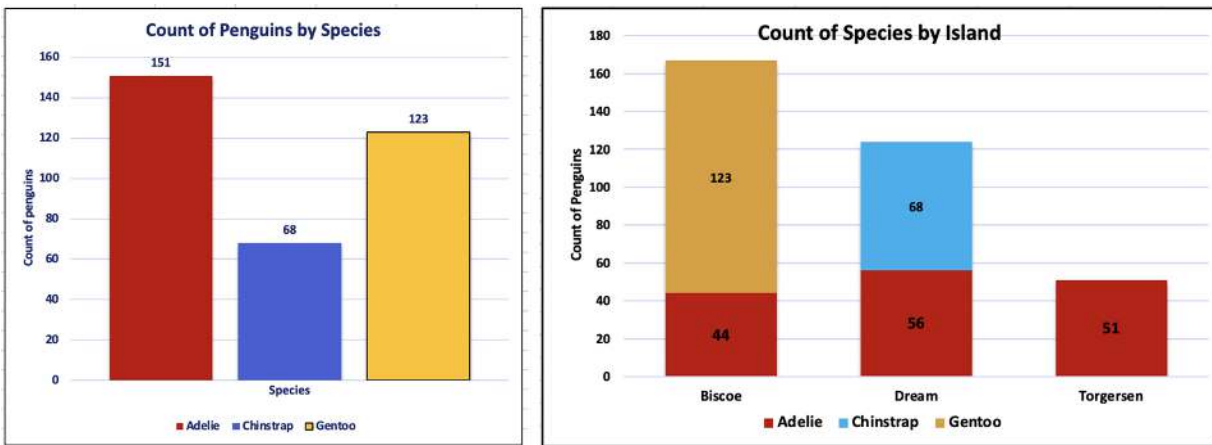


FIGURE 3.13: The bar graph on the left represents the counts of the three species of penguins found on the islands. The figure on the right represents a stacked bar graph of penguin species on each island. These bar graphs were created by using the pivot table in Excel. [↩](#)

Here are step-by-step instructions on how to create bar charts using the chart feature in Excel: [🔗](#). Microsoft provides a tutorial on how to use the COUNTIF function used in the creation of bar charts: [🔗](#).

A *stacked bar graph* or stacked bar chart extends the standard bar chart from one qualitative variable to two or more. Each bar in a bar chart is divided into a number of sub-bars stacked end to end, each one corresponding to a level of the second or third categorical variable. [Figure 3.13](#), right, shows a stacked bar chart of the number of penguins on each island, split further by

species. The advantage of a stacked bar chart is that we can visualize two or more qualitative variables at once. From [Figure 3.13](#), right, it is obvious that Torgersen Island only has Adelie penguins living on it. It would have taken more effort to infer this from merely examining rows and columns of an Excel spreadsheet.

Data Debate 3.8.2 *What can you infer from the stacked bar chart in [Figure 3.13](#)? Did a visual representation of the data help? How many variables do you think we can represent using a conventional bar chart?*

3.8.3 Histograms

A *histogram* is the equivalent of a bar chart for quantitative variables, particularly continuous ones. Data values corresponding to quantitative variables can be grouped together into intervals of values in a meaningful way. For instance, in [Example 3.1.2](#) the student heights can be grouped into ranges such as 45-50 inches, 50-55 inches, 55-60 inches and 60 inches or more.

The process of grouping data into intervals of values is called *binning*. Every data value must belong to one *bin*, and exactly one bin. There are many ways of determining appropriate intervals to create the bins. The easiest approach is to take the range of data values in the data set and divide it by the desired number of bins, and use this to determine the *bin size*, or *bin length*. Such histograms are called *equi-width* since every bin is of the same width.

Another option is to choose the quantiles to define the bin intervals, e.g., quartiles. This approach will ensure that each bin contains roughly the same proportion of data. Such histograms are called *equi-depth* since the height of the bar corresponding to each bin is roughly of the same height (or depth). Once the bins are determined, it is just a matter of counting the number of data values that fall in each bin, called the *bin count* or *bin frequency*.

The book *Multivariate Density Estimation: Theory, Practice, and Visualization* by David W. Scott [38] has an excellent overview of different types of histograms and their relative merits and weaknesses.

Example 3.8.1 *Suppose we wish to create a histogram with four bins using the data in [Example 3.1.2](#). We can compute the following:*

- $\text{range} = \text{max} - \text{min} = 70 - 48 = 22$
- $\text{number of bins} = 4$
- $\text{bin width} = 22/4 = 5.5$
- $\text{bin intervals (bins)} = [48, 53.5), [53.5, 59), [59, 64.5), [64.5, 70]$, where we start with the minimum value, add the bin width to get the upper boundary of the first bin, which is the lower boundary of the next bin, and so on until we equal the maximum value.
- $\text{bin counts (in order)}: 5, 4, 1, 1$. For example, the values 48, 51, 51, 51, and 50 lie in the interval $[48, 53.5]$ and contribute to the count of 5 for that bin.

Note that the bin boundaries overlap, but the bins are non-overlapping since by convention, any value that falls exactly on the upper boundary is not included in that bin, but in the next higher one. The exception to this rule is the maximum value.

[Figure 3.7](#), right, depicts an equi-width histogram based on the flipper length measurements of all penguins. The programming language *R* was used to create the histogram. Each bin is approximately 5 mm in width, hence equi-width. The bin corresponding to the interval of flipper length 205-210 mm has roughly a 30-bin count, implying that around 30 penguins have flipper lengths between 205 and 210 mm. If the bin intervals had been determined using the deciles represented by the red dots, there would be 10 bins, each with 34 or so

penguin count, but each bin would be of a different width as is evident from the distances between consecutive deciles represented by the red dots.

Data Debate 3.8.3 *Is the histogram in [Figure 3.7](#) equi-depth or equi-width? Draw an equi-depth histogram using the red dots as bin boundaries. What do you notice about the height of the bars?*

Histograms can be created using spreadsheet tools such as Google Sheets or Excel. Here is the link to step-by-step instructions on how to use Excel tools to create a histogram. [🔗](#). The plots in [Figure 3.14](#) were created using the chart feature in Google Sheets. The histogram of body mass on the left is not symmetric, with some outliers in the highest bin [6277.78-6500.00]. The histogram of flipper length on the right has two peaks, making it a *bi-modal* distribution. This histogram should resemble the one in [Figure 3.7](#) since it represents the same variable flipper length.

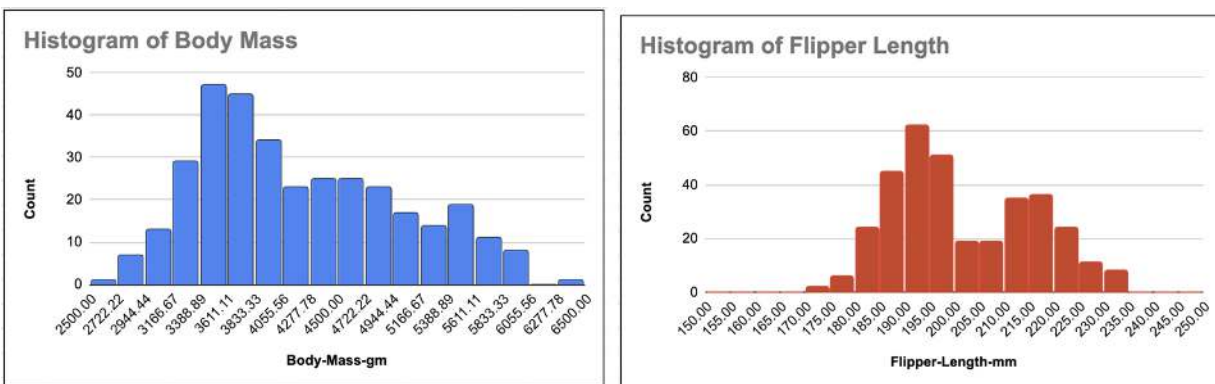


FIGURE 3.14: Histograms of Penguins data created using charts tool in Google Sheets. [🔗](#)

Data Debate 3.8.4 *Why do you think the flipper length histogram has two peaks? Are all the penguins of the same type?*

3.8.4 Boxplots

A *boxplot* or box-and-whiskers plot is used to illustrate the spread of data. It can be drawn horizontally as in [Figure 3.8](#) or vertically as in [Figure 3.16](#).

[Figure 3.15](#), created in Microsoft PowerPoint, shows a schematic diagram of a bar plot. The blue box is parallel to the number line and contains the middle 50% of the data. Since it is a horizontal bar plot, the blue box's left and right edges correspond to the first and third quartiles, respectively. The box is divided inside by a line that represents the median. The whiskers, shown as dotted lines, extend out beyond the first and third quartiles to represent values outside the middle half of the data. The whiskers end at what are considered acceptable limits for data values, called “fences”, denoted by small lines parallel to the quartile lines of the box. The acceptable limits can be determined in multiple ways, e.g., the 5th and 95th percentiles to mark the end of the whiskers. The three values outside the whiskers shown in red are called *outliers*. [Figure 3.8](#) shows an outlier circled in red, corresponding to the height of the tallest middle schooler in [Example 3.1.2](#).

Boxplot

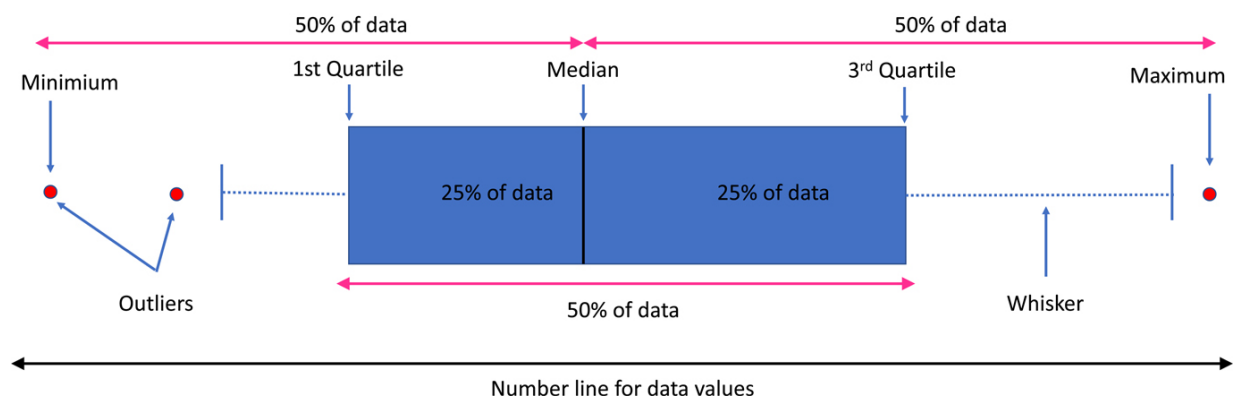


FIGURE 3.15: How to interpret a bar plot. This figure was created in Microsoft Power Point. [↗](#)

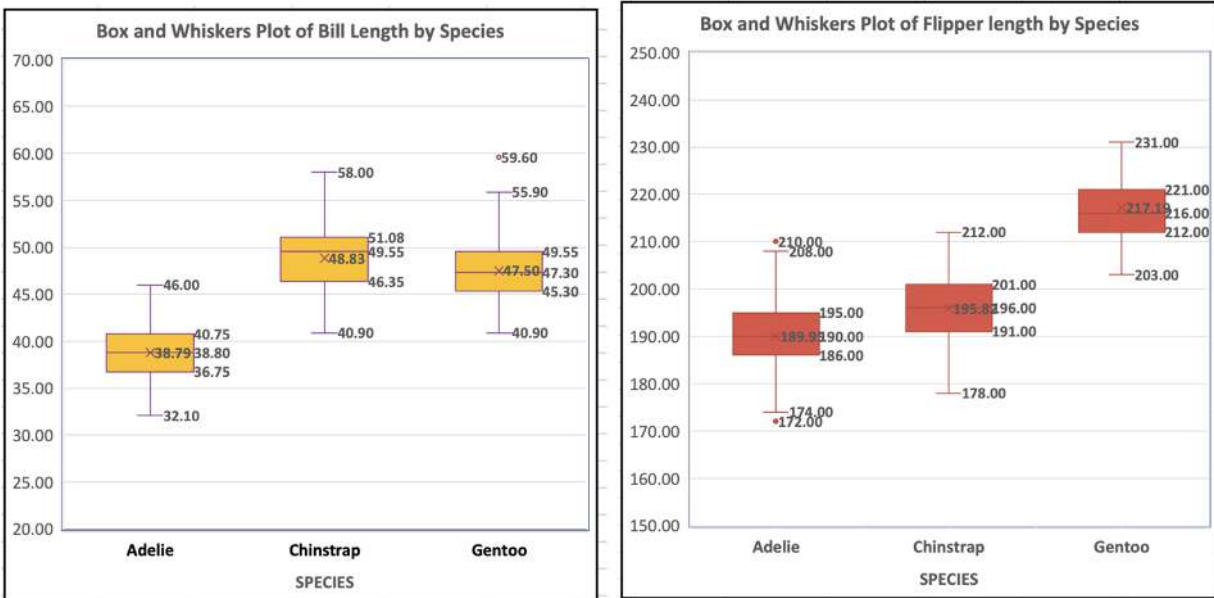


FIGURE 3.16: Box and whiskers plot created using Excel. The left chart displays quartiles for the variable bill length for each of the three species of penguins and the right chart displays the quartiles for flipper length. [📄](#)

The graphs in [Figure 3.16](#), created in Excel, display three box-and-whiskers plots, one for each species: Adelie, Gentoo and Chinstrap. The bar plots on the left correspond to bill length, and the bar plots on the right correspond to flipper length. The mean is marked inside each box with an “X”. The values of the mean, median, quartiles and the end of the whiskers, and outliers are shown. We can make the following observations:

- The mean, $Q1$, $Q2(\text{median})$, $Q3$ of bill length of Gentoo penguins are 47.50, 45.30, 47.30, 49.55, respectively. The ends of the whiskers are 40.90 and 55.90. The bill length of 59.60 is an outlier for Gentoo penguins.
- The flipper length of 210 is an outlier for Adelie penguins but not Chinstrap or Gentoo.
- The mean bill lengths of Chinstrap (48.83) and Gentoo (47.5) are close, but much higher than that of Adelie (38.79).

- The mean and median of flipper length are quite close for all three species indicating that the distributions are not skewed, confirmed by the skewness values in [Figure 3.11](#), which are close to 0.

The lengths of the two whiskers are approximately the same for almost all 6 bar plots, implying, again, a symmetry in the distribution of the values of each of bill length and flipper length for the three species.

The step-by-step process for creating a box-and-whiskers plot using Excel is available in Supplementary Resources , in [Chapter 3](#).

3.8.5 Dotplots and scatterplots

Dotplots are ideal for depicting the spread of a quantitative variable and for comparing multiple groups as shown in [Figure 3.7](#), left. It is clear that Gentoo penguins represented by orange dots have higher values of flipper lengths, compared to the other two species represented by pink and blue dots. A simpler instance of dotplots is in [Figure 3.8](#) where each blue dot corresponds to the height of one of the 11 students in [Example 3.1.2](#). Note the three dots stacked vertically that correspond to the mode at 51. Dotplots are useful for visualizing smaller data sets. When the volume of data is large, dotplots are too cluttered to be legible.

A *scatterplot* is a two-dimensional version of the dotplot. It is used to visualize the relationship between two quantitative variables, as shown in [Figure 3.17](#), with one variable depicted on each axis. Each data point is an ordered tuple of two values, one for each variable, represented as Cartesian (x, y) coordinates.

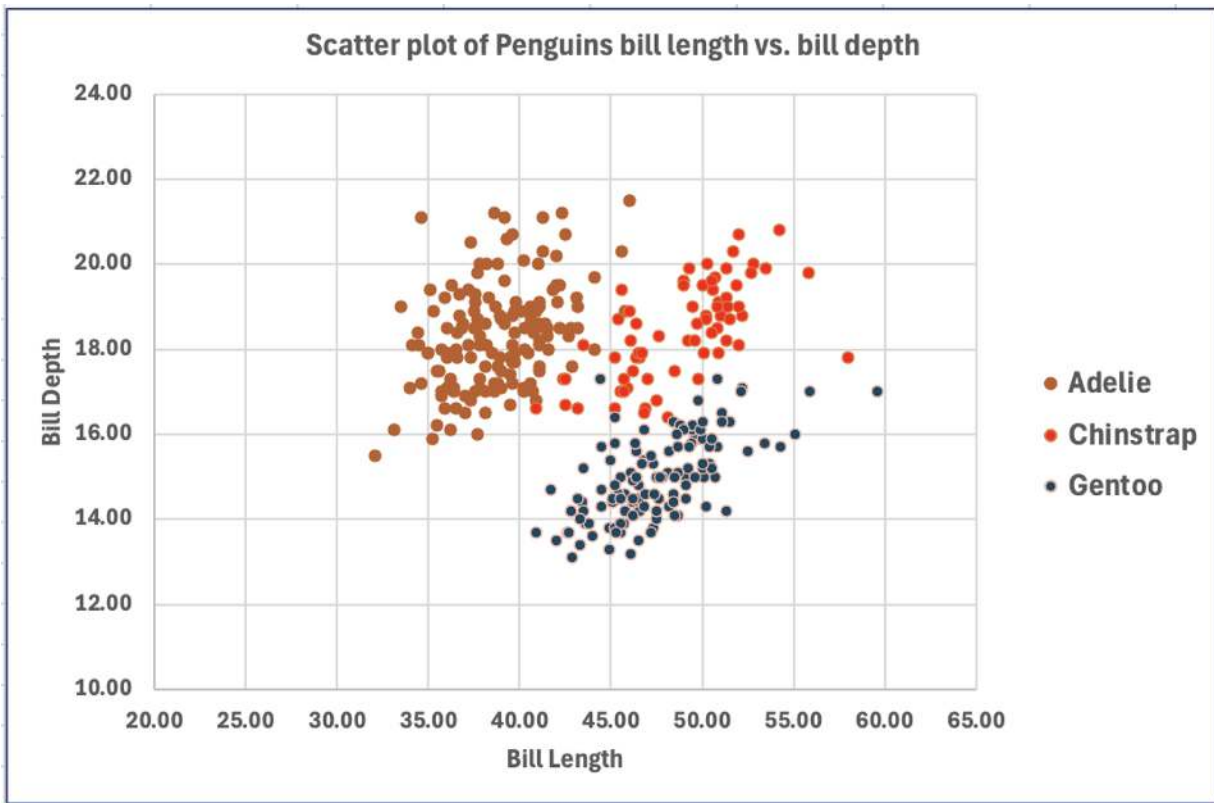


FIGURE 3.17: The scatterplot in [Figure 3.17](#) was created with Excel spreadsheets. It depicts the relationship between bill length and bill depth for the three penguin species. Each species (categorical variable) has a different color as shown in the legend. [↗](#)

In [Figure 3.17](#), the X -axis depicts the bill length and the Y -axis depicts the bill depth of penguins. Each dot corresponds to a penguin's bill length and depth, resulting in 342 dots on this scatterplot. Each color represents a species, as shown in the legend of the plot. Isn't it amazing that one look at the scatterplot reveals that each of the three species occupies distinct regions in the X,Y attribute space, resulting in clear clusters? What do you think that means? Would you have noticed this clustering of species just by looking at a spreadsheet of numbers?

Data Debate 3.8.5 *Is there a relationship between bill length and bill depth for the three species? What are some other facts you can gather by examining the graph?*

The step-by-step process for constructing scatterplots using Excel is explained in this document: 

Data Debate 3.8.6 *When is it inappropriate to use scatterplots? What happens when there are too many data points?*

3.8.6 Time series

A time series is a special kind of plot which shows how a quantitative variable changes over time. The *X*-axis of a time series plot is always time, and the *Y*-axis is the quantitative variable being measured at a given instance of time. Time series charts are widely used to track changes, study *trends*, i.e., a sustained change over time in the same direction (e.g., increasing number of likes on TikTok), and monitor *volatility* which is variability measured over time e.g., stock prices. [Figure 3.4](#) is an example of time series of ozone levels in Ann Arbor, MI. It depicts dangerous spikes in ozone during the summer months.

[Figure 3.18](#), left, depicts data from the United States Census Bureau that reflects the change in college enrollment, measured in thousands of people, from 1947 to 2020, by age group and gender. The red lines represent number of female students, and the blue lines represent male students. Different line types, e.g., solid, dotted, dashed, represent different age groups as marked on the graph. We can make some interesting observations:

- The number of students aged 14-24 enrolled in college increased steeply over time for both men and women up until 2010.

- Women surpassed men in enrollment in the 1980s in the age groups 25-34 and 14-25, but almost a decade earlier for the age group 35+.
- The enrollment declines after 2010 among all age groups and all sexes. It might have been a lagged effect of the financial crisis of 2008.

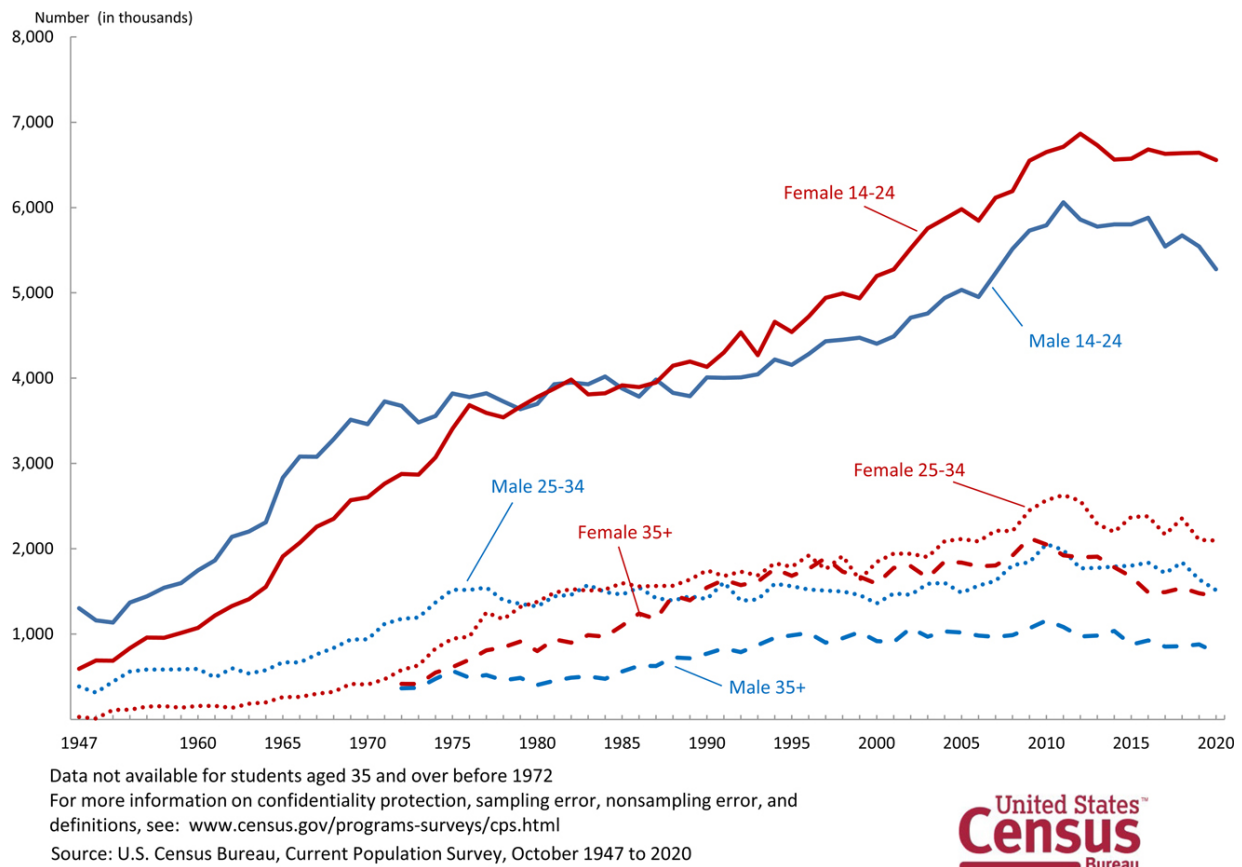


FIGURE 3.18: Time series plots depicts the values of a quantitative variable measured over a period of time, e.g., time series of college enrollment by gender and age group over the period 1947-2020. From the U.S. Census Bureau. [↗](#)

3.8.7 Heatmaps

Heatmaps show the intensity of a quantitative variable for several categories on a colored scale that helps identify regions of high and low intensity. For

example, the map in [Figure 3.5](#), left, shows the changes in average temperatures during the period 2011-2021 across the globe compared to the average during the period 1956-1976. Areas where the average change remained unchanged are represented by white, while those that became warmer are shown in shades of red, and those that became cooler on an average are shown in shades of blue. The deeper the shade of red, the hotter a region became, while the deeper the shade of blue, the cooler a region became. The color gradation is called an *intensity scale* since it corresponds to the magnitude of change.

A heatmap is not limited to depicting changes in temperature. It can be used to visualize the distribution of values of quantitative variable on a graded color scale across the values of a categorical variable, for example, population density across the globe or volume of posting on a social media platform. Certain times of day and certain days of the week have a greater volume of social media posts. Explore [GeminiBox 3.8.1](#) to learn more.

GeminiBox 3.8.1: AI tools employ heatmaps

“Give examples of AI heatmaps.” 

Heatmaps are used extensively by AI tools. For example, click heatmaps are used by companies to identify regions of a webpage that have high user engagement in order to optimize product placement. In the physical world, cameras track shoppers and generate heatmaps to determine which shelves attract most foot traffic for determining placement of products on shelves.

AI heatmaps are also used to track eye movements, to identify the most and least viewed parts of a website as shown in [Figure 3.19](#). Warm red hues denote areas where viewers lingered, while cooler blue hues represent less viewed areas. AI tools automatically generate such analytics. Explore AI heatmaps and their applications using the prompt in [GeminiBox 3.8.1](#)



FIGURE 3.19: Heatmap shows eye movements across the webpage. Image credit: Tschneider, Wikimedia Commons. [🔗](#) [📄](#)

3.8.8 Choice of graph

We have discussed a number of commonly used visualization techniques. There are several articles in [Section 3.11](#) that show how the boundaries of visualization technology are constantly being pushed by technical advances in graphics, rendering hardware, software and AI. Animation, holograms, simulations, and 3D graphics are quite popular.

The choice of chart or graph depends on the type of data and amount of data. Bar charts and pie charts are suitable for qualitative variables, especially when the number of unique categories is limited. In order to reduce the number, of categories, those that are not of interest or occur rarely could be grouped together into a single category called “Other” or “Miscellaneous”.

Dotplots and stacked dotplots are useful for displaying single quantitative variables, grouped by categories if needed. Scatterplots display the relationship between two quantitative variables, potentially grouped by a categorical variable e.g., in [Figure 3.17](#) which shows the scatterplot of bill length vs. bill depth for three species of penguins. Scatterplots too become cluttered and indecipherable if there is too much data.

Line plots display changes in one continuous quantitative variable with respect to another continuous quantitative variable. Time series are special type of line plots where the X -axis is time.

Heatmaps highlight regions of high and low intensity of a quantitative variable with respect to the categories of a qualitative variable. They are visually arresting and convey complex information in an intuitive fashion. They are used extensively by AI tools for monitoring online user engagement, for identifying abnormal cell activity in human organs, and for many other purposes.

The Visual Display of Quantitative Information by Edward Tufte [\[41\]](#) is an excellent introduction to data and information visualization.

3.9 EDA and Big Data

EDA is particularly useful in handling Big Data since it *summarizes* large volumes of data with just a small set of descriptive statistics. The sports industry is an example where Big Data is utilized as well as generated. In sports, organizing and analyzing data helps to find patterns, provide feedback to players and managers, make important decisions, and predict future performance of teams and players. *Moneyball* by Michael Lewis [\[27\]](#) is the true story of data-driven sports management that stunned the world of baseball. Fantasy sports have truly brought Big Data into the spotlight for

sports enthusiasts by supplying extremely detailed sports statistics about players and games in real-time as the sporting events unfold. This has led to an unprecedented participation of audience via fantasy leagues and sports betting. According to the Fantasy Sports and Gaming Association (FSGA) around 62.5 million people played fantasy sports in the USA and Canada in 2022. In 2023, the organization estimates that 19% of all American adults played fantasy sports. Click here  to read more.

In addition to sports, EDA of Big Data figures extensively in social media analytics, real-time marketing and advertising, and studying the spread of diseases. Unfortunately, EDA of Big Data analytics has also led to the use of misinformation to influence important outcomes like elections. In [Chapter 10](#), we will discuss the responsibility of using the power of big data and AI for improving rather than undermining society.

3.10 EDA and AI

A good foundation in EDA to cultivate data intuition is critical for the effective use of AI tools. Consider the following general criteria to take into account while working with large data and AI:

- *Ease of understanding*: Choosing models and methods that are easy to understand, interpret and communicate to others, without sacrificing accuracy or efficiency, whenever possible, is a good practice. Almost everyone intuitively understands the meaning of statistics like mean, median, minimum and maximum. However, results of many AI models like neural networks can be understood only by a very small group of experts. EDA is an effective way to develop intuition for data in the context of the real-world application it represents and to bridge that gap.

- *Computational considerations:* Computers need memory to store data, and processors with enough computing power to perform the necessary computations. In general, the more data, the more the computations needed to compute statistics and run models. EDA can guide the choice of appropriate ways to summarize the data, e.g., histograms, that can reduce the amount and complexity of data supplied to AI models while minimizing the loss of accuracy and precision.
- *Ease of updating:* The flow of data is constant. In order to reflect the most current information, statistics, models and other results need to be updated constantly. Thanks to AI, updating is done automatically whenever we request information through a prompt. However, some statistics and models are easy to update, while others are relatively hard, even in the simple case of mean vs. median as explored by [GeminiBox 3.3.2](#). EDA helps choose appropriate statistics and models. Work through the following simple example for an illustration of updating descriptive statistics.

Example 3.10.1 *Consider the following example. Suppose you buy 7 pairs of jeans whose median price is \$50 and whose average price is \$48. In your next trip to the mall, you buy four more pairs of jeans whose prices are \$20, \$30, \$40 and \$49. What is the new average price of the 11 pairs of jeans you possess? Can you compute the new median price?*

- *Robustness:* Ideally, the results and output of a data task should not be overly influenced by a small number of outliers that have extreme data values. This is especially true when there are data quality issues, e.g., a malfunctioning device can generate faulty 0's which could bias descriptive statistics like the mean. EDA can guide the choice of robust statistics depending on the application. Work through the following simple example to understand the effect of a single abnormal value on statistics.

Example 3.10.2 *Suppose your scores in five attempts in an athletic event, say, long jump measured in feet, are 15, 14, 16, 16, 14. However, due to human error, “16” is entered as “1” in your record. What happens to your average? What happens to your median?*

- *Beware of AI hallucinations:* AI tools can generate faulty output. The article *Why Meta's Large Language Model only survived three days*  explains how AI can write authoritative sounding “scientific research papers” on bogus topics like “Bears in space”. We know from our real-world experience that there are no bears living in space, so this article, however impressive seeming, cannot be true. In these early stages of AI tools, EDA provides the data intuition to spot inaccuracies.

3.11 Read & Reflect

- *How Amazon uses robots in its warehouses to improve shopping, optimize stocking and delivering products.* 
 - *How AI is bringing film stars back from the dead.* 
 - *A visualization series as a part of the New York Times' Learning Network.* 
 - *Visualization gallery generated by R code.* 
 - *The Role of Artificial Intelligence In Transforming Sports Broadcasting.* 
 - *Sportico's Charts of the Year: Best Data Visualization of 2023.* 
-

3.12 Exercises

1. Classification of variables

What are the objects in the Penguin data set? How many objects are there? How many variables are there? What are the names of the variables? Classify the variables as qualitative and quantitative.

2. Analyzing graphs from NASA

(a) [Figure 3.5](#) shows daily average temperature changes during the period 2011-2021 compared to the period 1956-1976. What do you conclude about the climate from NASA's [Figure 3.5](#), left?

(b) NASA's webpage GISS Surface Temperature Analysis (GISTEMP v4)  displays graphs pertaining to global temperatures. Examine the different graphs and list some observations that you gather by studying these graphs.

3. Descriptive statistics: Palmer Penguin data set

- (a) Open the penguin data set and copy it onto Google Sheets.
- (b) Sort the data by species and use the extension XLMiner to find the descriptive statistics of flipper length for each of the three species of penguins. Organize the statistics in a table similar to the one in [Figure 3.10](#). Here is a link to a video that demonstrates how to download the XLMiner Analysis Toolpak for Google Sheets: 
- (c) Compare the mean, median and range of flipper length of the three species.
- (d) What is the mode?
- (e) What information can you gather by studying these statistics?

4. Descriptive statistics and data spread: Serena Williams' aces

Use the definitions and formulas in [Section 3.4](#) to calculate statistics for the data set given below, sourced from kaggle  . Serena Williams' number of aces per year (2010-2020):

{248, 200, 484, 480, 452, 498, 324, 71, 170, 206, 178}.

- (a) Calculate the mean, median, mode and range of number of aces.
- (b) Discuss the range of this data set. Are there any outliers?
- (c) In which year did Serena score the most number of aces? The least number of aces?
- (d) Create a bar chart of number of aces per year.
- (e) What insight does this bar chart provide about Serena's performance over the years?
- (f) Which statistic is a good measure of Serena's ability to hit aces?

5. **Deviations based statistics: Variance, standard deviation, mean absolute deviation and median absolute deviation**

In this exercise, you will select a small data set to build data intuition. To understand the spread of data, we can use several statistics such as variance, standard deviation, mean absolute deviation and median absolute deviation. For this exercise, use a calculator to calculate descriptive statistics and answer the questions.

- (a) From the Penguin data set, select a sample of 10 Adelie penguins from the island Biscoe and copy it to a Google sheet. Use this data for all calculations.
- (b) Find the mean, median, and mode for bill length.
- (c) Find the range, minimum and maximum for bill length.
- (d) Find the variance and standard deviation for bill length.

- (e) What is the median absolute deviation for bill length?
- (f) What is the mean absolute deviation for bill length?
- (g) Repeat steps 1-5 for flipper length.
- (h) Repeat steps 1-5 for body mass.

6. Deviations based statistics; Cereals data set

In this exercise, you will practice how to calculate Variance, Standard Deviation, Mean Absolute Deviation, and Median Absolute Deviation using a tool such as Excel spreadsheet or Google sheet. The data we will use is a sample data set with the nutritional information about popular breakfast cereals. Here is the link to the data: 

- (a) Find the mean, median and mode for the calories.
- (b) Find the variance and standard deviation for calories.
- (c) What is the range of calories for this sample?
- (d) Calculate the mean and median carbs. What is the maximum amount of carbs in cereals?

7. Analyzing charts and graphs in the article *Is College Worth it?*

In this report by the Pew Research Center, economic trends for young adults and how young people view the importance of college education are analyzed. To read more about this trend, click this link: 

- (a) Read the report and examine the data represented in the graphs.
- (b) Explain the data presented in the pie chart.
- (c) According to this report, is a college degree helpful in obtaining a well-paying job?
- (d) Examine the graphs about the trend about young men without college degree and their earning. What do these graphs tell us?

8. Analyzing histograms

Examine the two histograms in [Figure 3.14](#) and answer the following questions for body mass and for flipper length:

- (a) Identify the independent variable, and dependent variable for each graph.
- (b) List the frequency (bin size) of each bin for both histograms.
- (c) How many penguins have a body mass less than 4500 gm?
- (d) Approximately how many penguins have a flipper length of at least 200 mm?
- (e) Which bin has the highest count of penguins (for body mass and flipper length)?

9. Analyzing graphs

Look at the graph created by the New York Times Learning Network here [🔗](#). What do you notice about New York City's temperatures? What about New Orleans?

10. Analyzing boxplots

Calculate Interquartile Range using the boxplot in [Figure 3.8](#), right. The red dot represents an outlier. What is the height of the student who corresponds to this outlier?

11. Comparison of descriptive statistics for penguin species

Examine [Figure 3.11](#) and answer the following:

- (a) What do you notice about bill lengths of the three species?
- (b) What is the median for each species?
- (c) Which species has an outlier? Why is this called an outlier?
- (d) What is the mean bill-length for each species?

(e) Which species has the smallest bill length?

12. Constructing scatterplots using Excel

Create a scatterplot of flipper length by body mass for the Gentoo Penguins. Follow the instructions in [Scatterplot link](#) to create the chart. After completing the scatterplot for flipper length and body mass, answer the following questions.

- (a) What do you notice about this scatterplot? Can you identify any relationship between flipper length and body mass?
- (b) How is it different from bill length by bill depth scatterplot?
- (c) Do you notice any outliers? What do these points represent?
- (d) What happens if you swap the X and Y variables?

13. Time series graphs

Find examples of time series graphs in your social media or news feed, with an example of trend and of volatility.

14. Analyzing a time series graph

In this exercise you will examine a graph about the rise in sea level over time and answer questions related to the graph in [Figure 3.20](#). You can read more about this graph in this report from the EPA .

- (a) What do you notice about the chart?
- (b) What is the independent variable? What is the scale?
- (c) What is the dependent variable? Explain.
- (d) When did satellite measurements of sea level begin?
- (e) What does the shaded area in the graph represent?
- (f) How was sea level measured before satellite imaging became possible? When did satellite measurements of sea level begin?

- (g) What is the rate of change from year 1900 to year 2000? What is the rate of change from year 2000 to year 2020? What do you infer from this?

15. Analyzing a heatmap

In this exercise, you will find a population density map of the world from Wikimedia commons . Examine and study the graph in [Figure 3.21](#) to answer questions.

- (a) Examine the map and discuss three features of the map.
- (b) Explain what you can understand by looking at this graph.
- (c) What is the scale used in this heatmap?
- (d) What is the meaning of different colors?
- (e) Name three countries with population density 10-25 people/sq. km.
- (f) Name three countries with population over 400 people/sq. km.

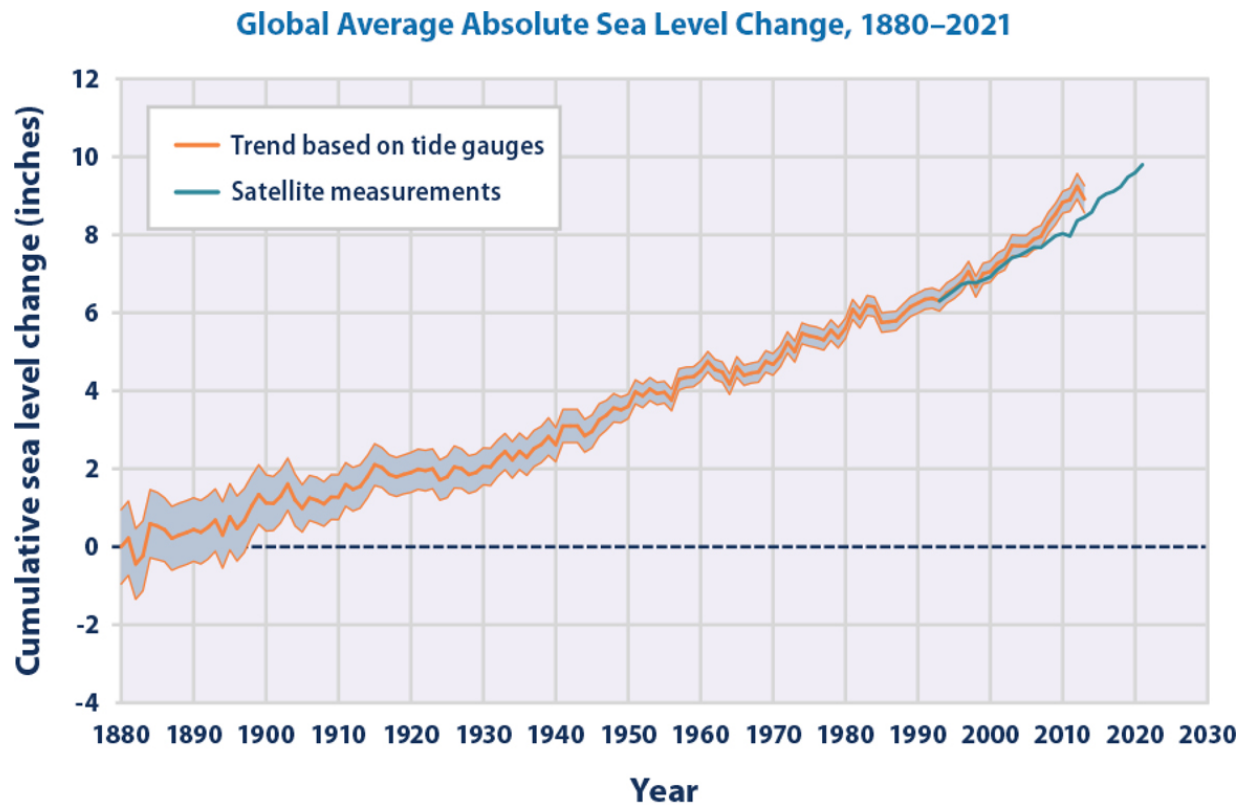
16. Big data and sports

Big data is ubiquitous in sports due to the widespread use of information technology, social media, the internet and availability of data for all kinds of sports. The sports data collection is so large and the data so complex that managing, storing, and analyzing this data is challenging. The analysis of data in sports has a big impact on the decisions made by players and teams.

In this exercise, you will read about different types of data related to sports, and choose a sport that you like. You can find data related to all kinds of sports on this website . Answer the following questions with as much detail as possible.

- (a) How has social media impacted sports data?

- (b) How would you classify information about an athlete in social media: reliable? valuable? accurate?
- (c) Choose a sport that you like or follow. What are some statistics that are collected for this sport? Explain and classify the statistics, and how it is calculated.
- (d) For the sport that you have selected, which statistic has the most impact? Why?
- (e) Choose a player or athlete from any sport. What kind of data is available for this athlete?
- (f) Find a data chart or data visualization for the sport that you have chosen. Describe and explain in detail what is displayed in this chart.
- (g) How does player performance data impact their compensation in team sports?
- (h) What are fantasy sports? Choose a fantasy sport that you are interested in and explain what statistics are used to select players for this fantasy sport.
- (i) What is the role of data visualization in sports?



Data sources:

- CSIRO (Commonwealth Scientific and Industrial Research Organisation). 2017 update to data originally published in: Church, J.A., and N.J. White. 2011. Sea-level rise from the late 19th to the early 21st century. *Surv. Geophys.* 32:585–602. Accessed September 2017. www.cmar.csiro.au/sealevel/sl_data_cmar.html.
- NOAA (National Oceanic and Atmospheric Administration). 2022. Laboratory for Satellite Altimetry: Sea level rise. Accessed March 2022. www.star.nesdis.noaa.gov/sod/lisa/SeaLevelRise/LSA_SLR_timeseries_global.php.

For more information, visit U.S. EPA's "Climate Change Indicators in the United States" at www.epa.gov/climate-indicators.

FIGURE 3.20: Time Series Graph-Data source: Climate Change Indicators: Sea Level. Absolute Sea Level Change. EPA.gov. U.S. Environmental Protection Agency (EPA) (July 2022). Archived from the original on 4 September 2023. [↗](#)

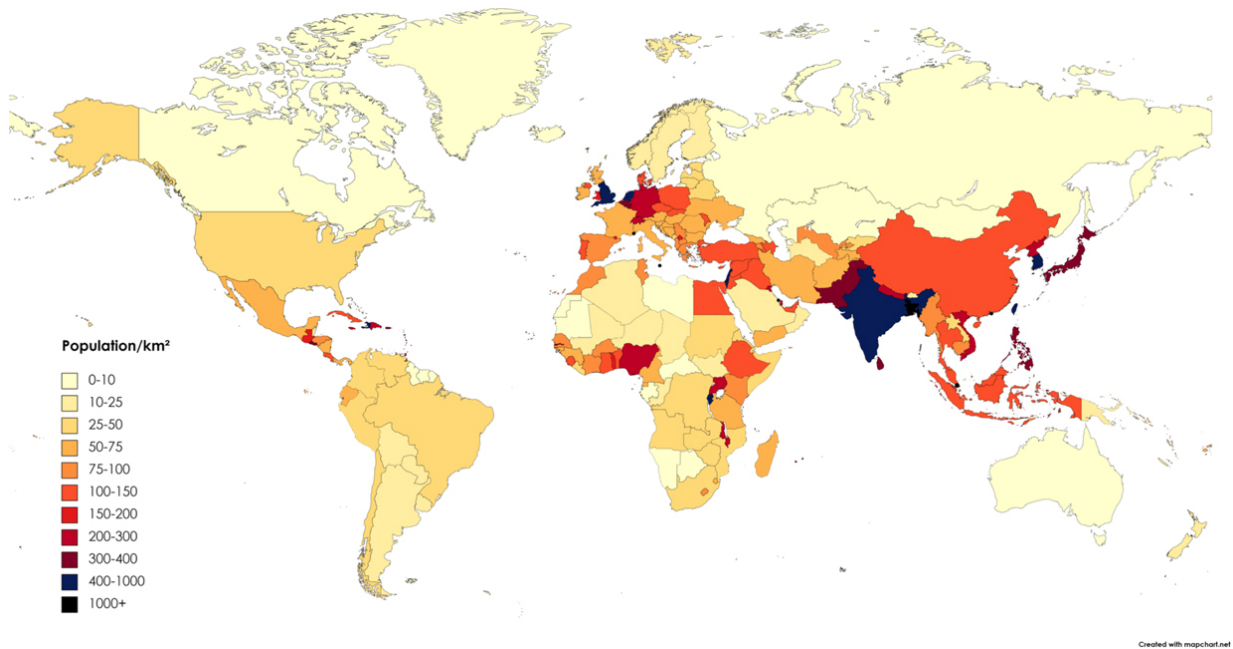


FIGURE 3.21: Source: Wikimedia commons. Heatmap of Population. [🔗](#)

3.13 Summative Project

Project 3.13.1 Exploring data sets: Ranking sports by skill requirements.

In this project, you will conduct exploratory data analysis to answer questions, analyze charts, and create different types of data visualizations using Google Sheets tools.

Overview: The data set lists 10 categories of athletic skills, such as endurance, strength, power, speed, agility, flexibility, nerve, durability, hand-eye coordination, and analytic aptitude to rank 60 different sports with the help of a panel of sports experts. Each skill is given a numerical score on a scale of 1 to 10, and a final total score which is the sum of these scores. Read about these traits in this link: [🔗](#) Here is the link to Google Sheets containing the data [🔗](#).



FIGURE 3.22: Image by [rawpixel.com](https://www.rawpixel.com) on Freepik.

1. Descriptive statistics

Examine the scores for the different sports, and answer the following questions:

- (a) Consider the skill hand-eye coordination, what is the median score for this skill? Which sports have scores approximate to this median score?*
- (b) What is the mean of the score for strength? Is this statistic meaningful for this data set?*
- (c) What is the minimum and maximum score for nerve? Which sport requires minimum nerve? And which sport requires maximum nerve?*
- (d) Consider the skill analytical aptitude. Which three sports have a score greater than 7 for this skill?*

- (e) Which is the toughest sport based on the skill scores in this table? Do you agree? Why?
- (f) What were some interesting and surprising information in this data set? Explain.

2. Data Visualization: Creating and analyzing charts using Google Sheets

- (a) Using the insert chart feature, create a scatterplot speed vs. agility. Give your graph appropriate title and labels for the axes. What do you notice?
- (b) Looking at the scatterplot, what is the relationship between the variables strength and power?
- (c) Examine the scatterplot shown in [Figure 3.23](#). What information can you gather? Which sport requires the most strength and power?
- (d) Examine the histogram shown [Figure 3.23](#). What do you notice?
- (e) Name two sports that need power that is greater than 8. Name two sports that need power less than 3.
- (f) In the histogram shown in [Figure 3.23](#), which interval has the highest frequency? What does that imply?
- (g) Find the mean, median and mode for the skill power.

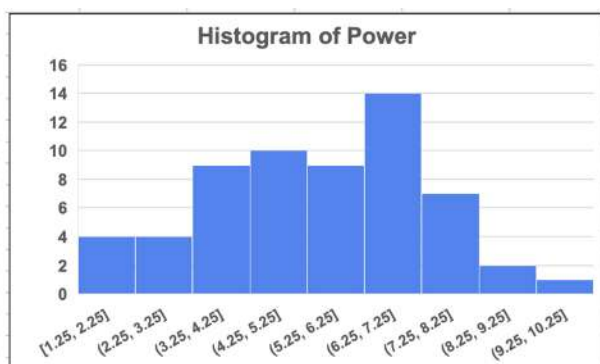
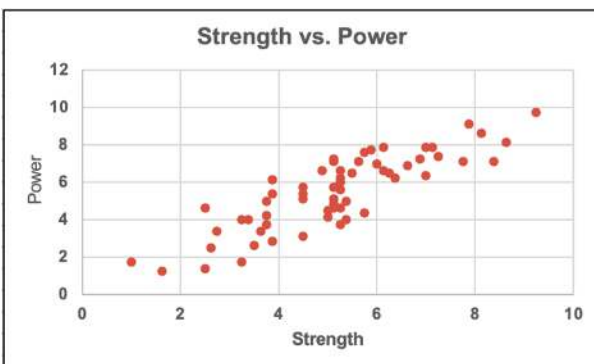


FIGURE 3.23: Left: Scatterplot of Strength vs. Power. Right: Histogram depicting the distribution of scores. Charts created using Google Charts tool.



3. Using AI for EDA

Use a Gemini AI prompt to find out how a data set from Google Sheets or CSV file can be uploaded to Gemini for data analysis, and then complete the following two tasks.

- (a) Use Gemini AI prompts to analyze the sports data set and find descriptive statistics for a specific sport of your choice.*
- (b) Use AI prompts to create two different charts and graphs for a sport of your choice. Check to see if the AI outputs are valid.*

4

Sampling: Less Is More

It is estimated that there are 10 quintillion insects, that is 10,000,000,000,000,000, 000 bugs, on this planet. We could not possibly study all of them. How did they arrive at that number in the first place? *Sampling*, where an appropriately chosen subset is used to make inferences about the entire population of interest, makes it possible to study insects without having to examine each and everyone individually. Sampling is used extensively in AI. This chapter motivates the concepts of populations and samples, the need for sampling, different types of sampling, variations caused by sampling and the impact it has on results, and how to ensure that results based on a sample are applicable to the population in general.

4.1 Why Sample?

Suppose that a truck carrying millions of loose Skittles breaks down and spills all its contents along a 100-yard stretch which is roughly the length of a football field. The truck driver says he needs to know the proportion of red skittles in the spill, urgently, and promises a big prize to the first person with the correct answer. What would you do? What if you only had two minutes to provide the answer?

Typically, when there are too many objects to study individually, whether they are insects, penguins or skittles, it is not possible to study all of them. It is enough to focus on a relatively small, appropriately chosen subset. The process of choosing a subset, called *sampling*, is used extensively in almost every area, by environmentalists that study changes in air, water, ice ([Figure 4.1](#)) and soil, to marine biologists who monitor migration patterns, to influencers and marketers who use social media to increase user engagement.



FIGURE 4.1: Would it be humanly possible to analyze and study all the ice and water on the planet? Left: Taking water samples, the U.S. Geological Survey. Right: Ice core sampling in Green Bay, Lake Michigan by NOAA Great Lakes Environmental Research Laboratory. [↗](#)

A *population* is the set of all objects of interest relevant to a problem. For example, to study popular sports on American campuses, the population would consist of all students enrolled in an American campus. A *sample* is a carefully chosen subset of the population that is representative of the population. The objective is to study just the sample and make inferences about the population in general. Students selected randomly from campuses across the United States balanced for gender, age and other factors would constitute a representative sample.

Sampling has many advantages:

- It is cost effective, since it is cheaper to observe a small set of objects. For instance, chasing down every whale in the ocean to measure its weight would take a lot of people in a lot of boats a lot of time. It might not be possible to track down every whale even then.
- Managing sample data is more efficient in terms of storage, processing and computing, since it is only a small fraction of the entire population.
- Sampling is better for the environment, in terms of physical impact as well as in terms of the energy requirements. It would be extremely damaging to dig up the entire ocean bed or all the mountain tops to study their mineral composition.
- In some instances, it is harmful to collect data. If a person is ill, it might be dangerous to take too much tissue for analysis.

The classic *Sampling Techniques* by William G. Cochran [9] provides a comprehensive coverage of sampling topics and their applications. Review the terminology and definitions in [ReferenceBox 4.1.1](#) thoroughly.

In the instance of the spilled Skittles, the population consists of the Skittles in the truck that were spilled on the street. Examples of other “populations” of Skittles include: all the Skittles sold in New Jersey, or all the Skittles produced in 2023, or all the Skittles bought in the USA in 2023. The notion of population therefore depends on the problem being studied.

ReferenceBox 4.1.1: Populations and Samples [↗](#)

- The set of all objects relevant to a study is called a *population*. The number of objects in the population is called the *population size*, denoted by N .

- The process of selecting a subset of objects from a population for the purpose of gathering data and studying is called *sampling*. The subset of data created through sampling is called a *sample*. The number of objects in a sample is called the *sample size*, usually denoted by n .
- The particular process used to create the sample from the population is called a *sampling method*. There are numerous sampling techniques to meet different types of study criteria.
- The proportion of the sample size to the population size, n/N , is called the *sampling proportion*.

Data Debate 4.1.1 *What are the advantages and disadvantages of working with a sample rather than the entire population?*

A popular method of sampling is to choose objects at random. For example, in a raffle draw, the tickets have unique numbers, and a subset of lucky winners is chosen randomly from among them. The “population”, in this instance, consists of all the tickets sold, and the “sample” consists of the winning tickets. What if the owner of the raffle wanted to reward his or her friends and chose their ticket numbers as the winners of the raffle? Or, chose the first (or last) five tickets sold to be the winners?

Data Debate 4.1.2 *Can two different random samples have the same objects in them? Is it possible for a sample to have bias?*

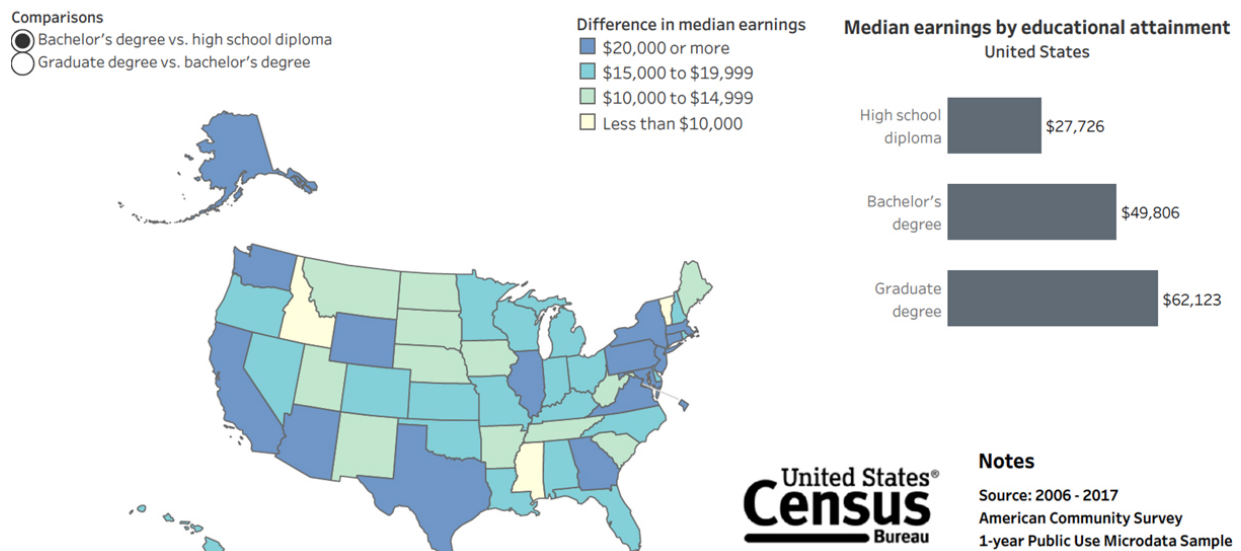
Analyzing statistical results based on samples versus populations has demonstrated that in most instances an appropriate sample approximates the results of the entire population quite well. For example, descriptive statistics such as the mean and the standard deviation computed from suitably chosen samples are close to the descriptive statistics based on the data of the entire

population. We can even quantify how closely the results based on a sample will approximate the results of the true population. [Chapters 5](#) and [6](#) discuss the statistical foundations of quantifying the confidence in results based on samples.

Sampling is a common practice in many data-driven studies. For example, [Figure 4.2](#) displays the results of disparities in income based on education attainment. The study uses only 1% of the census data. The graphic illustrates that:

- The median incomes, nationwide, for those with a high school diploma, bachelor's degree or graduate degree are \$28K, \$50K, \$62K, respectively.
- The disparities in income by educational level vary from region to region, shown by the heatmap.
- The disparities in income based on education levels are greatest in states depicted by the darker shades of blue, e.g., California, Texas and New Jersey, of \$20K or more per year.

Do young adults get what they paid for?



[Extended Description](#)

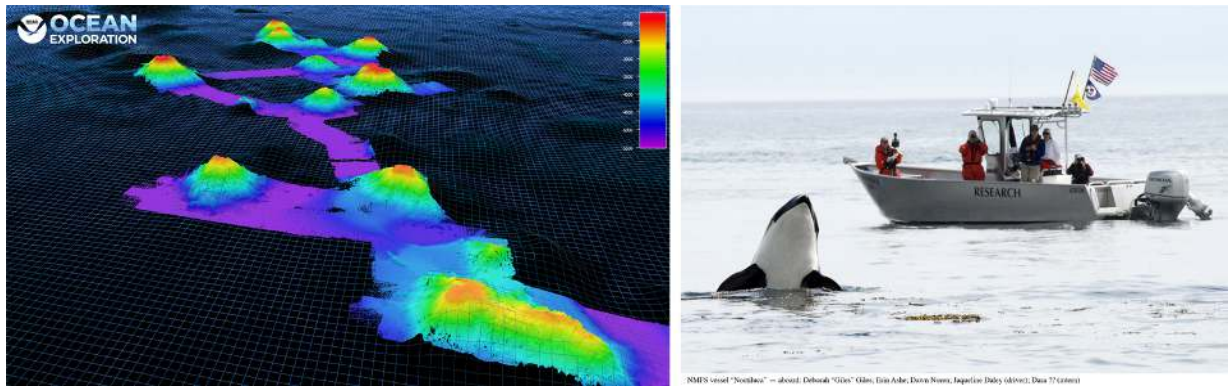
FIGURE 4.2: The United States Census Bureau provides small samples of census data to the public. The above visualization of income disparity based on education attainment is based on a 1% sample of the entire US population (335 million). Click for interactive graphics [🔗](#). [📄](#)

4.2 Illustrative Examples of Sampling Applications

Given the amount of data that is being collected daily by sensors in space, on the ocean, in cars and online, sampling is essential for real-time analytics and decision-making. Streaming companies sample viewing patterns to identify popular shows and segments. Different types of voices are sampled to create universally appealing synthetic voices for AI-assistants such as *Siri*. Satellites sample images of the earth to identify natural disasters like wild fires and floods. In this section, we discuss three illustrative examples of how sampling is used in the real world.

4.2.1 Sampling in environmental sciences

Sampling is widely used for research by biologists and environmental scientists because it is often difficult to study entire populations, e.g., whales. So, how would one collect biological samples from a killer whale swimming through the ocean? [Figure 4.3](#) shows how NOAA research scientists collect samples from orcas and dolphins.



[Extended Description](#)

FIGURE 4.3: Left: NOAA's image of bathymetry data from ocean floor mapping through sampling. Right: NOAA Research: Collecting samples from a boat. A pantropical spotted dolphin surfacing alongside the NOAA Ship Gordon Gunter. Credit: NOAA Fisheries (Permit No.14450). [↩](#)

Typically, biologists ride alongside whales or dolphins in a boat and use a long pole with petridishes attached at the end. The samples are used by marine scientists for killer whale conservation and research programs. For example, in a 2017 research paper, Linda Rhodes (manager of the Marine Microbes and Toxins Program at NWFSC) and her colleagues analyzed microbes in breath collected from Southern Resident killer whales. They found that some of the bacteria present were resistant to antibiotics, and certain bacteria posed a risk of disease. Read about Southern Resident Killer Whales and sampling in [Section 4.6](#).

Sampling is also used in the Gulf of Mexico to study the effect of oil spills, e.g., 2010 Deepwater Horizon oil spill and another one in 2016, on pantropical spotted dolphins shown in [Figure 4.3](#). According to NOAA scientists, ship-based surveys recorded fewer sightings and smaller group sizes. The survey team samples multiple groups or sightings to ensure an appropriate representation of the genetic diversity in the region. Read more about the pantropical spotted dolphins study in [Section 4.6](#).

Marine research and study of oceans is supported by another instance of sampling. Research organizations of NOAA map ocean floors through a special type of sampling to discover new topological features and provide accurate maps for safe navigation of ocean-going vessels. Ocean exploration is expensive and dangerous; such maps give researchers the best possible coverage of the ocean floor while remaining within time and resource constraints.

4.2.2 AI and sampling

The use of AI and machine learning models has helped make sampling more efficient. Machine learning models study patterns such as whale movements and identify high-probability locations and high-likelihood times when the pods are active. This enables scientists to place AI-equipped drones and camera sensors with image recognition capabilities at optimal locations to maximize the potential for capturing meaningful data rather than just an empty view of the ocean. AI makes data collection more cost-effective instead of wasting long periods of time and resources monitoring locations when there is no activity. To learn more, try the prompt in [GeminiBox 4.2.1](#).

GeminiBox 4.2.1: Sampling and AI

“How can AI be used for sampling in environmental sciences studies?”



Sampling is used extensively in driving user engagement on social media sites. Given the volumes of data generated by digital platforms, it would be impossible to react in a timely manner without sampling. Latency in data-driven decision-making such as those used by autonomous vehicles, e.g., “what is the speed of that moving object and what path is most likely to avoid

a collision”, could potentially have life-threatening consequences. Sampling reduces such latency by reducing the amount of data that needs to be processed.

4.2.3 Monitoring quality control through sampling

Occasionally the products we buy, like shirts, hair brushes or toasters, have little stickers that say “inspected by 007,” implying that they were inspected and cleared by a quality supervisor, Inspector 007. Sampling is used in manufacturing and many industries to ensure that products meet specifications, e.g., a hairbrush must have 20 bristles per square inch. Given how many thousands of hairbrushes are manufactured, it would be impossible to examine each and every one of them.

Industrial manufacturing has a systematic and statistical way of monitoring the quality of processes and goods. Edward Deming is considered to be the founder of the field of statistical quality control, and for revolutionizing industrial manufacturing monitoring through continuous sampling. Douglas Montgomery's *Introduction to Statistical Quality Control* provides a comprehensive overview of the topic [\[30\]](#).

Consider the following hypothetical example to understand how sampling is applied in the food industry. Such customer experience criteria are often used in restaurant rating apps such as *Yelp*.

Example 4.2.1 *The Farmhouse Bakery is famous for its cookies and cakes. In order to maintain their sales and reputation and grow their business, the owners want to ensure the quality of their food, the performance of their employees, the atmosphere and cleanliness of the premises, the quality of the ingredients, and the taste and repeat-purchase appeal of the baked goods.*

Here are some actions the owners could take, specific to each goal:

- Efficiency and performance of staff: randomly observe employees each week/month
- Cleanliness and atmosphere: inspect the premises several times a week, survey customers randomly to get their input
- Quality of ingredients from suppliers: select ingredients each week to check for freshness and expiration date
- Taste: select a batch of baked goods to see if they meet the flavor and taste that is expected, and repeat this step periodically

It is evident that sampling plays a big part in Farmhouse Bakery's effort at maintaining high quality. There are different types of sampling that the bakery can choose from, and selecting the appropriate one is the key to success.

Data Debate 4.2.1 *Why does quality control matter? Discuss a specific example.*

Data Debate 4.2.2 *Is sampling a good option for quality control? Are there other means of quality control in manufacturing?*

AI has brought new innovation and technology to industrial quality control that can both speed up the process as well as reduce human error in manual inspections. Read *Visual Inspection AI: a purpose-built solution for faster, more accurate quality control* in [Section 4.6](#).

4.3 Types of Sampling

To understand different types of sampling, consider the following hypothetical example. A robot factory makes five different types of robots in

five colors, with five different functions: play games (blue), perform household chores (brown), inspect cell towers (pink), take care of invalids (green), or monitor traffic (purple).

A team of research students interested in computer gaming strategies visit the factory and are allowed to interact with 5 robots from a room full of 25 robots, numbered 1 through 25 as shown in the box on the left in [Figure 4.4](#). The population size is $N=25$ since there are 25 robots in the room, and the sample size is $n=5$ since students are allowed to choose 5 robots to interact with.

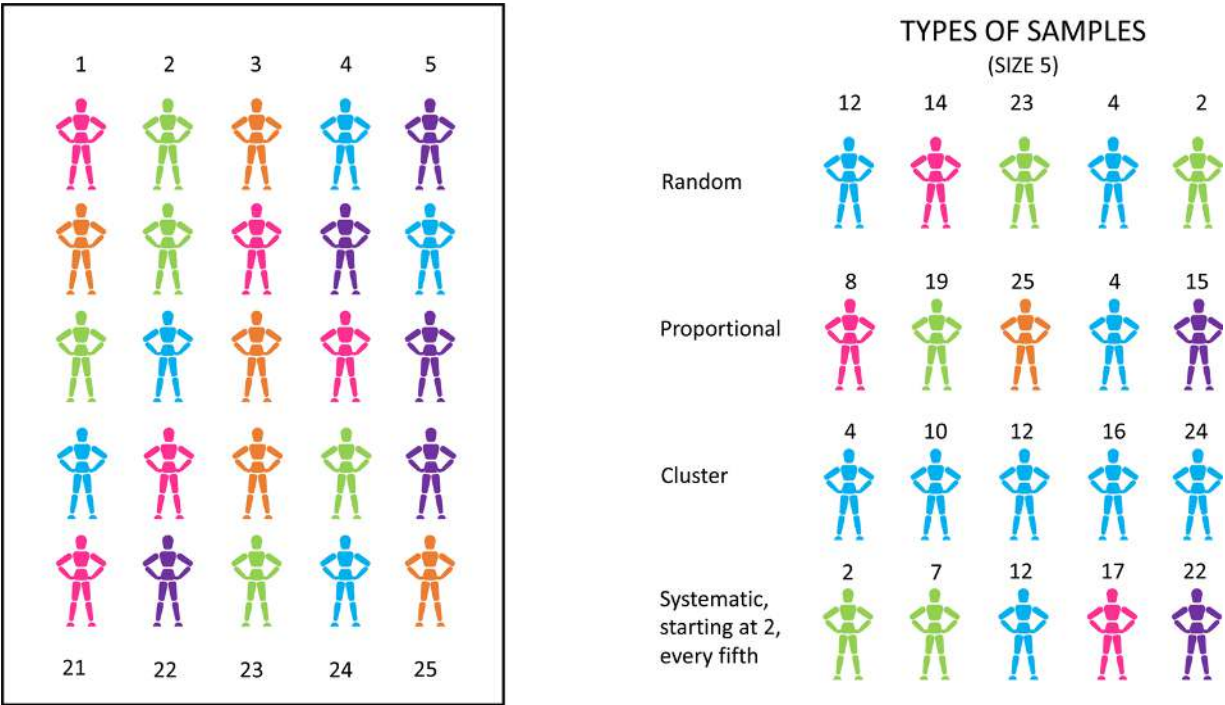


FIGURE 4.4: The box on the left shows a population of 25 robots, of 5 different types, denoted by a different color. They are numbered 1 through 25 in no particular order. Suppose you need to choose a sample of 5 robots. On the right-hand side there are 4 samples of 5 robots each, each created using a different type of sampling. [↗](#)

How should the students choose the 5 robots? One student suggests choosing 5 numbers between 1 and 25 randomly, by using a random number generator such as the Number Picker Wheel here . The random number generator spits out: { 12, 14, 23, 4, 2}, and the students choose the robots with those numbers from the box on the left, to create the first sample on the right in [Figure 4.4](#). Such a sample is called a *random sample*.

Another student suggests that since the random sample did not include each type of robot (brown and purple are missing), in order to experience interacting with all the five different types of robots the students should choose one of each type reflecting the proportion of each in the population, creating the second sample in [Figure 4.4](#). This is a *stratified sample*, which ensures that each group is represented in the sample. It is also a *proportional sample* since the proportion of each type is the same as that in the population.

Another student mentions that the blue robots are trained to play games with the students, and since that is the research focus of the student group, they should choose the entire group of blue robots as the sample, as shown in the third sample on the right in [Figure 4.4](#). Such a sampling method is called *cluster sampling*. Cluster sampling is useful when the sample is meant to serve a very specific goal, for example researching robots that play games. In such an instance, robots that are meant for other tasks would be of no interest to the sampling process.

Yet another student argues that the blue robots are very expensive, so the factory might not allow the students to use all five of their blue robots. The student goes on to suggest that the robots should be lined up any which way and every fifth robot should be chosen. In [Figure 4.4](#), the fourth example on the right is a *systematic sample* of every fifth robot starting with robot no. 2.

(*Periodic sampling* is a type of systematic sampling where samples are selected at regular intervals in time. For example, it is used to poll telecommunications network systems periodically to monitor performance

and detect unusual activity. Periodic sampling can be misleading if not applied properly, e.g., if the sampling interval matches the periodicity of the phenomenon being studied. For example, if we sampled a person's behavior at 3 AM every 24 hours, we would conclude that the person's sole function is to sleep.)

Data Debate 4.3.1 *Does the starting point make a difference to the sample? How many distinct systematic samples are possible for a sample of size n from a population of size N ?*

Another student suggests that they keep choosing robots until there are at least two blue robots in the sample. This could happen in just two tries if the first two robots that are picked happen to be blue, or might take more than 5 samples. Such a method of sampling is called *progressive* sampling.

Data Debate 4.3.2 *The worst case, i.e., the largest size of a sample needed for the progressive sampling of at least two blue robots is 22. Why?*

Review the definitions in [ReferenceBox 4.3.1](#).

ReferenceBox 4.3.1: Types of Sampling

- *A random sample* consists of objects selected from the population in no particular order or pattern, e.g., by using a random number generator or by drawing numbers blindly from a box like in a raffle lottery.
- *A stratified sample* consists of objects selected from each group in the population to make sure that every group is represented. If the proportion of each group in the sample is the same as the proportion in the population, it is a *proportional* sample.

- *Cluster sampling* involves choosing an entire group to be included in a sample, in contrast to proportional sampling.
- *Systematic sampling* involves choosing every k^{th} object in the population, resulting in a sample of size N/k .
- *Adaptive or progressive sampling* requires adding objects to a sample until a particular goal, e.g., at least two blue robots, is met.

The classic *Sampling Techniques* by William G. Cochran [9] provides a comprehensive coverage of sampling topics and their applications.

4.3.1 Hands-on sampling practice: Palmer penguins

Consider the example of Palmer penguins data discussed earlier. Suppose we need to choose a proportional sample that represents each species in proportion to its representation in the population. The population count and proportion of each species are provided in [Table 4.1](#).

TABLE 4.1: Count of penguin species in a proportional sample of size 30. [↩](#)

Species	Count	Population Proportion	Calculation of Sample	Proportional Sample	Rounded
Adelie	151	$151/342=0.44$	$.44 \times 30$	13.25	13
Chinstrap	68	$68/342=0.20$	$.20 \times 30$	5.96	6
Gentoo	123	$123/342=0.36$	$.36 \times 30$	10.79	11
Total	342	1.0			30

To create a proportional sample of size n , say, $n = 30$, follow the instructions below:

- First, compute the population proportion of each species by dividing the population count of the species by the total population size of $N=342$, as shown in Column 3 of [Table 4.1](#).
- Next, multiply the population proportion by the desired sample size as shown in Column 4, to find the count of penguins of each species to be included in the sample Column 5.
- Since fractional penguins sounds exceptionally cruel, round the results to the appropriate number as shown in Column 6, and make sure that the counts of species in the sample add up the desired sample size of $n = 30$.

It is important to work through the hands-on example above to get an intuition for proportional sampling. However, in practice, most data analysis software tools provide options for choosing samples automatically, e.g., Excel and Google Sheets have built-in functions for sampling from a large data set. Step-by-step instructions on how to generate random samples using Excel can be found at this link [🔗](#), and step-by-step instructions on how to generate random samples using Google Sheets are at this link [🔗](#). The instructions are also available in the textbook's Supplementary Resources [🔗](#).

4.3.2 To replace or not?

In the robot example, suppose we wish to study the strength of robots. Measurements regarding strength, e.g., maximum weight that the robot can lift, are collected from each robot in the sample. A robot is set aside once it has been chosen for inclusion in the sample and will not be selected again. However, sometimes it is useful to return an object to the population to be included again in the next sample selection.

To demonstrate, consider the following example, where the robots are numbered from 1 through 5. The numbers refer to the robot's ID number, not

its strength measurement. Suppose that there were only 5 robots in the population. If we need *measurements* from a sample of 4 robots, we only have 5 possible sample choices, robots no. (1,2,3,4), (2,3,4,5), (3,4,5,1), (4,5,1,2) and (5,1,2,3). Most of these samples resemble each other since any two samples have at least 3 robots in common, hence three strength measurements in common. Each sample object selection reduces the number of remaining robots to be chosen for the sample. That is, after the first sample robot is selected, there are 4 robots to choose from for the second sample object, three robots for the next sample object selection, and so on until only two robots are left for the selection of the last and fourth sample object. This method of sampling, where once an object is selected for a sample it is removed from sampling contention, is called *sampling without replacement*.

In order to produce reliable models and analytical results, AI and statistical studies often need more samples, and samples that are quite different from each other in order to represent as much of the population as possible. In such a situation, particularly if the population is small, the *measurements* of a sample object are allowed to be *duplicated* in the sample. This is done by replacing each robot that is chosen for the sample back in the population so that its measurements could be potentially included in the next sample selection. There will always be 5 robots to choose from. This method of sampling where an object is returned to the population after being selected for a sample is called *sampling with replacement*.

In the example of five robots, suppose that during the first sample selection, robot number 4 is selected. Then, for the selection of the next robot for a sample of size 4, the Population is $\{1,2,3,5\}$. If we continue to sample in this manner without replacing the robots back in the population, the population size will decrease by 1 each time. But if we restore the sampled robots back to the population after each selection, each robot's strength measurements have an equal chance of being selected during each sample

selection. If possible, the measurements are taken afresh each time the object is selected for the sample, called *repeated measurements*, as a way to control measurement error. We can continue to choose as many robots for as many samples as needed since the same robot's strength measurements can appear more than once in the sample.

[GeminiBox 4.3.1](#) shows an example of sampling with replacement. Robot #3 appears twice in Sample #1, so its strength measurements will be replicated in that sample. Explore further by using different populations instead of the number set (1, 2, 3, 4, 5), and different sample sizes.

GeminiBox 4.3.1: Sampling with replacement [↗](#)

In response to the prompt, “Create three samples of size 4 with replacement from (1, 2, 3, 4, 5),” Gemini AI-bot created the following output:

Sample 1: (2, 3, 3, 5)

Sample 2: (1, 1, 2, 4)

Sample 3: (3, 4, 5, 5).

ReferenceBox 4.3.2: Sampling with and without replacement

- *Sampling without replacement* refers to the process where an observation about an object in the population can be included at most once in a sample. Once an object is chosen for a sample, it is removed from sampling contention.
- *Sampling with replacement* refers to the process where an object in the population can be included multiple times in a sample. Once an object is chosen for a sample and its information is recorded, it is returned to the population and is available to be sampled again.

- *Resampling* is the process of creating multiple samples from one single sample using sampling with replacement. Samples created in this manner are used for studying variability of the results of statistical and AI models.
- *Bootstrapping* is a type of resampling that samples with replacement when there are too few objects in the population.

[Figure 4.5](#) shows 10 samples of size 5 drawn with and without replacement from the set of numbers 1 through 25. When the samples were created with replacement, numbers repeat and occur multiple times in a sample, as in Samples 3 and 6. The samples on the right were created without replacement so that no number occurs more than once in any given sample. Gemini computed the mean of each sample as well. The results are shown in [Figure 4.5](#). Here the population consists of the numbers 1 through 25. Notice that each sample is different. Since the samples are different, the descriptive statistics computed from the samples are also different.

- | | |
|--|---|
| • Sample 1: {3, 7, 11, 15, 19} Mean: 11 | • Sample 1: {2, 5, 11, 16, 18} Mean: 10.4 |
| • Sample 2: {1, 5, 10, 15, 20} Mean: 10.2 | • Sample 2: {1, 4, 8, 13, 19} Mean: 9 |
| • Sample 3: {2, 2, 8, 14, 18} Mean: 8.8 | • Sample 3: {3, 7, 10, 14, 20} Mean: 10.8 |
| • Sample 4: {5, 9, 12, 16, 20} Mean: 12.4 | • Sample 4: {6, 9, 12, 15, 17} Mean: 11.8 |
| • Sample 5: {4, 6, 10, 13, 17} Mean: 10 | • Sample 5: {2, 5, 8, 14, 19} Mean: 9.6 |
| • Sample 6: {1, 8, 8, 15, 19} Mean: 10.2 | • Sample 6: {1, 7, 11, 16, 20} Mean: 11 |
| • Sample 7: {3, 7, 11, 14, 20} Mean: 11 | • Sample 7: {4, 6, 10, 13, 18} Mean: 10.2 |
| • Sample 8: {6, 9, 12, 17, 18} Mean: 12.4 | • Sample 8: {3, 9, 12, 15, 17} Mean: 11.2 |
| • Sample 9: {2, 5, 10, 16, 19} Mean: 10.4 | • Sample 9: {2, 8, 10, 14, 16} Mean: 10 |
| • Sample 10: {4, 8, 13, 15, 20} Mean: 12 | • Sample 10: {5, 7, 13, 18, 20} Mean: 12.6 |

FIGURE 4.5: Ten samples of size 5 from the set of numbers 1 through 25, and the corresponding sample means, created using Google Gemini. Left: With replacement. Right: Without replacement. [↩](#)

Data Debate 4.3.3 *What happens to the size of the population pool available for sampling when sampling with replacement and sampling without replacement? Give examples of situations where you would use sampling with replacement and sampling without replacement.*

Data Debate 4.3.4 *Is there a difference in descriptive statistics computed by sampling with replacement and without replacement? Examine [Figure 4.5](#), compare the sample means to the population mean and variance, state your conclusions, and explain why.*

GeminiBox 4.3.2: Cluster Sampling with replacement

“Explain cluster sampling with and without replacement.” 

The most significant use of sampling with replacement is studying the variability of the performance of statistical and AI models. *Resampling* is used to artificially create multiple samples from an existing sample using sampling with replacement. Analyses and statistics based on these multiple samples are used to study possible variation of results across different sections of the data. *Bootstrapping* is a particular type of resampling that is used in machine learning and AI to understand the variability in performance of AI models. For those interested in pursuing machine learning, Bradley Efron's original work, *The Jackknife, the Bootstrap and Other Resampling Plans* [[15](#)] is an excellent introduction to resampling.

4.4 Sampling Variability and Distributions

The variation in the mean from sample to sample is called *sampling variability*. This is true of any descriptive statistic, model or visualization

computed from samples, since samples are subsets of the population and each subset is bound to be different from others. Being able to quantify the variability, and guarantee that the true population values can be approximated by the sample-based values is critical. For example, when an autonomous car is trained on a sample of highways, it is important to be able to guarantee that it will perform equally well on highways it has not been tested on. Statistics provides a strong foundation for providing such guarantees. In this section, we will provide an empirical explanation of sampling variability and sampling distributions. [Chapters 5](#) and [6](#) will present a formal probability-based discussion.

Sampling variability of the sample mean can be studied with EDA techniques from [Chapter 3](#). For example, compute the mean of the sample means, check the dispersion of the values of sample means by computing their standard deviation, and build a histogram to see how the sample mean values are spread across the number line.

To demonstrate this approach, we created a 100 samples of bill length from the Palmer penguin data, each sample of size 100, using the *R* programming language. We computed the means of each of these 100 samples and plotted a histogram of the 100 sample means, shown in [Figure 4.6](#), left. The vertical blue dashed line represents the population mean of all 342 penguins, an average bill length of 43.8. From [Chapter 3](#), we know that the bars of the histogram represent the proportion of values, in this case sample means, that fall in the corresponding bins. According to the histogram, most values of the sample mean are close to the population mean, and fall in the interval [43, 45].

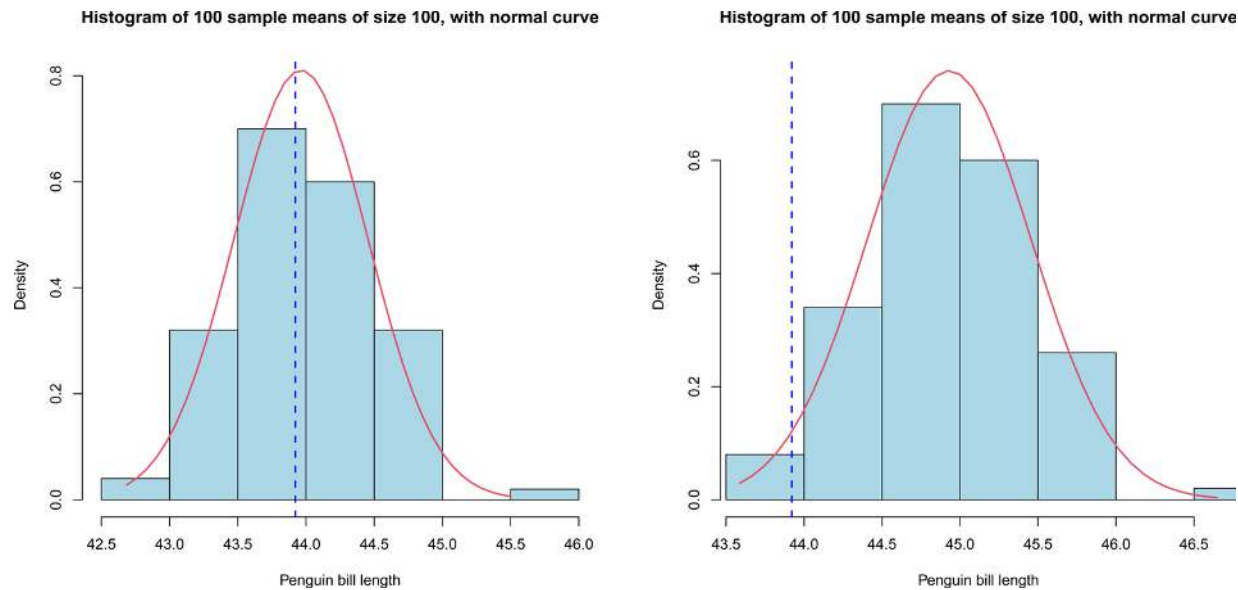


FIGURE 4.6: Left: Histogram of 100 sample means of size 100 each, is symmetric around the true population mean of penguin bill length denoted by the vertical dashed blue line. As the sample size increases, the histogram will resemble the red bell-shaped curve. Right: Histogram of trimmed sample means where the smallest 10% of values were dropped from the computation of the mean. Most of the values are higher than the population mean as evidenced by the shift of the histogram to the right, away from the true population mean denoted by the dashed blue line at 43.8. The figure was created using R. [↩](#)

If we were to increase the sample size of the means from a 100 to a very large number, the histogram would eventually resemble the red bell-shaped curve which, as we will learn in [Chapter 5](#), follows the 68-95-99.7 rule. The rule states that for any variable whose data values are spread across the number line in the bell-shaped distribution, 99.7% of the values will fall within three standard deviations of the mean of the values. In this instance each value is a sample mean. Therefore, if we compute the mean of the sample means, i.e., the mean-of-means, and the standard deviation of the sample means, known as the *standard error*, we can expect that 99.7% of the

sample means will fall within 3 standard errors of the mean-of-means. The computations are shown below.

The mean-of-means is given by:

$$\overline{\overline{X}} = \frac{\sum_{i=1}^n \overline{x}_i}{n}, \quad (4.1)$$

where

$$\overline{x}_i = \frac{\sum_{i=1}^n x_i}{n}, \quad (4.2)$$

and n is the sample size. The standard deviation of the sample means, also called the *standard error* is given by

$$s_{\overline{x}} = \sqrt{\frac{\sum_{i=1}^N (\overline{x}_i - \overline{\overline{X}})^2}{n}}. \quad (4.3)$$

Standard error refers to the error in estimating the population mean using the sample mean.

Therefore, using the 68-95-99.7 rule, we would expect 99.7% of sample means will be contained in the interval

$$[\overline{\overline{X}} - 3 * s_{\overline{x}}, \overline{\overline{X}} + 3 * s_{\overline{x}}] \quad (4.4)$$

This interval and its boundaries are often referred to as *margin of error* and *error bounds*, respectively. We will learn more about the bell-shaped curve and these error bounds in [Chapters 5](#) and [6](#).

Just as the sample mean has a data distribution associated with it, every statistic computed from a sample will also have an associated sampling distribution due to sampling variability. Try the prompt in [GeminiBox 4.4.1](#) to study the sampling distribution of the sample maximum. Explore sampling distributions of other statistics such as the median and quartiles.

GeminiBox 4.4.1: Sampling distribution of the maximum value of a sample

“Plot the sampling distribution of the maximum value of a sample.” 

4.4.1 Effect of sample size

In general, if the samples are large, they will tend to resemble each other as well as the population, resulting in lower variability. This is to be expected since large samples contain a larger chunk of the population. On the other hand, the smaller the size of the sample with respect to the population, the greater the difference between the samples, and the greater the variation in descriptive statistics computed from the samples. As the sample size increases, the descriptive statistics in most cases increasingly resemble the population values, because the sample, as it gets bigger, will increasingly resemble the population. In fact, descriptive statistics that have this property are desirable since the goal of using samples is to estimate the population values as well as possible with less data.

In order to study the effect of sample size on sampling variability, we conducted the following empirical studies:

- For a given sample size, we created 20 samples of bill length from the Palmer penguin data
- Computed the mean $\bar{x}$ of each of the 20 samples as shown in Equation 4.2
- Computed the mean-of-means $\bar{\bar{X}}$ of the 20 sample means as shown in Equation 4.1
- Computed the standard deviation $s_{\bar{x}}$ of the 20 sample means, also known as standard error, as shown in Equation 4.3
- Computed the error bounds shown in Equation 4.4

We repeated the above steps with different sample sizes and created [Figure 4.7](#), left, where the X -axis represents the size of the sample, the blue curve represents the mean-of-means from Equation 4.1 for the corresponding sample size n on the X -axis, the red dotted lines represent margin of error from Equation 4.4, above and below the mean-of-means. The horizontal purple line represents the true population mean of the bill length of all 342 Palmer penguins.

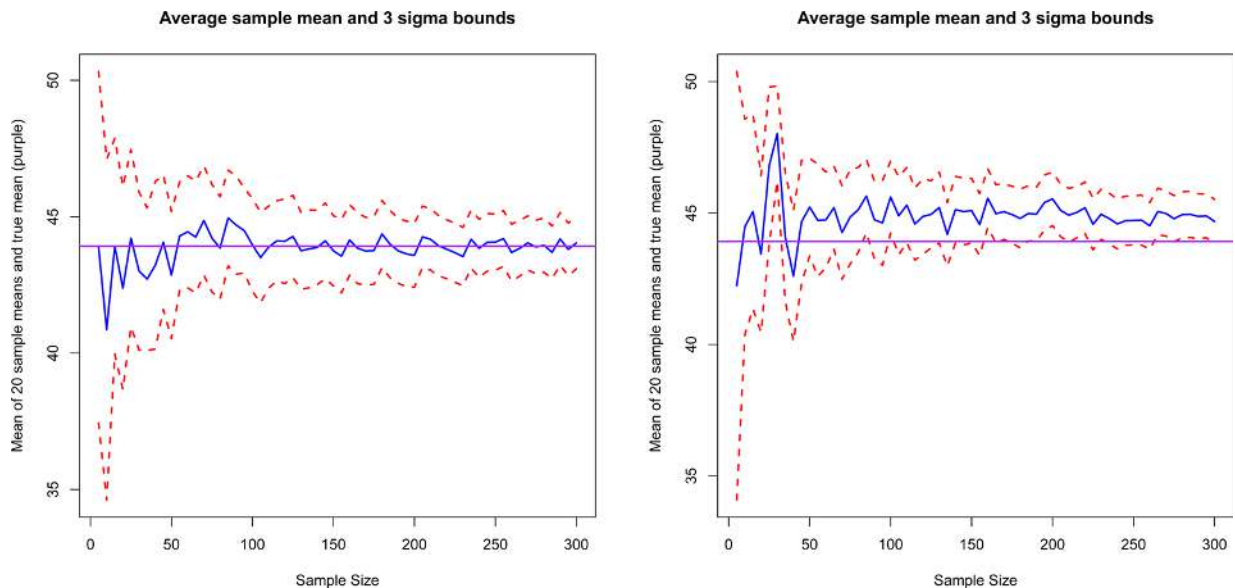


FIGURE 4.7: Bill length of Palmer penguins. Left: As the sample size of the sample means increases (X -axis), the average of 20 sample means gets closer to the true population mean represented by the horizontal purple line, and the margin of error (represented by the 3-standard error limits (dashed red lines)) decreases. Right: The experiment is repeated, but with only the top 90% of values in the sample. Such a mean, called a *trimmed mean*, is a *biased* estimate of the population mean and will never approximate the true population mean (horizontal purple line), no matter how big the sample size. The figures were created using R code. [↗](#)

Intuitively, [Figure 4.7](#) shows that as the sample size increases, the mean-of-means (dark blue line) gets closer to the true population mean. This is to be expected since larger samples contain a bigger subset of the population and will resemble the population more. In addition, the variation in the sample means lessens as the sample size increases, denoted by the narrowing of the margin of error, the dashed red lines around the mean-of-means. This implies that for a large enough sample size, we can expect margin of error to be smaller. Such a descriptive statistic whose value gets consistently closer and closer to the population value as the sample size increases is said to be an *unbiased* estimator of the true population value.

4.4.2 Statistical bias

Next, we repeated the experiment but dropped 10% of the smallest values from each sample before computing the mean. Such a mean is sometimes called a *trimmed mean* since we trimmed the lower 10% of the data from the sample. The histogram of a 100 trimmed means based on samples of size 100, [Figure 4.6](#), right, shows that most of the sample mean values are higher than 43.8, the true population mean denoted by the dashed blue line. This is to be expected since we dropped the smallest 10% of values from each sample. No

matter how big the sample, the histogram will always be shifted to the right of the true population mean, and never be representative. [Figure 4.7](#), right, shows the changes in the trimmed sample means as the sample size increases. Compared to the plots in [Figure 4.7](#), left, the average of the trimmed sample means is consistently higher than the true population mean represented by the horizontal line, irrespective of the sample size. This is to be expected since we only choose the top 90% of sample values and ignore the bottom 10% of sample values. The sample trimmed mean is a *biased* estimate of the population mean and will never be representative of the true population mean no matter how large a sample we choose.

Data Debate 4.4.1 *What do you think will happen if we compute the sample median with only the top 80% of the values in a sample?*

ReferenceBox 4.4.1: Sampling variation, error bounds and bias

- *Sampling variability* is the variation in statistics and models from sample to sample. Each sample contains a different set of objects from the population, hence any statistic or model computed from a sample will vary from sample to sample, as well as differ from the population value.
- *Sampling distribution* is the data distribution of the values of sample statistic for a fixed sample size, e.g., a histogram, of the values of the sample statistic over multiple repetitions of samples of the same size. The most well-known sampling distribution is that of the sample mean which has the classic bell curve. [Figure 4.6](#), left, shows the sampling distribution of a 100 sample means, each based on a sample of size 100. A formal probability-based definition of sampling distributions is presented in [Chapters 5](#) and [6](#).

- *Margin of error* is the interval of the sampling distribution that is likely to contain a high proportion of values of the sampling statistic. The boundaries of this interval are called *error bounds*. A formal probability based definition of error bounds and margin of error is presented in [Chapters 5](#) and [6](#).
- *Statistical Bias* is the difference between a statistic computed from a sample and the population value. It is desirable that as the sample size increases, the descriptive statistics computed from the sample should be closer to the population values, and the statistical bias of the descriptive statistic should approach 0.

In [Chapters 5](#) and [6](#), we will learn the probability-based explanation of the bell-shaped distribution of sampling means and use it to compute error bounds and confidence intervals.

4.5 Bias and Ethics of Sampling

Sampling involves choosing subsets. It is possible to choose the subset to reflect certain desired outcomes. For example, in order to determine the most popular sport on campus, the soccer captain could just sample everyone in the team. What would happen to the outcome? Sometimes, the consequences of bias in sampling could be quite serious, more than just propping up soccer as the most popular sport on campus. Choosing samples from poor countries to conduct clinical trials for new drugs leads to under-reporting of side effects, because the test subjects do not want to lose access to the free medicines. Once the medicine is released to the public it could have serious side effects and the medicine could be recalled. Or, a social media or television channel

can show selective content of a certain kind to mold public opinion. When samples are biased, the AI applications that use the sample data will also exhibit bias. For instance, an AI algorithm trained only on male resumés will disregard female-seeming resumés. We will discuss the consequences of using biased data for training AI in [Chapter 10](#). Here are some examples of how samples might be biased.

Example 4.5.1.  **Selection Bias:** *Suppose an experiment was conducted to find the distribution of sizes of fish in a lake. The net that was used to catch a representative sample of fish had a mesh size of 1 cm so that fish smaller than 1 cm would not be represented in the sample. This is an example of selection bias in sampling where only the big fish make it into the sample. Check  for details.*

Note that *selection bias* is due to the way the sample is selected. It is different from statistical bias which refers to the bias in the descriptive statistic with respect to the population value. Selection bias would manifest as statistical bias that cannot be fixed with larger sample sizes.

Example 4.5.2.  **Response bias:** *There is a food festival on campus and you want to operate a taco stall to raise funds for your soccer team. You give the students who buy tacos a little card to fill out with a box to tick against a  or . The students who love the tacos don't bother to fill it out; they just walk off with the food enjoying their snack. But some students complain that they got too little filling or that the taco was not fresh. Such students do fill out the card and tick the  box. What do you think about the responses you collected?*

This is called *response bias* since typically unhappy customers respond to surveys with greater frequency than happy customers.

Example 4.5.3 Unconscious bias: *It is your parents' 35th wedding anniversary and you are organizing a surprise celebration to include your large extended family. You borrow your sibling's guest list of relatives that she has made for her wedding planned for later that year. After all, you share the same grandparents, aunts, uncles, cousins and other extended family. What you don't realize is that she has dropped your Cousin Kate and her family from the guest list. You go ahead and use her guest list, and then your father uses your list for your engagement celebration and so on. Cousin Kate and her family are now excluded from numerous family celebrations without your being aware of it or intending to. People fault her for not showing up (unaware that she wasn't on the guest list to start with) and drop her from all future family events.*

Unconscious bias is more insidious than overt bias because it frequently goes undetected. It might seem innocuous at face value but accounts for many unintended yet unfair practices.

ReferenceBox 4.5.1: Bias in samples

- *Selection bias* refers to a bias built into the sample during the sampling process through judgment or for convenience, and it results in conclusions that are not representative of the population of interest.
- *Response bias* refers to the bias in responses on surveys, questionnaires, polls, focus groups and one-on-one interviews, e.g., people are more likely to respond when they are dissatisfied rather than when they are happy, skewing the responses in the negative direction.
- *Unconscious bias* refers to bias that is not intentional and often goes undetected for that reason. It could be due to oversight, or even data

quality errors where data records are dropped by mistake and excluded from the sample. In [Chapter 8](#) we will learn about two famous economists that unintentionally wreaked havoc on the economies of some European countries by inadvertently dropping 15 rows of data from an Excel spreadsheet.

Bias in sampling, training data, selection of experimental subjects, or while framing questions for responses, can all lead to conclusions that are not representative of the population of interest. While it is not possible to avoid all bias, and sometimes it is unconscious, it is important to vet samples and data as much as possible.

4.6 Read & Reflect

- **Doing more with less:** *Young adults and higher education.* 
- *PEW research short reads: When surveying small populations, some approaches are more inclusive than others.* By Anna Brown. 
- *Lab pulls microscopic details from Southern Resident Killer Whale samples. (Research supported by a grant from the National Fish and Wildlife Foundation's Killer Whale Research and Conservation Program.)* 
- *Embarking on Research to Study Pantropical Spotted Dolphins,* NOAA. 
- *NOAA's Ocean Exploration brochure on Deep Sea Sampling.* 
- *Visual Inspection AI: a purpose-built solution for faster, more accurate quality control.* 

4.7 Exercises

1. Generating random samples.

The built-in functions in Excel and Google Sheets can be used to generate samples from given populations.

- (a) **With Excel:** This exercise will help you practice using Excel spreadsheet features to generate random samples from a given data set. You will use the CSV file Palmer Penguins data set, which is available in the Resources section  of the textbook for this exercise. Suppose you want to generate five random samples of penguins of size 10. How would you do this? How do the proportions of species in the samples compare with the population proportion of each species? Here is the link to the step-by-step instructions for completing this exercise: 
- (b) **With Google Sheets:** This exercise will help you practice using Google Sheets features and functions to generate random samples from a given data set. Generate three samples of 20 penguins from the penguin data set with replacement. Save each of these samples on a separate sheet for computing statistics. Next, generate three samples of 20 penguins from the population data set without replacement. Save these samples on a separate Google Sheet to compute statistics. Use these steps to create Random samples in Google Sheets with replacement and without replacement. 
- (c) Examine the mean, median and standard deviation of the samples generated with replacement and without replacement in part (b). What do you notice?

2. Generating proportional, cluster and systematic samples.

Use the Palmer Penguin data set as the population data for the following exercise. Save each of these samples on separate Google Sheets for computing statistics.

- (a) Choose a proportional sample of size 30 from the three species. How do the proportions of species in the samples compare with the population proportion of each species? Generating proportional samples: 
- (b) Cluster sampling involves dividing a population into clusters or groups by certain characteristics, and then randomly selecting a cluster. From the Palmer penguins population, divide the data set into clusters by species and island. Calculate the mean and median of body mass of a sample of 20 Adelie penguins from the three islands. Was there any difference in these statistics?
- (c) Choose a systematic sample of size 30, starting at a certain row number (for example, every 5th row) in the spreadsheet. How do the proportions of species in the sample compare with the population proportion of each species?
- (d) Make pie charts for species for the three samples created above from the Palmer penguins data and compare with the pie chart for the entire population. What do you notice? Can you explain?

3. How does sample size affect statistics?

In this exercise we will select random samples of size 10, 30 and 100 of the species Adelie from the Palmer Penguin data set, available in the Resources section  of the textbook. We will then calculate the mean bill length for each of these samples and compare it to the population of Adelie penguins in the data set. 

Record your calculations in [Table 4.2](#):

TABLE 4.2: Sample means of Adelie penguins from Palmer Penguin data set [📄](#)

Mean bill length	Sample size 10	Sample size 30	Sample size 100
Sample #1 mean			
Sample #2 mean			
Sample #3 mean			

- (a) Sort the Palmer Penguins data set by species and select the rows which contain the Adelie penguins. Copy and paste these rows to a new sheet.
- (b) Using the function “Randomize range” from the menu Data, randomize the data and then select a sample of 10 penguins. Since we are calculating mean bill length, you can copy the columns that contain species and bill length. Calculate the average bill length. Repeat this for two more samples. What do you notice about these three averages?
- (c) Now repeat the same process for 30 penguins. What do you notice?
- (d) Repeat once more with a sample of 100 penguins. What do you notice?
- (e) Now calculate the mean bill length for the population of Adelie penguins. How does sample size impact the statistic mean bill length?

4. Sampling methods

- (a) A student counselor in a university selects 25 freshmen, 25 sophomores, 25 juniors and 25 seniors for a sample of 100 students for a research study. What sampling method is this?
- (b) The cashiers at a grocery store hand out a survey about their shopping experience to the customers as they pay for the groceries and ask them to drop off the survey at the exit. They will receive a \$5.00 coupon for their next shopping purchase if they complete the survey. What method of sampling is this?
- (c) The water department mails out a survey about water quality to all the residents in a county and asks them to send it back. What method of sampling is this? What are some possible biases?
- (d) What is the difference between proportional and stratified sampling?
- (e) A high school principal uses student ID numbers of 9th grade students to select every 10th student to participate in a study about bullying. What method of sampling is this?
- (f) Describe a scenario where simple random sampling can be used to collect data.

5. Response bias

- (a) In trials for a new medicine discussed in [Example 4.5.1](#), what type of bias is there in the responses from people who were given the free medicine? Do their responses make the medicine look better or worse?
 - (b) What about the responses to the taco satisfaction survey discussed in [Example 4.5.2](#)? Do they make the tacos look better or worse?
6. What method of sampling will be best when testing a new drug for flu virus for adults? Explain and discuss some challenges and biases.

7. A statistical research organization wants to test the hypothesis “30 minutes of aerobic exercise three times a week can reduce the incidence of heart attacks in adults”. They decide they will send an email to all the students in a local university to complete and return a survey to collect data for this research. What are some possible biases while collecting data in this manner?
8. In this exercise, you will read about a research study that explains how sampling is used for conservation of species. Scientists from NOAA are collecting samples of genetic material to better understand pantropical spotted dolphins in the Gulf Coast. Click here to read the story and answer the question: [NOAA spotted pantropical Dolphins](#)
- (a) What method of sampling do the researchers use in this study?
 - (b) What is the population?
 - (c) How many samples are they planning to collect?
 - (d) What is the objective of this research study?
 - (e) How are the researchers ensuring the validity of the study?

9. Using AI tools to understand variability.

In this exercise, you will use the ElectricCarData CSV file in the Resources section  of the textbook to examine the variability of means in sample sets. Using the AI bot Gemini you will generate samples, calculate the means of the samples, and compare this with the population mean. (Hint: Use the resource “How to upload data files into Gemini for data analysis.” )

We will examine the variable “Range in km” of electric cars for this exercise.

- (a) Use the Gemini bot to generate 5 sample sets of size 10 from the data set.
- (b) Ask Gemini to calculate the mean of each sample set. List the samples and the means.
- (c) Calculate the mean of the sample means.
- (d) Ask Gemini to calculate the mean of the entire data set. Compare this with the sample means. Describe what you find.
- (e) Ask Gemini to generate 5 random sample sets of size 40 and the sample means. List the samples and the means. Compare the means to the population mean. Describe what you notice.
- (f) Verify the results using a calculator!

10. **Sampling with and without replacement.**

In this exercise you will investigate the effect of sampling with replacement and without replacement on descriptive statistics. The data you will use is women's tennis champion Serena Williams' total number of aces per year from 2010 to 2020. This is your population data. Using the AI bot Gemini, you will generate samples, perform EDA, with and without replacement and then analyze the impact of the two methods of sampling on these statistics. Data: Serena Williams number of aces per year (2010-2020); aces = 248, 200, 484, 480, 452, 498, 324, 71, 170, 206, 178.

- (a) What do you notice about the data? (hint: outlier)
- (b) Ask AI bot Gemini to generate four samples of size 6 with replacement by giving Gemini the appropriate prompt (remove the outlier before sampling).
- (c) Ask Gemini to generate four samples of size 6 without replacement.

- (d) Calculate the means of each sample from part (a) and part (b). Compare the means of the samples from both sampling methods with the population mean.
- (e) Is it possible to generate a sample of size 15 without replacement? Why?

4.8 Summative Projects

Project 4.8.1 Sampling Exploration with M & M

In this exploration you will understand the concept of sampling with a hands-on activity. You will differentiate between sample and population, understand the necessity for sampling, create a frequency distribution, and make predictions and understand sample variability.

The samples are fun size M & M packs, and you will create a frequency distribution of flavors in your sample. You will then make a hypothesis about the flavors based on your sample.

- 1. Your sample is the funpack of M & M. Complete [Table 4.3](#) and also calculate the percentage of each color in your funpack.*
- 2. What is the hypothesis based on your sample?*
- 3. Form a group with four other students and pool your data. What is the pooled data hypothesis? What do you observe about the pooled data? Create a table for recording this data similar to [Table 4.3](#).*
- 4. Compare your data with the 1-lb, 2-lb and 3-lb bags of M & M. What do you notice?*

5. Mars, Inc., is quite precise about the percentages of colors for the different M & M products, and the percentages are different for the various products. For plain M & Ms, according to Mars Inc., the color distribution for plain M & Ms in 2023 is: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown. Compare this data with your data and the pooled data. What do you observe?
6. State your conclusion.

TABLE 4.3: Color frequency table. [📄](#)

	Blue	Brown	Brown	Orange	Yellow	Red
f						
%						

Project 4.8.2 Random Sampling and Variability Using Penguin Body Mass

For this project, you will use the Palmer Penguin data set as a population and examine the data for the body mass of the penguins. Using the option “Randomize range” from the pull-down menu Data, generate 5 samples of sizes 10, 15, 30 and 100. Refer to the resource, “How to generate random samples in Google Sheets,” in the Resources section [🔗](#). Perform an EDA on each sample and complete [Table 4.4](#). Examine the variability in standard deviation (SD) of samples of increasing sizes. You will compare the SD of the samples and compare each sample SD to the population SD.

TABLE 4.4: Sample standard deviations for different sample sizes. [📄](#)

Sample size	S #1 SD	S #2 SD	S #3 SD	S #4 SD	S #5 SD	Avg SD
10						

Sample size	S #1 SD	S #2 SD	S #3 SD	S #4 SD	S #5 SD	Avg SD
15						
30						
100						

Answer the questions:

- 1. What did you observe while examining the SD values in this table? Explain.*
- 2. What happens when the sample size increases?*
- 3. Based on the SD values, which sample size is most accurate?*
- 4. Create a histogram for each sample size SD. How does the mean SD for each sample size differ?*
- 5. What is the SD of the population? Compare it to the mean SD of various sample size means.*
- 6. What is your conclusion?*

Project 4.8.3 Oceanography: Data collection to predict sea conditions

In this project you will explore data related to sea conditions using the NOAA websites listed below. This project is adapted from the lesson on Oceanography from NOAA's website on Education. Source:  The data and research will help answer some key questions related to sea conditions: How do researchers, sailors, fishermen and others accurately forecast conditions at sea, past and present? What factors influence sea conditions? How do onshore and offshore weather conditions differ?

1. *What type of sampling is used to collect data? Is this method appropriate? Why?*
2. *To answer these questions, investigate using the following websites and collect data for the buoys listed in the data sheet in this link: [🔗](#).*
 - (a) *Current data buoy information is found on: [🔗](#)*
 - (b) *Look at the Beaufort Wind Scale: [🔗](#)*
 - (c) *Monitor National Marine Sanctuary map online: [🔗](#)*
3. *How do you classify the wind speeds at the present time based on the Beaufort wind scale for the locations of the buoys?*
4. *Select a station off the coast of NC and select “view details” at the bottom of the chart displayed, and then copy the combined chart of wind speed, wind gusts and atmospheric pressure. Explain what you see in this chart.*

5

Probability in the Age of AI

Uncertainty is a big part of our lives, whether it is the outcome of an election, a football game, a medical prognosis, or the performance of a stock. Probability is a way of quantifying the uncertainty. AI-powered applications use probability laws extensively to quantify uncertainty, make predictions, recommend decisions and provide margins of error. For example, in order to choose the best route, an AI-powered autonomous car has to assign probabilities to different routes, to the occurrence of events such as traffic congestion, accidents and roadwork, take into account the desired time window of arrival, and finally recommend a route that has the highest chance of meeting the criteria. This type of decision-making is called *probabilistic reasoning* and is the foundation of AI.

This chapter covers the fundamentals of probability distributions, their applications in AI, and their use in providing confidence guarantees with respect to sampling variability.

5.1 Uncertainty, probability and AI

Uncertainty arises in many ways. A simple example is a coin toss shown in [Figure 5.1](#) where there are two equally likely outcomes, heads or tails, but the coin can land only on one of the two faces. Until the coin lands, the outcome is unknown. AI applications like the self-driving car shown in [Figure 5.1](#), right, use *probabilistic reasoning* to integrate knowledge from different sensors, predict the trajectory of other moving vehicles, and assess risk associated with environmental factors to choose the best route as described in detail by AI pioneers Russell and Norvig [37]. The following example gives a brief summary.



FIGURE 5.1: Left: A coin toss has equal chances of two outcomes. Right: A self-driving car uses probabilities for making complex decisions. [↗](#)

Example 5.1.1 *A self-driving taxi needs to deliver a customer to the airport on time. It can leave 90 minutes ahead of time, drive within speed limits so as not to get a ticket, and deliver the customer so that the wait time at the airport is not too long. But the outcome cannot be guaranteed since there could be circumstances that could get in the way, e.g., accidents, traffic congestion en route, early flight departure, and even a meteorite strike. The taxi could leave 180 minutes ahead of time and ensure getting to the airport on time, but that would most likely entail a long unproductive wait at the airport. So weighing all the probabilities of uncertain events, including the unlikely event of a meteorite collision, the autonomous taxi will correctly choose the 90-minute plan. Explore [GeminiBox 5.1.1](#) to learn more about the use of probabilities by autonomous cars.*

Sports betting, fantasy football, and office pools all use probabilities to make predictions about uncertain outcomes. Sampling variation discussed in [Chapter 4](#) is an example of uncertainty in how well the sample estimates the true population value.

GeminiBox 5.1.1: Probabilities and Autonomous Cars [↗](#)

“How is probability used in autonomous cars?” [🗨️](#)

In this chapter, we will discuss fundamental concepts in probability and focus on their relationship to AI with real-world examples. Additional detail and further reading include *Introduction to Probability Models* by Sheldon M. Ross [[36](#)], an excellent and comprehensive introduction to probability theory and probability distributions.

GeminiBox 5.1.2: Uncertainty and Probability

“What is the connection between uncertainty and probability?” 

5.2 Events and Outcomes

Coin tosses, medical prognoses, and stock movements are examples of *events* that could have one or more *outcomes*, e.g., heads-tails, recover-remission-terminal, up-down-steady. Some events have fixed or predetermined outcomes, e.g., an object, when dropped will fall to the ground on our planet. For other events, the outcomes are *uncertain*.

The chance of correctly guessing an outcome is related to the uncertainty of an event. The greater the uncertainty, the lower the chance of a correct guess and vice versa. For *deterministic* events, uncertainty is near 0, and the likelihood of guessing the correct outcome is close to 1. In the coin toss example, heads and tails are equally likely, guessing the correct outcome has only a 50% chance of success. On the other hand, for a common cold, the likelihood of recovery is close to 99% so the chance of correctly guessing the outcome is high.

The uncertainty of the outcomes of an event is quantified through a *probability distribution* that captures the relative chances of the occurrence of each outcome. Probability distributions can range from simple, e.g., 50-50 for a fair coin toss, to complex mathematical functions like the Normal, or empirically constructed algorithms like neural networks. They are used extensively in AI applications, for example, in assisting robots to perform tasks such as cleaning and reconnaissance, as shown in [Figure 5.2](#).



FIGURE 5.2: Robots use probability distributions to recognize objects and navigate surroundings. Left: Heathrow Airport has eight ultraviolet-disinfecting (UVD) robots that roam the terminals to clean and disinfect (Mx. Granger, Wikimedia Commons). Right: A Vision 60 Q-UGV ground robot does a simulated route mission at Holloman Air Force Base, New Mexico, April 17, 2023 (U.S. Air Force photo by Airman 1st Class Isaiah Pedrazzini, Wikimedia Commons). [📄](#)

ReferenceBox 5.2.1: Probability Terminology [📄](#)

- An *event* E is any activity or phenomenon whose *outcome* is of interest, e.g., a coin toss, a sporting event or election, movement of a stock, or getting a job after graduation.
- A *Sample Space*, S , is the set of all outcomes of an event E . For example, the sample space of a coin toss $S = \{H, T\}$ since there are only two outcomes, heads (H) or tails (T),
 1. The outcomes in S are mutually exclusive, i.e., non-overlapping. The coin has to land either on heads or tails.
 2. The set of outcomes in S is *exhaustive* to include every possible outcome.
 3. Ideally, the sample space must not include irrelevant or impossible outcomes. For example, the outcome “6” is relevant for the roll of a dice but not for a coin toss.

- An event E is *deterministic* if its outcome is known in advance, or known with near certainty. For example, the sun rises every morning, a ball falls down toward the ground on our planet when dropped, and a human will die without oxygen.
- An event E is *probabilistic* if its outcome cannot be predicted with a high degree of certainty.

Events can have multiple outcomes associated with them, e.g., rolling more than a “4” on a single roll of dice. Can you determine the sample space of the roll of dice and outcomes that support the event above?

Data Debate 5.2.1 *Equally likely outcomes:* What is an example of an event that has equally likely outcomes? If you wanted to place a bet on a sporting event, would you prefer equally likely outcomes? Why or why not? In what situation would you prefer an event with equally likely outcomes?

To understand the difference between deterministic and probabilistic events with examples, explore [GeminiBox 5.2.1](#).

GeminiBox 5.2.1: Deterministic vs. probabilistic events

“What's the difference between deterministic and probabilistic events?” 

Data Debate 5.2.2 *AI and uncertain events:* Self-driving cars handle a multitude of uncertain events.

- Will a pedestrian cross the road when the light is green?
- Will the car ahead brake suddenly?
- Is there traffic congestion ahead?
- Is the road surface slippery?

Discuss the challenges faced by an autonomous vehicle.

5.2.1 Probabilities of events

In order to understand how to compute probabilities, suppose that an ice cream parlor has a raffle where you can walk in and push a buzzer, and you will get a scoop, for free, of

whatever flavor pops up. Suppose that the flavors offered by the shop are:

$S = \{\text{Chocolate, FroYo, Honey, Mango, Matcha, Mint, Rocky Road, Strawberry, Vanilla, Vegan}\}$.

Let the following subsets represent flavors liked by your friends, A, B, C, D, and E.

$A = \{\text{Chocolate, Mango, Matcha, Mint, Strawberry, Vanilla}\}$

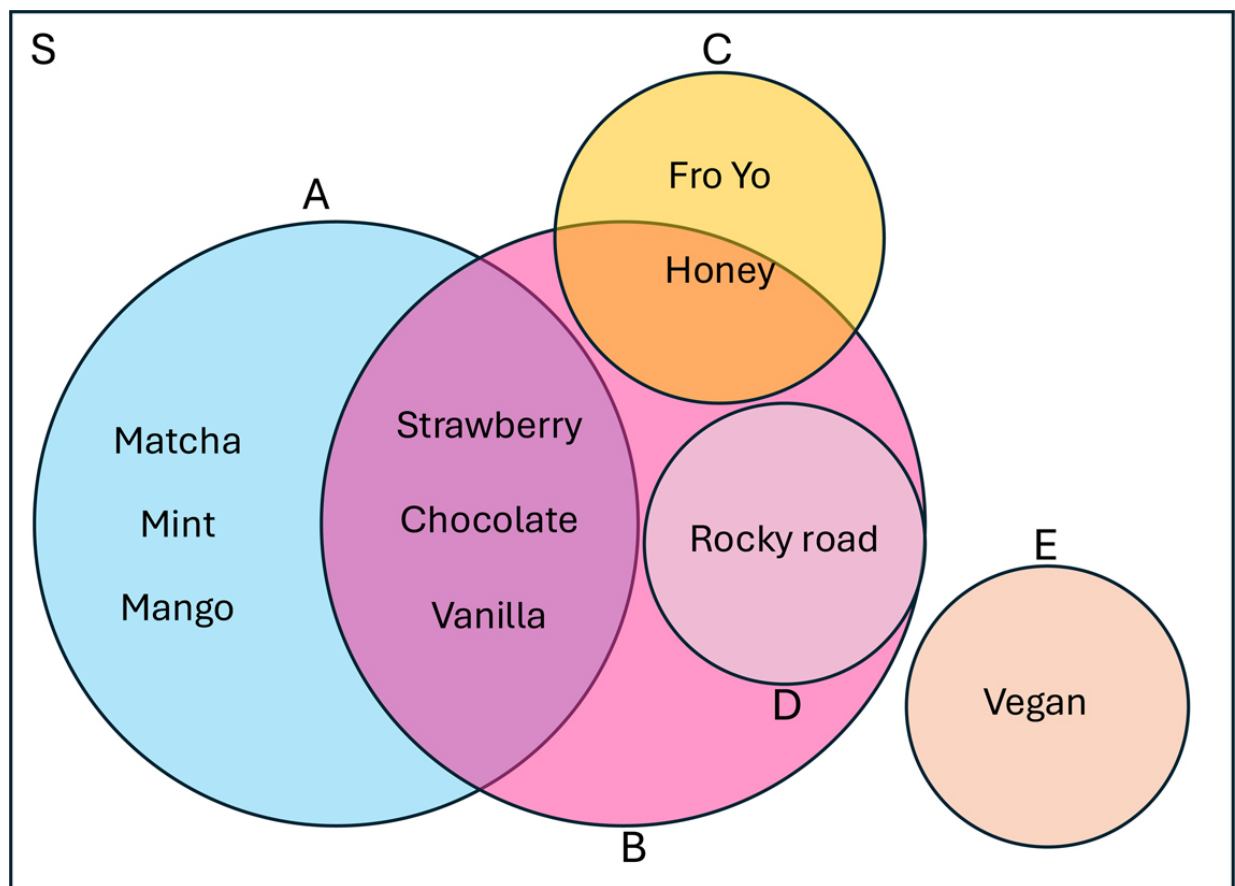
$B = \{\text{Chocolate, Honey, Rocky road, Strawberry, Vanilla}\}$

$C = \{\text{FroYo, Honey}\}$

$D = \{\text{Rocky road}\}$

$E = \{\text{Vegan}\}$

[Figure 5.3](#) shows the sets in relation to each other. Suppose that each flavor in S has an equal chance of popping up. What are the chances that your friend A will like the flavor?



[Extended Description](#)

FIGURE 5.3: Events, outcomes and the probabilities of their occurrence. [📄](#)

Instinctively, the answer is 6/10, because there are 6 flavors in set A , out of a total possible 10 flavors in sample space set S , and they are all equally likely. Therefore, the *probability* of a flavor from set A popping up is denoted by $P(A)$:

$$P(A) = \frac{|A|}{|S|} = \frac{6}{10},$$

where $|A|, |S|$ represent the cardinality of the sets A and S . [Section 5.9.1](#) in the Appendix has a brief note on set theory terminology and operations. Herbert B. Enderton's *Elements of Set Theory* [16] provides a great introduction.

The probability of two events co-occurring is given by the probability of the intersection of the set of outcomes corresponding to the two events:

$$P(A \cap B) = \frac{|A \cap B|}{|S|}, \tag{5.1}$$

For example, the probability that a flavor that pops up is liked by both your friends A and B , i.e., is in sets “ A AND B ”, is given by the probability of the set $A \cap B = \{Chocolate, Strawberry, Vanilla\}$:

$$P(A \cap B) = \frac{|A \cap B|}{|S|} = 3/10.$$

The probability that the flavor is liked by either B or C , i.e., from sets either “ B OR C ” is given by $B \cup C = \{Chocolate, FroYo, Honey, Rocky road, Strawberry, Vanilla\}$. Note that the flavor Honey occurs only once in $B \cup C$, even though it occurs in both B and C because the outcomes have to be distinct. Therefore,

$$P(B \cup C) = P(B) + P(C) - P(B \cap C) = \frac{|B|}{|S|} + \frac{|C|}{|S|} - \frac{|B \cap C|}{|S|}, \tag{5.2}$$

where $P(B \cap C)$ is subtracted from the sum of the probabilities of B and C to avoid double counting.

$$P(B \cup C) = 5/10 + 2/10 - 1/10 = 6/10.$$

Review [ReferenceBox 5.2.2](#) thoroughly and work through the example in [Section 5.9.1](#).

ReferenceBox 5.2.2: Probabilities of Events [↗](#)

Probability is an objective way of quantifying the likelihood of the occurrence of an event E . In the following definitions, we will **represent an event E by the set of its outcomes A** .

- If $|S|$ is the cardinality of the sample space S , and $|A|$ is the cardinality of the set of outcomes that support an event, then the probability of that event is given by:

$$P(A) = \frac{|A|}{|S|}.$$

- The probability of an empty set ϕ is 0.

$$P(\phi) = \frac{|\phi|}{|S|} = 0/|S| = 0.$$

- Probability of the sample space S is 1.

$$P(S) = \frac{|S|}{|S|} = 1.$$

- Probability of two events, A AND B , co-occurring is given by:

$$P(A \cap B) = \frac{|A \cap B|}{|S|}.$$

- The probability of at least one of two events, A OR B , happening is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{|A|}{|S|} + \frac{|B|}{|S|} - \frac{|A \cap B|}{|S|}.$$

5.3 Probability Distributions

Suppose that it is summer and that you are a frequent visitor to that ice cream parlor we mentioned earlier. On one visit, you might choose chocolate; on another, you might choose

strawberry. However, you are more likely to choose some flavors over others. You might choose chocolate on 80 visits, strawberry on 15, and mocha on 5. Can you figure out the flavor you are most likely to pick? The one you are least likely to pick?

5.3.1 Random variables

Your choice of ice cream flavor is a *random variable* whose value varies from visit to visit. However, you can associate probabilities with your flavor choices by examining the history of your choices during your past visits to the ice cream parlor. The probabilities of outcomes of a random variable can also be governed by mathematical functions.

ReferenceBox 5.3.1: Random Variables and Probabilities

- A *random variable* X represents a variable whose value is uncertain and is determined by the outcome of a probabilistic event, e.g., choice of a flavor of ice cream.
- The *domain* D of a random variable is the set of all possible outcomes. D is said to be *exhaustive* since it includes all possible values of X . The domain is related to the sample space S discussed in [ReferenceBox 5.2.1](#). If X is associated with a single event E , then D and S are the same.
- The values of X and the associated probability p of occurrence of the outcomes that result in that value of X together constitute the *probability distribution* of X .
- *Probabilities range* between 0 and 1 since they represent the chances of a particular outcome. A value of 0 means that an outcome will never occur, while a value of 1 implies that an outcome is certain.
- *The sum of probabilities* associated with all the possible values of the random variable X must sum to 1, since D is exhaustive and includes every possible outcome.
- The probability that X takes a particular value can be determined from observed data. For example, the probability that you will pick chocolate ice cream is 0.8. Such a probability distribution based on empirical observation is called an *empirical probability distribution*.

- In some instances, if the nature of the process that determines the outcome is well known, e.g., that of a fair coin toss, the probability can be expressed theoretically by a mathematical function without the need for empirical observation.

Data Debate 5.3.1 *If X is the random variable corresponding to your friend B's flavor choice, what is the domain D of X ? Is it the same as the sample space S ?*

5.3.2 Probabilities of random variables

The values of a random variable X and the probability of occurrence of those values together are called a *probability distribution* of X . If X is a continuous variable, it is called a *probability function*. For example, the probability distribution of the random variable X that represents your choice of ice cream flavors (mentioned at the beginning of this section) is:

$$P(X = \text{Chocolate}) = 80/100 = 0.8 \quad (5.3)$$

$$P(X = \text{Strawberry}) = 15/100 = 0.15 \quad (5.4)$$

$$P(X = \text{Mocha}) = 5/100 = 0.05 \quad (5.5)$$

$$P(X = \text{OtherFlavors}) = 0 \quad (5.6)$$

Other Flavors (OF) represents flavors other than strawberry (S), chocolate (CC) and Mocha (M). Such an estimate computed from data gathered through observation or experimentation is called an *empirical probability distribution*. Here are some interesting properties of probabilities:

- If an outcome is certain, then the probability of that outcome for that event is 1. That event is deterministic and not random.
- If there is no chance of a particular outcome occurring, e.g., going home without buying ice cream, then the probability of that outcome is 0, the smallest value a probability can

take.

- The sum of probabilities of all the values in the domain of a random variable X , e.g., choice of ice cream flavors, is 1. This is because the domain of X must by definition include every possible outcome and therefore every value of X .

Data Debate 5.3.2 *Is it meaningful for a probability to be negative? Why or why not?*

ReferenceBox 5.3.2: Types of Probability Distributions

- The *probability distribution* of a random variable X is the collection of probabilities of occurrence of its values. It is denoted by:

$$P(X).$$

Equation 5.3 represents the probability distribution of your choice of ice cream flavors. Probability distributions and their estimation are key to data science and AI.

- The *joint probability distribution* of two random variables X and Y specifies the chances of the co-occurrence of the set of all possible joint outcome values of X and Y . It is denoted by:

$$P(X, Y).$$

For example, [Figure 5.5](#) shows the joint probabilities (in black) of ice cream choices of you and your friend, inside the cells of the table. [Section 5.3.3](#) describes joint probability distributions in detail.

- The *marginal probability distribution* of a random variable Y is simply its probability distribution irrespective of the values of other random variables, e.g., your flavor choice irrespective of what your friend chooses. It is described in [Section 5.3.4](#), denoted in red in [Figure 5.5](#).
- The *conditional probability distribution* of a random variable Y is its probability distribution given a specific value of another random variable. For example, probabilities of your friend's flavor choice, given that you chose CC, is shown in [Figure 5.5](#), first row, in green.

5.3.3 Joint probabilities

Suppose you visit the ice cream parlor with your friend whose choice of flavors is represented by the random variable Y . Equation 5.7 shows both your choices on 100 visits. The notation, for example, “ $X=CC$ & $Y=Mint$ ” means that you chose chocolate ($X=CC$), while your friend chose mint ($Y=Mint$).

$$\begin{aligned}P(X = CC \quad \& \quad Y = OF) &= 0.5 \\P(X = S \quad \& \quad Y = S) &= 0.1 \\P(X = CC \quad \& \quad Y = CC) &= 0.1 \\P(X = CC \quad \& \quad Y = Mint) &= 0.2 \\P(X = S \quad \& \quad Y = OF) &= 0.05 \\P(X = Mocha \quad \& \quad Y = Mint) &= 0.05\end{aligned}$$

where OF represents flavors other than strawberry (S), chocolate (CC), Mocha and Mint.

The probability distribution of the co-occurrence of values of X and Y is called the *joint distribution* of X and Y denoted by $P(X, Y)$. See [Figure 5.5](#) where each cell represents the joint probability (shown in black) of the flavor choices of you and your friend. Confirm that the probabilities have values in the interval $[0,1]$ and add up to 1.

Data Debate 5.3.3 *Is it possible to calculate the probability distribution of X shown in [Section 5.3.2](#) from the joint probabilities in [Section 5.3.3](#)? How about that of Y ?*

Example 5.3.1 Joint distributions and AI in sports *The sports industry has embraced probabilities and AI in a big way in all aspects of sports such as fan engagement; scouting, training and performance of athletes, shown on the left in [Figure 5.4](#); and sports analytics and strategies. For example, how will a particular batter perform against a particular pitcher on a particular baseball field? While the sports industry has always relied heavily on statistics, e.g., *Moneyball* by Michael Lewis [[27](#)], the use of AI has taken it to a completely different level. Read the article *The Game Changer: How AI Is Transforming The World Of Sports Gambling* in [Section 5.6](#). According to a *Forbes* article in [Section 5.6](#), AI is the new MVP in sports betting, since it can compute these probabilities in real-time with up-to-date player statistics.*



FIGURE 5.4: Left: What is the probability that a player will score during a free throw? (Photo from Wikimedia. Michael Barera. Orlando Magic vs. Milwaukee Bucks December 31, 2024). Right: Google's *DolphinGemma* uses conditional probabilities to predict dolphin communications. (Wild Atlantic spotted dolphins. Ricardo Liberato, CC BY-SA 2.0 via Wikimedia Commons.) [↗](#)

		Your Friend (Y)					
		Chocolate Chip	Strawberry	Mint	Mocha	OF	
You (X)	Chocolate Chip	0.1 (0.125)	0 (0)	0.2 (0.25)	0 (0)	0.5 (0.625) (0.91)	0.80 (1.0)
	Strawberry	0	0.1	0	0	0.05 (0.09)	0.15
	Mint	0	0	0	0	0 (0)	0.00
	Mocha	0	0	0.05	0	0 (0)	0.05
	OF	0	0	0	0	0 (0)	0.00
		0.1	0.1	0.25	0	0.55 (1.0)	1.00

[Extended Description](#)

FIGURE 5.5: Probability distributions of choices of ice cream flavors for X and Y : Joint (black), conditional $P(Y|X = CC)$ (green), marginal of X and Y (red), and conditional updated using Bayes' rule $P(X|Y = OF)$ (highlighted in yellow). [↗](#)

To learn more about how AI and probability theory are used in sports, explore the prompts shown in [GeminiBox 5.3.1](#).

GeminiBox 5.3.1: AI in Sports

“How is AI used in Sports? How is probability theory applied in sports AI?” 

5.3.4 Marginal probabilities

The probability distribution of a random variable Y , irrespective of that of other random variables, is called the *marginal distribution of Y* . When we refer to the probability distribution of a random variable, it is *by default* its marginal distribution since we ignore the probability distribution of its co-occurrence with other random variables. The bottom row of red probabilities in [Figure 5.5](#) represents the marginal probability distribution of Y , your friend's ice cream flavor choice, and the last column of red probabilities represents the marginal probability distribution of X , your ice cream flavor choices.

5.3.5 Conditional probabilities

Is your friend's ice cream flavor choice affected by yours? This is an important question in marketing where taste-makers and influencers are hired to pump up sales. Put another way, what flavor would your friend choose given that you chose chocolate chip? This is called a *conditional probability distribution* because it refers to your friend's choice conditional upon your choice. It is possible that your friend would choose a different flavor so that the two of you can swap halfway through and enjoy both flavors.

In the instance shown in [Figure 5.5](#), you chose chocolate on 80 visits (marginal distribution of 0.8 in red), and your friend chose chocolate on 10 of those 80 visits (joint distribution of CC and CC in black) when you chose chocolate. Therefore, the probability that your friend will choose chocolate given you chose chocolate is $0.1/0.8 = 0.125$ shown in green. See [Appendix 5.9.2](#) for the full conditional distribution of your friend's choice Y given you chose CC.

Conditional probability of “ Y given X ” can be written in general as:

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(X \cap Y)}{P(X)}, \quad (5.7)$$

where $P(X, Y) = P(X \cap Y)$ is the joint distribution of X and Y , the probability of the co-occurrence of values of X and Y . $P(X \cap Y)$ is the same as $P(Y \cap X)$ since it represents the outcomes that are common to both X and Y .

Conditional probabilities are used extensively in AI, e.g., by autonomous vehicles, by lunar landers like the Blue Ghost described in [Figure 5.7](#), and even for predicting dolphin communications as shown in [Example 5.3.2](#).

Example 5.3.2. [🔗](#) **Chatty dolphins, conditional probabilities and AI:** Google's *DolphinGemma* uses probabilities to predict the next sound in the complex language of clicks and whistles used by dolphins, shown on the right in [Figure 5.4](#). It can even generate dolphin-talk. You can read more about *DolphinGemma* in [Section 5.6](#). Explore how probabilities are used by *DolphinGemma* using the prompt in the [GeminiBox 5.3.2](#).

GeminiBox 5.3.2: Talking with dolphins using probabilities and AI [🔗](#)

“Does DolphinGemma use probabilities?” [🔗](#)

5.3.6 Bayes' rule

The conditional and marginal probability distributions of random variables are related by a famous rule called *Bayes' Rule*:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (5.8)$$

which, using Equation 5.20, can be simplified to:

$$P(Y|X) = \frac{[P(X \cap Y)/P(Y)]P(Y)}{P(X)} = \frac{P(X \cap Y)}{P(X)}. \quad (5.9)$$

Similarly,

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}. \quad (5.10)$$

Bayes' Rule plays a big role in machine learning and in updating hypotheses when new information becomes available. For example, if we know that your friend chose the flavor *OF*, we can make a more accurate guess as to what you had chosen. Using Equation 5.10 and the data in [Section 5.3.3](#), the probability distribution of your flavor choice given your friend's choice is given by:

$$\begin{aligned} P(X = CC|Y = OF) &= 0.5/0.55 \\ &= 0.91 \\ P(X = S|Y = OF) &= 0.05/0.55 \\ &= 0.09 \end{aligned}$$

Note that the fact that your friend chose the flavor *OF* implies an increased likelihood of 0.91 that you had chosen chocolate compared to the marginal probability of 0.8 described in [Section 5.3.2](#). The updated probabilities of your choice of ice cream, after your friend's choice is made, are shown in [Figure 5.5](#), highlighted in yellow. They are called *posterior probabilities* since they reflect the updated probability after a particular event has occurred. For example, given that 90% of the traffic is flowing through a side street as opposed to the usual 10%, what is the posterior probability that an accident occurred on the highway? Would it be higher or lower than if you had not had the information about the 9-fold increase in traffic along the side street?

Data Debate 5.3.4 *Does the conditional probability distribution of your flavor choices given that your friend chose OF, as shown in Equation 5.11, meet the sanity check of a probability distribution? What if your friend had chosen mint, would the conditional probability of that distribution meet the criteria of a probability distribution?*

Example 5.3.3 Robots and conditional probabilities in AI: *Robots use conditional probabilities to estimate their location, to navigate around objects, and to choose actions based on the most current sensor reading. See [Figure 5.2](#) for two great examples. Robots are used to clean airports, to load and unload objects, for reconnaissance purposes, and even to deliver food. Read the article Robots in Everyday Life in [Section 5.6](#).*

The game show *Let's Make a Deal* originally hosted by its creator Monty Hall is an excellent illustration of Bayes' Theorem. Contestants have to choose one of three closed doors, one of which conceals a car and the other two conceal goats. A contestant chooses a door, and Monty opens another door. If that reveals a goat, the contestant is given the

option of sticking with the original choice or switching doors. What do you think would be a good strategy? Use [GeminiBox 5.3.3](#) to explore.

GeminiBox 5.3.3: Monty Hall and Bayes' Theorem

“Bayes' rule three-door problem.” 

Bayes' rule is widely applied in AI, particularly in Machine Learning. A popular algorithm that uses Bayes' rule is called the *Naive Bayes Classifier*. It is trained on data with labels and associated attributes, and creates patterns of attributes that co-occur with label values with a high probability. When new information becomes available, the attribute patterns are used to predict the labels. This is known as *supervised learning* since the labels are provided to train the model.

5.3.7 Independent events

Dependence between random variables can be quantified. Intuitively, if two random variables X and Y are completely independent, we would expect the conditional probability of $P(Y|X)$ to be the same as the marginal distribution of $P(Y)$ because we would expect the values of X to have no influence on the values of Y . For example, the number of letters in a person's first name and their ice cream flavor choice. Which means that:

$$\begin{aligned} P(Y) &= P(Y|X) \\ &= \frac{P(Y \cap X)}{P(X)} \implies \\ P(Y \cap X) &= P(X)P(Y) \end{aligned}$$

The first implication is due to Equation 5.20, and the second is due to simple algebra. Since $P(Y \cap X)$ is the same as $P(X \cap Y)$,

$$P(X \cap Y) = P(X)P(Y). \tag{5.11}$$

This is the *multiplicative rule of probabilities* of independent events. Equation 5.11 provides a simple way to test the independence of two random variables X and Y , where $P(X)$ and $P(Y)$ are the marginal distributions and $P(X, Y)$ is the joint distribution.

Example 5.3.4 *If your choice of ice cream flavor (X) was independent of your friend's (Y), then Equation 5.11 has to be true for EVERY pair of values of X and Y . Can you establish or contradict this from the table of probabilities in [Figure 5.5](#)? Hint: Consider X =Strawberry and Y =Strawberry.*

A simple hands-on example of how probabilities in general, and Bayes' Rule and the multiplicative rule of probabilities of independent events are used to build a simple AI application, the Naive Bayes Classifier is demonstrated in [Chapter 9](#).

5.4 Common Probability Distributions

There are several ways to define a probability distribution. Two common ways are through empirical observation and through compact mathematical functions. Explore the prompt in [GeminiBox 5.4.1](#) to learn more.

GeminiBox 5.4.1: Probability Distributions

“What are the different ways of determining a probability distribution?” 

5.4.1 Empirical distributions: learning from data

Empirical probability distributions are based entirely on observed data rather than pre-defined theoretical formulations. [Section 5.3.2](#) is a good example of an empirical frequency distribution of ice cream flavors based on the observation of 100 visits to the ice cream parlor. This approach is essential because, in many real-life scenarios, the distribution of a random variable is unknown. In [Chapter 3](#), we learned about frequency distributions, counts, proportions, bar charts ([Section 3.8.2](#)) and histograms ([Section 3.8.3](#)). These are the foundations of empirical probability distributions. [Figure 5.5](#) is another example of an empirical distribution.

If a random variable X has values $a_1, a_2, \dots, a_k$ with frequencies $n_1, n_2, \dots, n_k$ respectively in $N = \sum n_i$ instances of the observation of X , then the empirical probability, $\hat{p}_i$ that X will have the value a_i is given by:

$$P(X = a_i) = \frac{n_i}{\sum_i n_i},$$

the proportion of the value a_i in N instances of X .

5.4.2 Probability distributions and AI

AI relies heavily on probability theory and distributions. Empirical distributions are used extensively by AI applications. Probability distributions underlying real-world phenomena are often unknown or too complex to capture accurately with predefined mathematical models. Some examples of the ways in which AI uses empirical distributions are:

- *Neural networks* use empirical distributions learned entirely from data to approximate complex probabilistic distributions of hundreds of attributes. For example, neural networks can remove noise from an image as shown in [Figure 5.6](#) by learning the noise distribution, and use that knowledge to de-noise images. The figure in the middle is the noisy image with a grainy background. After de-noising, the background is no longer grainy, but the spotted texture on the bird, e.g., its wings and head has been smoothed out as well when compared to the original image on the left.
- *AI language models* such as Natural Language Processing (NLP) models are trained to learn the empirical probability distribution of sequences of words from data such as articles and books to mimic human language. We will discuss NLPs briefly in [Chapter 9](#). Large language models (LLMs) that are behind the current explosion in the use and integration of AI into work productivity tools are powered by empirical probability distributions computed from vast corpuses of text.
- AI uses empirical distributions to create baseline distributions and identify low probability events with respect to this distribution as *anomalies*. For example, the interconnectedness of sensors in power grids is modeled using empirical distributions and low probability occurrences, e.g., power surges and outages are detected as departures from the normal baseline patterns of the power grid. Read an *MIT News* article, *Using artificial intelligence to find anomalies hiding in massive datasets: pinpointing potential power grid failure* on this topic in [Section 5.6](#).

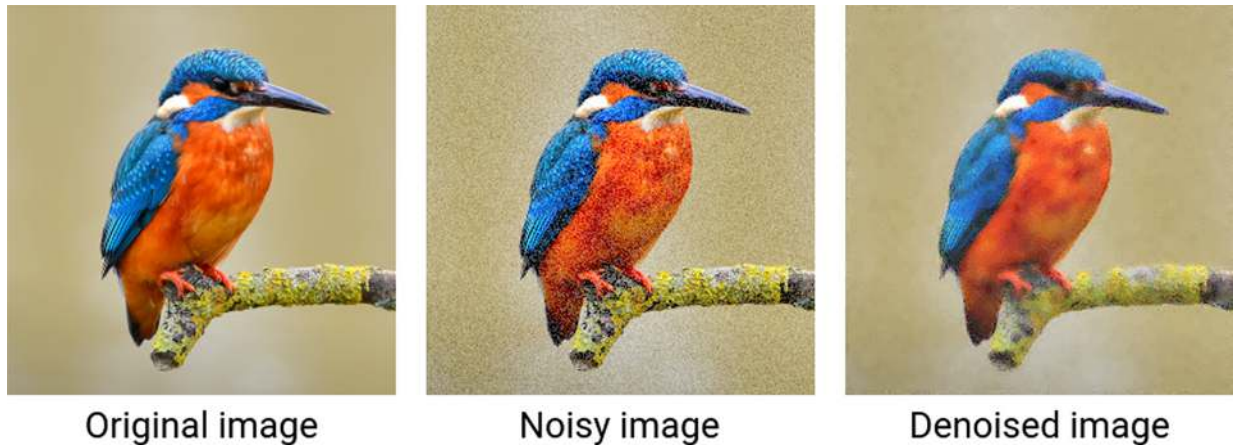


FIGURE 5.6: Empirical distributions are used by neural nets to learn noise distributions to de-noise images. Photo credit: Kriomet, CC0, via Wikimedia Commons. [🔗](#)

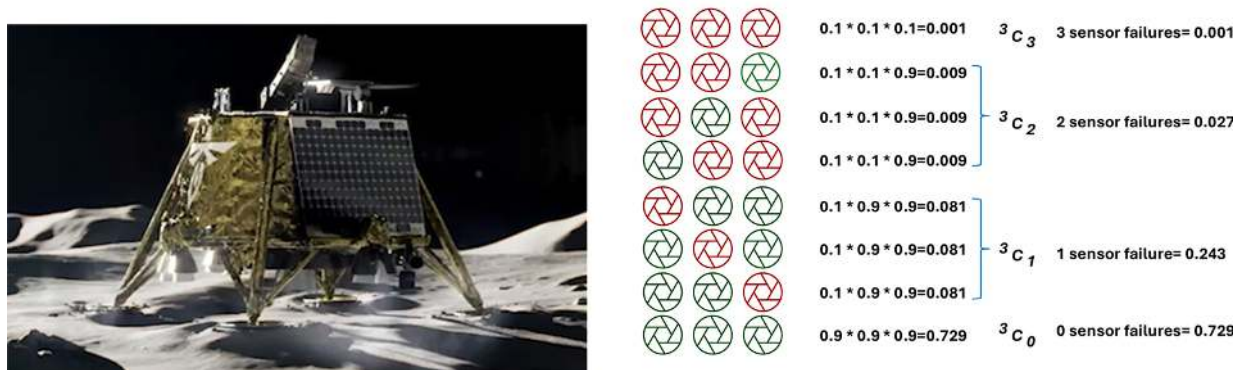


FIGURE 5.7: Blue Ghost, a lunar lander from FireFly and NASA used sensors to stick its autonomous moon landing on the first attempt without toppling over. Right: Binomial probabilities of hypothetical sensors failing. [🔗](#)

5.4.3 Theoretical probability distributions

Random variables can be discrete or continuous. Probability distributions too are discrete or continuous depending on the nature of the random variable they represent. Mathematical equations are often used to express probability distributions and functions associated with random variables. Such theoretical formulations are compact, and when applicable, enable the derivation of important properties of the distribution without the help of massive amounts of data and computation. Prior to the increase in computing power, theoretical results were the mainstay of probabilistic approaches. We will review one discrete (Binomial) and one continuous probability distribution (Normal) as examples. *Introduction*

to *Probability Models* by Sheldon M. Ross [36] has a comprehensive discussion of probability distributions.

5.4.4 Illustrative example: Blue Ghost and the Binomial distribution

Recently, a lunar lander called [Blue Ghost](#) made a perfect *autonomous* landing on the moon in its first attempt, a rarity, and collected valuable data for research. A rendering of the lander is shown in [Figure 5.7](#). In order to do so, Blue Ghost had to use many sensors to gather data about the landing surface and use AI to find a proper spot and trajectory so as not to topple over and jeopardize its mission.

The Blue Ghost had four sensors on each of its footpads but ignored the data from one of them and still landed successfully. Suppose, hypothetically, that a three-legged lander needs at least two of its three sensors to successfully function in order to avoid smashing into the surface of the moon. Each sensor functions independently of the other and has a 10% chance of failing at any given time.

The probability distribution of the number of functioning sensors is shown in [Figure 5.7](#), right. Sensors in red denote malfunctioning sensors, while those in green are functional. The first row represents the probability of all three sensors failing (red) independently of each other, i.e., $0.1 \times 0.1 \times 0.1 = 0.001$ using Equation 5.11 of probabilities of independent events.

The next three rows represent the probability of exactly two sensors failing, given by $0.1 \times 0.1 \times 0.9 = 0.009$ for each row. There are three such scenarios, since sensor pairs (1,2) or (1, 3) or (2, 3) could fail, leading to exactly two sensors failing. (The number of combinatorial scenarios of choosing exactly two from three is given by $\binom{3}{2}$.) Therefore, the probability of exactly two sensors failing is:

$$\binom{3}{2} * 0.009 = 3 * 0.009 = 0.027.$$

The next three rows represent the probability of exactly one sensor failing ($\binom{3}{1}$), and finally, the last row represents the probability of none of the sensors failing (all green) ($\binom{3}{0}$).

The lander will be successful as long as none of the sensors fail (probability = 0.729), or exactly one sensor fails (probability=0.243). Therefore, the probability that the lander will succeed is $0.729 + 0.243 = 0.972$, using Equation 5.2 and noting that “exactly one sensor” and

“none of the sensors” are mutually exclusive events, and the probability of their intersection is 0.

Lunar landers (and autonomous cars) have hundreds of sensors, and it is important to compute and monitor the probability that there are no mission critical failures. The Binomial distribution is widely used in such circumstances.

5.4.5 Binomial distribution

An event that has only two outcomes is called a *binary event*, e.g., a coin toss. The probability distribution of a random variable with binary outcomes, say 0 or 1, is given by:

$$P(X = 0) = p \tag{5.12}$$

$$P(X = 1) = q \tag{5.13}$$

$$= 1 - p \tag{5.14}$$

Since there are only two outcomes, one of which must occur, the sum of the probabilities of the two outcomes is 1. That is:

$$P(X = 0) + P(X = 1) = 1 \implies q = 1 - p.$$

In general, a random variable X which represents the number of outcomes of type 0 (or 1) of n independent binary outcomes, all with the same probability, can be represented by a Binomial distribution. The probability of k of n outcomes being 0, where each has an independent probability p of being 0, is given by:

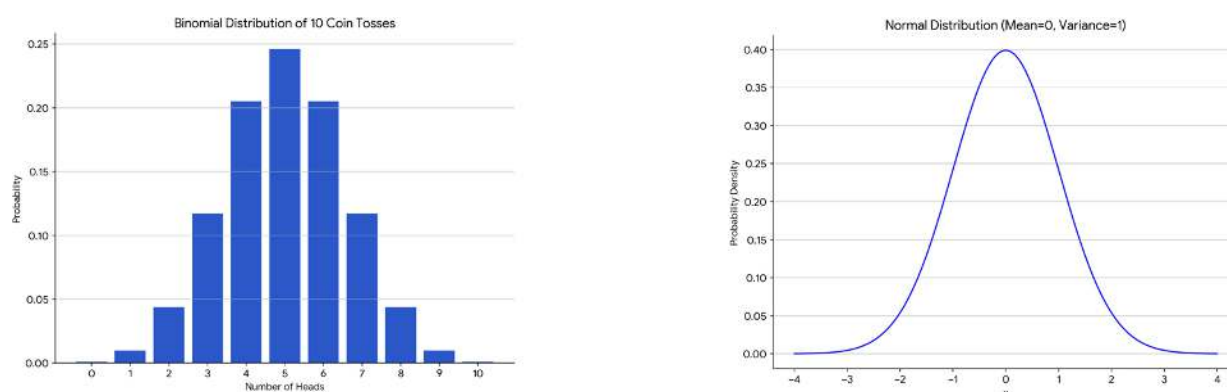
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{5.15}$$

where X is the random variable that represents the number of 0's.

In the above hypothetical example of Blue Ghost, 0 represented the outcome that a sensor was functioning (not failing), and the Binomial probabilities were computed for

various numbers of the three sensors failing.

The Binomial distribution is critical to many AI applications where independent binary inputs of numerous AI-agents are combined by complex probabilistic reasoning models to predict the next state of a system, and recommend an action, e.g., autonomous cars. The distribution has a nice symmetric shape when the two outcomes are equally likely as in a coin toss. [Figure 5.8](#), left, generated using the prompt in [GeminiBox 5.4.2](#), shows the distribution of the number of heads in 10 coin tosses where the probability of heads= $1/2$ for any given toss.



[Extended Description](#)

FIGURE 5.8: Plots generated by Gemini AI using the prompts shown in [GeminiBoxes 5.4.2](#) and [5.4.3](#). Left: A Binomial distribution, a classic example of a discrete probability distribution, depicts the probability of number of heads in 10 tosses of a fair coin with equally likely outcomes. Right: The Normal (Gaussian) distribution with mean 0 and variance 1. The Normal distribution is one of the most widely used distributions. [↗](#)

GeminiBox 5.4.2: Binomial Distribution [↗](#)

“Plot a Binomial distribution of 10 coin tosses where probability of heads= $1/2$.” [🔗](#)

GeminiBox 5.4.3: Normal Distributions [↗](#)

“Plot a Normal distribution with mean 0 and variance 1.” [🔗](#)

Examples of other discrete distributions include:

- Poisson distribution used to quantify the probability of occurrence of number of discrete events, e.g., number of vehicles crossing an intersection in a given time interval.

- Multinomial distribution, the multi-outcome equivalent of the Binomial distribution.

Please refer to [36] for detailed descriptions.

5.4.6 Illustrative example: Hurricanes and the Normal distribution

Hurricane forecasts are some of the most useful statistical data. They are closely watched to protect human lives by issuing evacuation orders, and to contain damage to property through preventative measures, e.g., deploying storm shutters and sandbags. [Figure 5.9](#), left, created by Nilfanion, shows storm tracks superimposed over a NASA earth image. [Figure 5.9](#), right, shows the projected path of Hurricane Earl. The white bubble represents the uncertainty in the hurricane trajectory factoring in a complex set of dynamic conditions like vertical wind shear, humidity and other complex factors. Read the article *Determining uncertainty: a review of hurricane intensity predictability* published by NOAA in [Section 5.6](#).

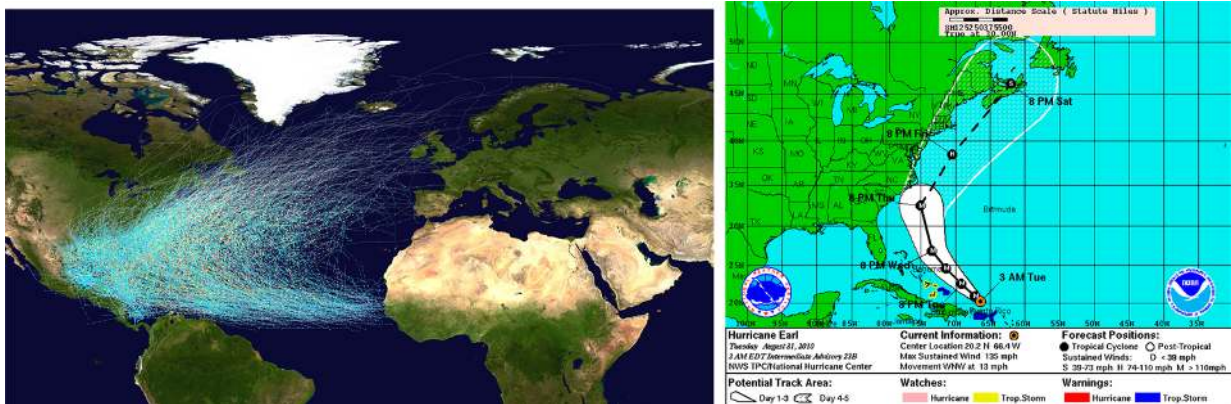


FIGURE 5.9: Left, tracks of all Atlantic hurricanes which formed between 1851 and 2019. The points show the locations of the storms at six-hourly intervals. (Image credit: Nilfanion). Normal distributions are used to model hurricane tracks and to predict potential future trajectories (right) shown by the white area around the storm track. (NOAA.) [↗](#)

Continuous probability distributions are used extensively to estimate uncertainty in storm trajectories and in many other situations that involve quantifying uncertainty. While there are many continuous probability distributions, the Normal or Gaussian distribution is one of the most widely used.

5.4.7 Normal probability distribution

If X is a continuous random variable that follows a Normal distribution, then the domain of X is the entire real number line, $[-\infty, \infty]$, but in general 99% of the values are clustered in a relatively small interval of six standard deviations centered on the mean, giving rise to the famous bell-shaped curve that determines the way the data is distributed across the number line. Specifically, the shape of the Normal distribution centered on mean μ and variance σ^2 is given by the function:

$$f(x) = \frac{e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}} \quad (5.16)$$

where $x \in [-\infty, \infty]$.

The area under the Normal curve in the interval $[x_1, x_2]$ corresponds to the *probability* that the Normal random variable with mean μ and variance σ^2 has a value in an interval $[x_1, x_2]$. It is given by:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \frac{e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}} dx. \quad (5.17)$$

Linear Statistical Inference and its Applications by C. R. Rao [35] is a foundational statistics textbook that provides a thorough grounding in the Normal distribution and its applications. Students with a mathematics and calculus background will find the book valuable. We have summarized important properties of the Normal distribution in [ReferenceBox 5.4.1](#).

ReferenceBox 5.4.1: Important properties of the Normal distribution

- The Normal distribution is symmetric about the mean μ .
- The mean median and mode of the Normal distribution are identical, equal to μ .
- When $\mu = 0$ and $\sigma = 1$, it is called the *Standard Normal* distribution given by:

$$f(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}$$

where $z \in [-\infty, \infty]$.

- Any Normal distribution can be converted to the Standard Normal distribution by subtracting the mean μ from the random variable X and dividing by the standard deviation σ . That is, if X is distributed according to a Normal distribution with mean μ and standard deviation σ , then the random variable:

$$Z = \frac{X - \mu}{\sigma},$$

is distributed according to the Standard Normal distribution with mean 0 and standard deviation 1.

- The area under the curve in the interval $[\mu - \sigma, \mu + \sigma]$ is 68% of all the area under the curve.

$$\int_{-\sigma}^{\sigma} \frac{e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}} dx = 0.68.$$

Using Equation 5.17:

$$P(-\sigma \leq X - \mu \leq \sigma) = P(-1 \leq Z \leq 1) = 0.68.$$

- Similarly, the area under the curve in the interval $[\mu - 2\sigma, \mu + 2\sigma]$ is 95% of all the area under the curve.

$$P(-2\sigma \leq X - \mu \leq 2\sigma) = P(-2 \leq Z \leq 2) = 0.95.$$

- The area under the curve in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ is 99.7% of all the area under the curve.

$$P(-3\sigma \leq X - \mu \leq 3\sigma) = P(-3 \leq Z \leq 3) = 0.997.$$

- The area under the curve sums to one, making it a probability distribution.

$$P(-\infty \leq X - \mu \leq \infty) = P(-\infty \leq Z \leq \infty) = 1.$$

- The distribution of the area under the Standard Normal distribution, and therefore probability, is shown in [Figure 5.10](#).

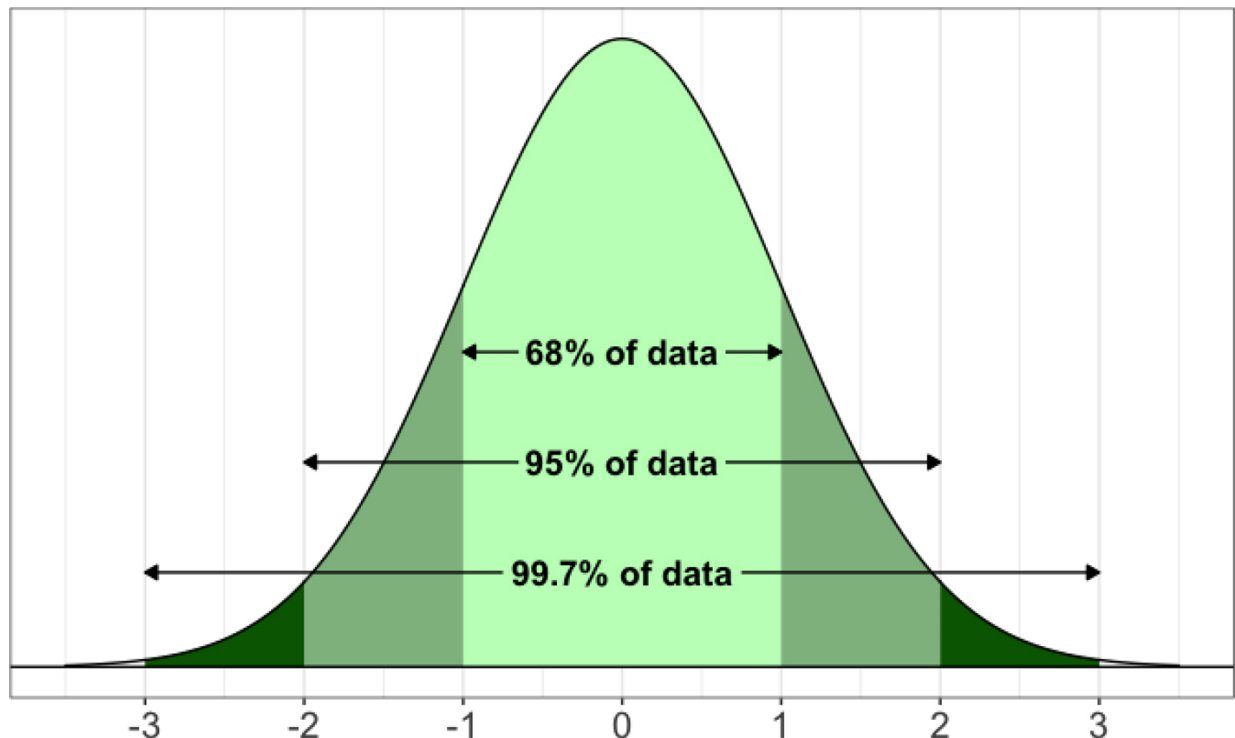


FIGURE 5.10: Consider the Standard Normal curve with mean 0 and standard deviation 1. The Normal curve has a special property: it is symmetric about the mean, 68% of the data lies within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations of the mean. Courtesy: ASPiegel, Wikimedia Commons. [↗](#)

Data Debate 5.4.1 Skewed distribution: *A Normal distribution is symmetric in shape and its mean, median and mode are identical. What can cause a distribution to become skewed, i.e., not symmetric?*

Other significant continuous distributions include:

- The χ^2 distribution (🔗), an asymmetric distribution defined over $[0, \infty)$, is used commonly for modeling squared error probability distributions. We will use it for testing independence of categorical attributes in [Chapter 6](#).
- The F -distribution (🔗), also an asymmetric distribution defined over $[0, \infty)$, is used for modeling the ratio of squared errors, used commonly for testing dependence between continuous and categorical variables in [Chapter 6](#).

Details of these distributions are provided in [36]. For the purposes of this book, it is enough to learn how to look up the values of these distributions using reference tables or software tools, as explained in [Chapter 6](#).

5.5 Margin of Error and Confidence Intervals

The Normal distribution is the rock star of probability distributions because it allows us to provide margins of error in a variety of applications from descriptive statistics to deep learning AI models. [Chapter 4](#) discussed statistical bias, the difference between a statistic (or models) estimated from the sample and the true population value, and the standard error in the estimation. In this section we will discuss *confidence intervals*, which are based on the probability distribution of the sample statistic.

Error bounds are used extensively in AI. For example, [FrontFinder AI](#), an AI tool, detects cold, warm, stationary and occluded weather fronts across the continental United States and provides error bounds in the form of confidence intervals. Other AI models like deep learning networks use confidence intervals for model selection, model accuracy, model performance and generalization. Such bounds are grounded in probability theory and Normal distributions.

The process of using samples to estimate the population mean and specifying the margin of error in the form of confidence intervals using the Normal distribution is outlined in [Figure 5.11](#).

- Step 1: A population consists of many instances; we cannot measure each one and compute the population mean, standard deviation or other descriptive statistics. We typically rely on samples from the population.
- Step 2: We learned in [Chapter 4](#) that each sample is different, and hence descriptive statistics vary from sample to sample. This is called *sampling variation*. [Figure 5.11](#) demonstrates sampling variation shown in Samples S1 through S5, and sample means T1 through T5.
- Step 3: If we choose many such samples, the sample means are distributed across a range of values. The distribution of the values of a sampling statistic is called its *sampling distribution*. Sample averages of any random variable, e.g., wind speed, humidity, weight, IQ, home prices, all follow the Normal probability distribution, no

matter what the underlying distribution of the random variable. This is important because that means we can always use the Normal distribution as the sampling distribution of the sample mean no matter what continuous random variable we are measuring! This is formalized by the *Central Limit Theorem*.

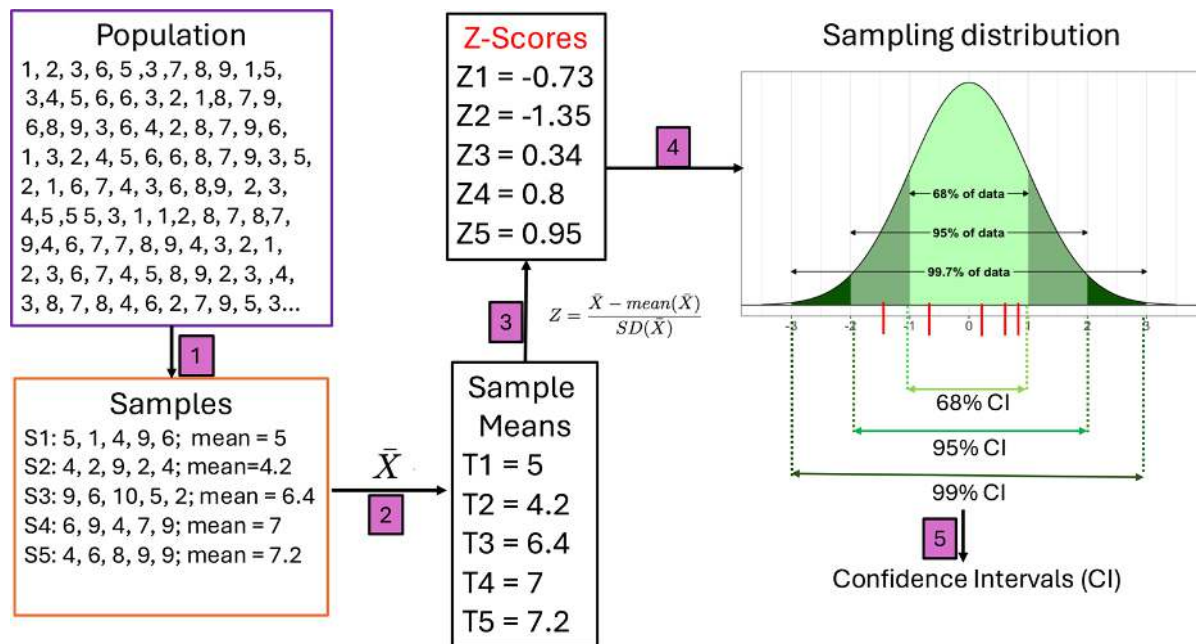


FIGURE 5.11: Normal distribution is used to estimate the error in using a sample to estimate population characteristics in the form of confidence intervals. Step 1: Sample from the population to estimate population characteristics without having to use all the populations data. Step 2: Each sample is different and results in different values of descriptive statistics, e.g., sample mean. Step 3: Convert the statistics to their standard values known as Z-scores by subtracting the mean and dividing by the standard deviation (SD). Step 4: The distribution of Z-scores (red lines) is the sampling distribution of the sample statistic. For the sample mean, because of the Central Limit Theorem, the sampling distribution is a Normal distribution. Step 5: Use the properties of the Normal distribution to find the *confidence intervals* that enclose a specified proportion of the Z-scores. The confidence intervals quantify the inaccuracy in approximating the population mean using the sample mean. [📄](#)

Definition 5.5.1 The *Central Limit Theorem (CLT)* states that the sample average $\bar{X}$ of n independent values, sampled randomly from the same population, has a Normal distribution for large n , with mean μ and variance σ^2/n where μ is the population mean,

and σ^2 is the population variance, **irrespective** of the distribution of the population values. $\sigma/\sqrt{n}$ is the standard error in approximating the population with the sample mean.

- Step 4: Since the values of $\bar{X}$, according to CLT are distributed according to a Normal distribution with mean μ and standard deviation σ , the quantity

$$Z = \frac{\bar{X} - \text{mean}(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

called a *Z-score*, is distributed as a Standard Normal distribution with mean 0 and standard deviation 1.

The sample means T1 through T5 in [Figure 5.11](#) are “standardized” to Z-scores Z1 through Z5. These Z-scores are shown as red ticks on the sampling distribution of the sample mean. Since the sampling distribution is the Normal distribution (bell-curve), 68% of Z-scores will lie in the interval $[-1, 1]$, 95% in the interval $[-2, 2]$ and 99% in the interval $[-3, 3]$.

- Step 5: The fact that the sample mean's Z-scores are distributed according to the Normal (bell) curve means that the probability that the interval

$$[\bar{X} - 3\sigma/\sqrt{n}, \bar{X} + 3\sigma/\sqrt{n}]$$

encloses μ is 99%. This interval is called a *99% confidence interval*. That is, if 100 intervals are computed from a 100 samples, then 99% of them will contain the true value of the population mean, μ . (Additional properties of the Normal distribution are described in [ReferenceBox 5.4.1](#).)

ReferenceBox 5.5.1: Sampling Variation and Margin of Error

([Figure 5.11](#))

- *Sampling variation* occurs because each sample is chosen to contain only a subset of the population values, and each subset is different. Therefore, descriptive statistics, e.g., the sample mean, computed from the samples will also vary from sample to sample.

- The *sampling distribution* of the sample mean captures the distribution of the sample mean values across different samples. It is a probability distribution that quantifies the uncertainty in the value of the sample mean caused due to sampling variation.
- The Central Limit Theorem (CLT) states that the sample mean $\bar{X}$ of n independent values, sampled randomly from the same population, is a random variable with a Normal distribution for large n , with mean μ and variance σ^2/n where μ is the population mean, and σ^2 is the population variance, *irrespective* of the distribution of the population values.
- A Z-score refers to a *standardized value* of a random variable by subtracting its mean and dividing by its standard deviation. Standardization creates a unit-free measurement that makes different variables comparable. For example, atmospheric pressure, temperature, wind speed when standardized are free of units of measurement like inches of mercury (Hg), Fahrenheit, meters/second, etc. and can be used in multivariate analytics like clustering. The Z-score of a random variable X is given by

$$Z = \frac{X - \mu}{\sigma},$$

where μ is the population mean of X and σ its standard deviation.

- Since, according to the CLT, the sampling distribution of the sample mean is a Normal distribution, then its Z-score,

$$Z = \frac{\bar{X} - \text{mean}(\bar{X})}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

is a Standard Normal distribution, and has a bell-shaped probability distribution like [Figure 5.10](#).

- The bell shape of the Standard Normal distribution implies that the probability that the sample mean falls in certain ranges called *confidence intervals* can be theoretically determined by the area under the Normal curve.
- The upper and lower bounds of the confidence are called *error bounds* and quantify the inaccuracy, or the *margin of error*, in estimating the population mean by the sample mean.

5.5.1 More on confidence intervals

The US Census Bureau has a succinct explanation of confidence intervals. Check *A Basic explanation of Confidence Intervals* in [Section 5.6](#). Here are some salient observations:

Note 1: The larger the sample size n , the narrower the interval $[\bar{X} - 3\sigma/\sqrt{n}, \bar{X} + 3\sigma/\sqrt{n}]$.

Note 2: The acceptable uncertainty or inaccuracy in the estimation of the population mean by the sample mean is quantified by the probability limit

$$\alpha = 1 - \text{confidence},$$

known as *level of significance*. These are often prespecified by the applications. In quality control in manufacturing, the error tolerance for estimating the average reliability of a part could be as low as 0.000001.

Note 3: The upper and lower limits of the confidence interval are called *error bounds* or *tolerance limits* since they represent the amount of error tolerated in the estimation of the population mean.

5.6 Read & Reflect

- *CDC's Vision for Using Artificial Intelligence in Public Health.* 
- *Planning a summer vacation? Let AI be your travel guide!* 
- *The Game Changer: How AI Is Transforming The World Of Sports Gambling.* 
- *DolphinGemma: How Google AI is helping decode dolphin communication.* 
- *Using artificial intelligence to find anomalies hiding in massive datasets: pinpointing potential power grid failure.* 
- *Robots in Everyday Life.* 
- *Determining uncertainty: a review of hurricane intensity predictability.* 
- *A Basic explanation of Confidence Intervals.* 

- *AI and machine learning are improving weather forecasts, but they won't replace human experts.* 

5.7 Exercises

1. **Probability of independent events.** You and some friends visit a fair and decide to play games to win prizes in the arcade. You decide to play three games, and you want to calculate the probability that you will win a prize. Here is a link to the three games: 
 - (a) Target Archery: In this game, you win a prize if the arrow that you shoot lands on the yellow, red or blue areas on the target board [Figure 5.12a](#). The diameter of the board is 100 cm, the width of each color is 10 cm.
 - i. Assuming that your arrow lands on the board, what is the probability that you will win a prize? (Hint: Calculate areas of the different colored circular strips).
 - ii. If the arrow lands on the yellow area of the target, you win the grand prize. What is the probability that you will win the grand prize, assuming that the arrow always hits the dart board?
 - iii. What is the probability that you will not win anything?
 - (b) Spinner Game: The second game stall you visit is the spinner. To win a prize, the pointer has to land on a double number (the outer number has to match the inner number) [Figure 5.12b](#).
 - i. What is the probability that you will win a prize on 1 spin?
 - ii. Each player is allowed 3 spins. What is the probability that you will win at least 1 spin?
 - (c) Lucky Dice Throw 8: Two Red Dice. The next stall is our favorite dice stall, where we win if we get a sum of 8 in one throw of the two dice. Let's see if we have more chances to win here. What is the probability that you will a prize? What is the probability of getting a sum of 8 twice in exactly two of three throws? [Figure 5.12](#)



FIGURE 5.12: Left: Archery target. Image from Wikimedia Commons. [🔗](#). Center: Spinner. Wikimedia Commons. Image Attribution: BornintheGuz at the English-language Wikipedia [🔗](#). Right: Two red dice. Image from Wikimedia Commons. [🔗](#). [📄](#)

2. **AI in games.** AI is used in many computer games, and probability theory plays a role in determining moves, and predicting outcomes to make the game competitive. AI has mastered games such as Chess, Go and the card game Bridge. Many games require human-like communication, understanding rules and predicting the moves of the human players. Read this article from *The Guardian* (March 2022) about the AI called “NooK” that was able to beat eight professional master players of the card game Bridge. [🔗](#)

Choose a popular computer game and find out how AI is applied and what role probability theory plays in developing and creating this game.

3. **Probability.** The Flying Club has 120 members, 72 are men and 48 are women. The club wants to form a committee of 5 members for a special air show event, and they decide to randomly choose 5 members (pulling names from a hat), one at a time.
 - (a) What is the probability that all 5 committee members are men?
 - (b) What is the probability that 4 are men and 1 is a woman?
 - (c) Given that the first two draws are men, what is the probability that the next draw will be a woman?
 - (d) The club members want a fair representation from both men and women and decide that they will select 3 men and 2 women. How should they go about doing this?

4. **Conditional probability.** Mrs. Murthy conducted a survey of 7th grade students who played school individual sports and team sports. There were 80 boys and 59 girls who participated in the survey. The results of the survey are listed in [Table 5.1](#). Calculate the probabilities for the following:

TABLE 5.1: Survey of 7th grade students [↗](#)

	Team	Individual	Both	No Sports
Boys	80	52	35	25
Girls	59	34	18	36
Total	139	86	53	61

- (a) Random student chosen is a girl and plays only team sports.
 - (b) Student selected at random is a boy who plays both team and individual sports.
 - (c) Student selected at random plays no sports.
 - (d) Given that a student selected is a girl, what is the probability that she plays team sports?
5. **Application of Binomial distribution.** A pharmaceutical company has developed a certain medication, and it was found that 8% of adults who take this medication experience negative side effects. If we select a random sample of 100 patients who take this medication, use the Binomial probability distribution to find the following probabilities:
- (a) Exactly 10 people will suffer negative side effects.
 - (b) More than 6 people will suffer negative side effects.
 - (c) More than 20 people will suffer negative side effects.
 - (d) No one has any side effects.
6. **Binomial Distribution.** Consider the following scenario in basketball: A basketball player takes four independent free throws with a probability of 0.7 of getting a basket on each shot. The number of baskets made is recorded. Here each free throw is a trial, and trials are assumed to be independent. Each trial has two outcomes: basket (success) or no basket (failure). The probability of success is $p = 0.7$, and the probability of failure is

$q = 0.3$. This is an example of a Bernoulli experiment with 4 trials. If the variable X represents the number of successes, find the Binomial distribution of X .

7. **Confidence intervals and error bounds.** Confidence intervals are frequently used in many aspects of real life when the outcomes are uncertain, and we need to quantify the uncertainty. This is a link to a problem set for practice. Solve the first 5 problems from this set. Check your solutions with the answers provided for selected problems! 

8. **Create a histogram using NBA data about team average points per game.** Here is a data table with team points per game for the 2024-2025 NBA season: 

Create a histogram for the average points per game for 2024 using a bin size of 2 and using the insert chart function in Google Sheets. Find the mean and standard deviation for the data. Is this distribution a Normal distribution? Why? Examine the histogram and state three things you notice about the distribution of average points scored.

9. **Mean and Standard deviation.** In this exercise, you will examine the data for the fastest woman sprinter Gabby Thomas. Gabby has won many championships, including the Gold medal for the 200 m race in Paris Olympics in 2024. Here is the link to the document with details about her performance: 

- (a) What are some challenges when creating a Normal distribution with this data set.
- (b) Calculate Mean and Standard deviation.
- (c) Visualize (Histogram)-create a histogram to see how the data is distributed and compare it to a theoretical Normal curve.

10. **Applying Bayes' Theorem.** Note: The following problems are selected from the problem set on the website . A person uses his car 30% of the time, walks 30% of the time and rides the bus 40% of the time as he goes to work. He is late 10% of the time when he walks; he is late 3% of the time when he drives; and he is late 7% of the time he takes the bus.

- (a) What is the probability he took the bus if he was late?
- (b) What is the probability he walked if he is on time?

5.8 Summative Project:

How the sports industry uses AI Artificial intelligence is widely used in the sports industry for various purposes. Teams collect and analyze large amounts of data, and AI is used to analyze and predict performance about individual athletes and teams. Data from tracking devices are used to train athletes. All kinds of data are used to predict performance and outcomes for future matches and competitions. Fan engagement is tracked through smart devices and social media. Streaming channels and TV bring in revenue for networks and advertisers, and AI plays an important role in targeting the audiences for different sports. The marketing of sports goods and fantasy sports have proliferated since smartphones became widely used. Probabilistic AI models are being developed to make sports more exciting and interesting.



FIGURE 5.13: All kinds of sports are using AI applications for analyzing data

1. Use Gemini to find data for a sports team or player of your choice. What kind of statistics are used to rank players or teams in your chosen sport? Explain. How is probability applied to make predictions about the performance of the team or player? Examine the data and check the accuracy of the Gemini output!

2. In this example, a Gemini prompt for the New York Mets baseball team data for last 25 years produced the following output: . How does this compare with the actual results for the 2025 season?
3. This is the output for a prompt to find the probability of New York Mets winning the 2025 season: . Ask Gemini for the probability of the Mets winning the 2026 season championship. What data is used to compute this output?
4. Fantasy Sports has become increasingly popular with sports fans, and now AI algorithms are making this a multi-billion dollar industry. Choose a fantasy sport that interests you and find the various factors that you have to consider while creating your fantasy sports team. Explain the rankings for the teams and how probability is used in creating these rankings. Here is the link to NBA rankings for Fantasypros for basketball. What information has been used to create this ranking chart? 

5.9 Appendix

This appendix contains detailed explanations of a select set of topics on probability.

5.9.1 Sets of outcomes

Simple concepts from set theory provide a useful way of thinking about outcomes and probabilities. [ReferenceBox 5.9.1](#) provides definitions of commonly used terms in *Set theory*. Herbert B. Enderton's *Elements of Set Theory* [16] is a great introduction. The definitions presented here have many variations, e.g., when applicable to more than two sets, but the discussion here is limited to concepts that are useful for explaining probability theory.

ReferenceBox 5.9.1: Set Theory Terminology

- *A set A* is a collection of distinct objects, called its *elements*.
- *The cardinality $|A|$* of a set A is the number of elements in the set. It is 0 for an empty set with no elements.

- The complement $\overline{A}$ of a set A consists of all the elements in the sample space S that are NOT in the set A .
- The intersection of two sets A and B , denoted by $A \cap B$, consists of all the distinct elements that are present in both A AND B .
- The union of two sets A, B denoted by $A \cup B$, consists of all the distinct elements that are in A OR in B , i.e., present in at least one of the sets.
- A set A is a (proper) subset of another set B , denoted by $A \subset B$, if all its elements are present in B .
- Mutually exclusive sets have no element in common. That is, the intersection of two mutually exclusive sets is the empty set ϕ .

As a quick example to illustrate set theory terms, consider the following set of ice cream flavors $S = \{\text{Chocolate, FroYo, Honey, Mango, Matcha, Mint, Rocky Road, Strawberry, Vanilla, Vegan}\}$, and its subsets:

$A = \{\text{Chocolate, Mango, Matcha, Mint, Strawberry, Vanilla}\}$

$B = \{\text{Chocolate, Honey, Rocky road, Strawberry, Vanilla}\}$

$C = \{\text{FroYo, Honey}\}$

$D = \{\text{Rocky road}\}$

$E = \{\text{Vegan}\}$.

[Figure 5.3](#) shows the sets in relation to each other. Using the definitions in [ReferenceBox 5.9.1](#):

The following pairs of sets are *overlapping* and have elements in common, (A,B) , (B,C) , and (B,D) . For example, the flavors that are “in set A AND in set B ”, denoted by $A \cap B = \{\text{Chocolate, Strawberry, Vanilla}\}$.

The following pairs of sets are non-overlapping or mutually exclusive: (A,C) , (A,D) , (A,E) , (B,E) . For example, the intersection of the two sets has no elements and is empty, $A \cap D = \phi = \{\}$, where ϕ is the Greek symbol traditionally used to represent an empty set.

The set D is a subset of the set B , denoted by $D \subset B$.

The complement of the set A is given by all the flavors in the sample space that are “NOT IN” A, $\overline{A} = S - A = \{FroYo, Honey, Rockyroad, Vegan\}$.

The union of flavors in sets “B OR C” is: $B \cup C = \{Chocolate, FroYo, Honey, Rocky road, Strawberry, Vanilla\}$. Note that in order to avoid double-counting, the flavor Honey is represented only once in the union of the sets B and C, even though it is present in both. So, $B \cup C = B + C - B \cap C$.

The cardinality of the set of A is the number of elements of in set A, given by $|A| = 6$.

5.9.2 Conditional probability distribution of $Y|X = CC$

- **Computing marginal probabilities from joint point probabilities:** The individual probabilities of two random variables X and Y can be computed from the joint probabilities of X and Y . For example, the individual probabilities of flavor choices for your friend Y can be computed from [Section 5.3.3](#) in the following manner:

$$\begin{aligned}
 P(Y = OF) &= P(X = CC \ \& \ Y = OF) + P(X = S \ \& \ Y = OF) \\
 &= 0.5 + 0.05 \\
 &= 0.55 \\
 P(Y = S) &= P(X = S \ \& \ Y = S) \\
 &= 0.1 \\
 P(Y = CC) &= P(X = CC \ \& \ Y = CC) \\
 &= 0.1 \\
 P(Y = Mint) &= P(X = CC \ \& \ Y = Mint) + P(X = Mocha \ \& \ Y = Mint) \\
 &= 0.2 + 0.05 \\
 &= 0.25
 \end{aligned}$$

Sanity check: All probabilities lie in the range $[0,1]$ and sum to 1.

- **Computing marginal probabilities from conditional probabilities:** Suppose we want to know how often your friend chooses Mint. We simply sum his choice of Mint given all your possible flavor choices. From [Figure 5.8](#), we know that he chooses Mint 0.2 proportion of the times you choose CC, which is 0.8 times in total; therefore, he chose Mint $0.2/0.8=0.25$ of the times you chose CC. Similarly, he chooses Mint 0.05 proportion of the times you choose Mocha, which is 0.05 of the times in total; therefore

0.05/0.05=1, i.e., he chose Mint every time you chose Mocha. These are the only two occasions when he chooses Mint. Therefore,

$$\begin{aligned}
 P(Y = \text{Mint}) &= P(Y = \text{Mint}|X = \text{CC})P(X = \text{CC}) \\
 &\quad + P(Y = \text{Mint}|X = \text{Mocha})P(X = \text{Mocha}) \\
 &= 0.25 * 0.8 + 1 * 0.05 \\
 &= 0.25
 \end{aligned}$$

given your flavor choices. This logic can be stated formally for any two random variables X and Y as:

$$P(Y) = P(Y|X)P(X) \tag{5.19}$$

- **Computing conditional probabilities and deriving Bayes' Rule:** How often does your friend choose CC when you choose CC? From [Section 5.3.3](#), you two choose it simultaneously only 10% of the time. So of the 80% of the time you choose CC, your friend also chooses it 10% of the time. Similarly, you both choose Strawberry simultaneously 10% of the time out of the 15% of the time you choose Strawberry. This can be stated as:

$$\begin{aligned}
 P(Y = \text{CC}|X = \text{CC}) &= 0.1/0.8 = 0.125 \\
 &= \frac{P(Y = \text{CC} \ \& \ X = \text{CC})}{P(X = \text{CC})} \\
 P(Y = \text{Mint}|X = \text{CC}) &= 0.2/0.8 = 0.25 \\
 &= \frac{P(Y = \text{Mint} \ \& \ X = \text{CC})}{P(X = \text{CC})} \\
 P(Y = \text{OF}'|X = \text{CC}) &= 0.5/0.8 = 0.625 \\
 &= \frac{P(Y = \text{C} \ \& \ X = \text{CC})}{P(X = \text{CC})}
 \end{aligned}$$

Sanity check: $P(Y|X)$ is a probability distribution since all the probabilities lie in the interval $[0,1]$ and sum to 1.

The above reasoning can be stated in general for any two random variables X and Y as:

$$P(Y|X) = \frac{P(Y, X)}{P(X)},$$

(5.20)

where $P(Y, X)$ denotes the joint distribution of Y, X and can also be written as $P(Y \cap X)$. Similarly,

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}, \quad (5.21)$$

Rearranging Equations 5.20 and 5.21, and noting $P(X, Y)$ is the same as $P(Y, X)$, we get:

$$P(Y|X)P(X) = P(Y, X) = P(X, Y) = P(X|Y)P(Y).$$

Finally, with one more algebraic step, we get the famous Bayes' Rule:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \quad (5.22)$$

The full conditional distribution of your friend's choice Y given you chose CC is given by:

$$\begin{aligned} P(Y = CC|X = CC) &= 0.1/0.8 = 0.125, \\ P(Y = S|X = CC) &= 0, \\ P(Y = Mint|X = CC) &= 0.2/0.8 = 0.25, \\ P(Y = Mocha|X = CC) &= 0, \\ P(Y = OF|X = CC) &= 0.5/0.8 = 0.625 \end{aligned}$$

Other Flavors (OF) represents flavors other than Strawberry (S), Chocolate (CC), Mocha and Mint.

- **Testing independence of two random variables:**

From [Section 5.3.5](#), we can make the following computations for your flavor choices X , and your sister's choices Z .

The joint probability distribution of X and Y is given by:

$P(X, Y) = (CC \& CC=0.1, CC \& Mint=0.2, S \& S=0.1, CC \& OF = 0.5, S \& OF = 0.05, Mocha \& Mint= 0.05, All else = 0)$

The marginal probability distribution of X is given by:

$P(X) = (CC=0.8, S=0.15, Mint=0, Mocha=0.05, OF=0)$ The marginal probability distribution of Y is given by:

$P(Y) = (CC=0.1, S=0.1, Mint=0.25, Mocha=0, OF=0.55)$

The product of probability distributions $P(X)P(Y)$ is given by:

$P(X)P(Y) = (CC \ \& \ CC=0.1*0.8=0.08, S \ \& \ S=0.1*0.15=0.015, CC \ \& \ S = 0.8*0.1=0.08, All \ else = 0)$

Therefore, since $P(X, Y) > P(X)P(Y)$, it contradicts Equation 5.11, and we conclude that X and Y are not independent.

Data Debate 5.9.1 *Is $P(Z)P(X)$ a probability distribution? Why or why not?*

Note that $P(Z)P(X)$ is the product of two probability distributions and is NOT a probability distribution, hence it does not have to follow the rules of probability distributions. It does not sum to 1.

Data-Driven Hypotheses Testing

Is it better to exercise in the morning or at night? Which is more popular, coffee or tea? Does an AI algorithm favor one group over another? In order to provide objective and data-driven answers to such questions, we collect samples for gathering data, formulate and test hypotheses to draw conclusions within an acceptable probability of error. This process is called *statistical hypothesis testing*.

Social media companies extensively use hypothesis testing in the form of *A/B testing* to increase user engagement by comparing ease-of-use and attractiveness of icon designs and placement. NASA relies on hypothesis testing to distinguish between flight trajectories of suspicious drones and those of birds. Self-driving cars recognize moving objects like bicycles, humans and cars by testing them against statistical profiles based on known features. AI models and algorithms are audited for fairness and bias using hypothesis testing. In this chapter, we will learn about statistical hypothesis testing and its applications in AI.

6.1 Automation and Hypothesis Testing

Companies increasingly rely on AI to screen job applicants, determine parole eligibility for the incarcerated, and to establish compatibility on

dating apps. They use statistical hypothesis testing to perform sanity checks called *AI system audits*, e.g., to ensure that female applicants are not excluded while AI algorithms screen resumès for hiring, or compliance with regulatory requirements. Social media companies use hypothesis testing to compare different designs and products in order to find the best way to keep users engaged on their platforms. *Statistical hypothesis testing* provides an objective way to answer questions, compare groups, and derive conclusions and make inferences based on sampled data. The framework provides guarantees on the validity of results through margins of error and probability that the conclusions are correct. AI applications use testing-based reasoning extensively, e.g., to decide whether feature A or feature B has better user engagement. In this chapter, we will provide an overview of statistical hypothesis testing and its applications, and highlight its relevance to AI. Our objective is to cover the necessary concepts and provide Gemini prompts, additional material in the Appendix, and further references to help those interested in deeper study. A comprehensive and rigorous discussion of this topic can be found in *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II* by Bickel and Doksum [2], and *Statistical Inference* by Casella and Berger [7].



FIGURE 6.1: Left: AI is often used to screen resumès during the hiring process. Does it favor one group over another? Right: Are puppies adopted at higher rates than other animals at rescue centers? Statistical hypothesis testing provides data-driven answers to such questions.

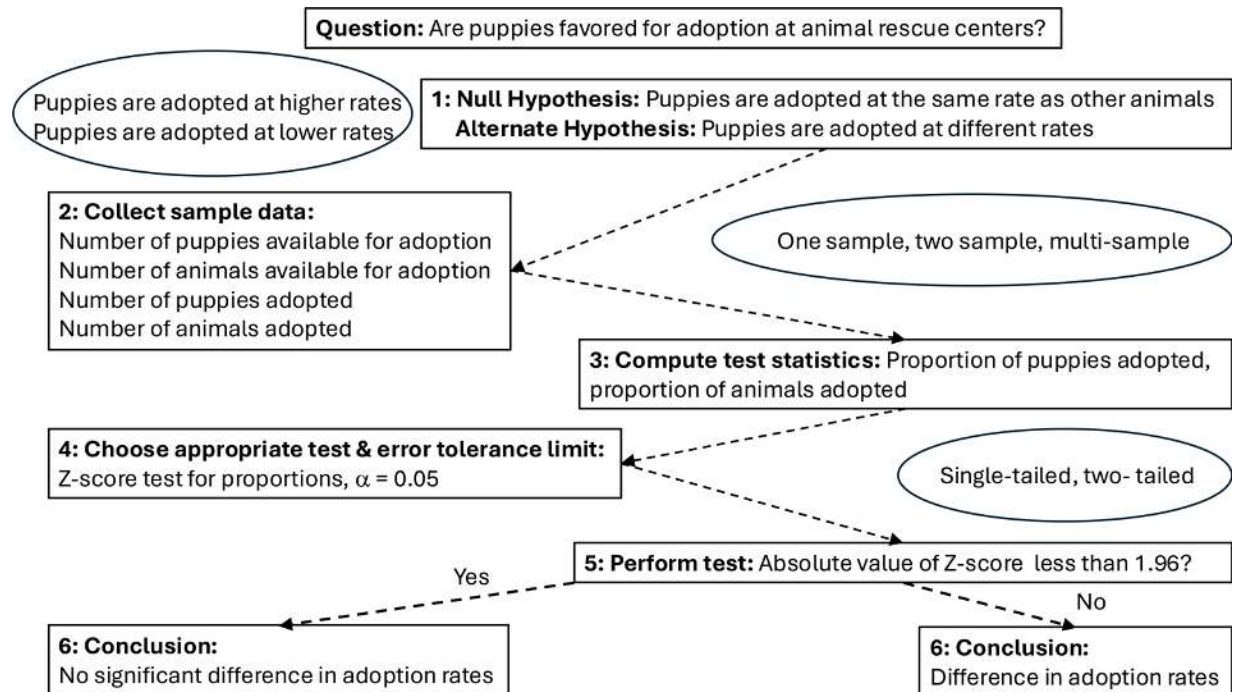
6.1.1 Process of hypothesis testing: Key steps

Suppose an animal rescue organization wants to check if there is a preference towards adopting puppies over other animals at their rescue centers across the country. Since it is expensive and time-consuming to collect data from every rescue center, they would choose a few representative centers across the country and gather data over a period of time regarding animals available for adoption, and check if the proportion of puppies that were adopted during that period was “similar” to the proportion of all animals adopted.

However, what constitutes “similar” or “different”? Are adoption rates of 0.4 (puppies) and 0.39999 (all animals) different? How about 0.4 and 0.389? Where would one draw the line? The two proportions will never be identical due to sampling variation and other chance happenings. If an influencer announces “Adopt a Peacock” day at the time the samples are collected, there could potentially be a one-time spike in peacock adoptions that could produce misleading results. The process of *statistical hypothesis testing* is an objective, data-driven way to determine whether the two proportions are significantly different (or not) after accounting for the noise of such chance variability.

[Figure 6.2](#) shows a simple process diagram of hypothesis testing in the context of the puppy adoption question. [ReferenceBox 6.1.1](#) lists the key steps, and [ReferenceBox 6.1.2](#) provides definitions of key terms and concepts. [Section 6.2.3](#) shows a step-by-step worked-out illustrative

example of hypothesis testing. Please refer to these as you read through the chapter.



Extended Description

FIGURE 6.2: Are puppies favored for adoption over other animals at animal rescue centers? The process of statistical hypothesis testing is designed to answer such questions in an objective, data-driven manner. [↗](#)

ReferenceBox 6.1.1: Key Steps in Hypothesis Testing [↗](#)

1. Formulate null hypothesis based on question to be answered. State the alternative hypothesis.
2. Collect samples and gather relevant data.
3. Compute a suitable descriptive statistic from the sample.
4. Choose an appropriate sampling distribution of the descriptive statistic and specify level of significance.

5. Test if the sample statistic is consistent with the sampling distribution dictated by the null hypothesis, i.e., is it a typical value or an unlikely, low-probability value?
6. State conclusion and answer question.

We describe below the key steps mentioned in [ReferenceBox 6.1.1](#) and illustrated in [Figure 6.2](#). The terminology and definitions are explained in [ReferenceBox 6.1.2](#) and should be thoroughly studied. The key steps in hypothesis testing are:

1. **Formulate null and alternative hypotheses:** Based on the question that needs to be answered, compose a hypothesis that you wish to confirm or contradict, e.g., “Puppies are adopted at the same rate as other animals at the animal rescue center.” By default, in the absence of any prior evidence, we assume the status quo, or *null hypothesis*, that there is no difference, and we use the sampled data to either confirm or contradict this hypothesis.

An *alternative hypothesis* is a specific statement that holds if the null hypothesis is not true. “Puppies are adopted at a different rate than other animals at rescue centers” is an example of an alternative hypothesis. The bubble next to Step 1 in [Figure 6.2](#) shows two other examples of alternate hypotheses, of higher and lower rates of puppy adoptions. It is important to state the alternative hypothesis before collecting data so that it is not influenced by what we observe in the data.

2. **Select a sample and gather data:** Collect data relevant to the problem. For example, sample animal rescue centers across the country, and at each center count: the number of animals available for adoption during the past month, number of puppies available for adoption during the past

month, number of animals adopted in the past month, and number of puppies adopted in the past month. It can be any duration, but applied consistently across all data that is collected. Choose the most suitable sampling technique to get reliable results.

- Make sure that the data collected is *relevant* to the problem. There is no point in collecting data about tigers in jungles or migrating wildebeest in the Serengeti, [Figure 6.3](#), however tempting.
- In addition, ensure that the sampling process provides representative data of animal rescue centers across the country, and on different days to mitigate the effect of abnormal data such as “Adopt a Peacock” day sponsored by an influencer. What would happen if you were to collect data only on “Adopt a Peacock” day?



FIGURE 6.3: Data relevance: To answer questions about adoption rates of puppies at rescue centers, would it make sense to collect data about: Tigers in jungles? Wildebeest in the Serengeti? What would happen if you were to gather samples at rescue centers only on “Adopt a Peacock”

day? Photo credits: [Tiger](#), [Wildebeest](#) and [Peacock](#), courtesy Unsplash.



Data Debate 6.1.1 *What type of sampling method should we use to collect samples to test if puppy adoption rates are the same as those of all animals at rescue centers?*

3. **Compute a descriptive statistic from the sample:** Determine the appropriate descriptive statistic $T(X)$ to be computed from the sample, e.g., sample mean, standard deviation, or proportion. In the instance of pet adoptions, we would use the proportion of all animals that are adopted, and the proportion of puppies that are adopted. $T(X)$ is a random variable since it varies from sample to sample. $T(X)$ is called a *test statistic* because it is used to test the validity of the null hypothesis.
4. **Choose an appropriate test and level of significance α :** Use the sampling distribution of the test statistic $T(X)$ based on the null hypothesis to identify regions of high likelihood for typical values. For simplicity, we will drop $T(X)$ and use the notation $P_{\mathcal{N}}(X)$ to represent the probabilities of the sampling distribution of the test statistic under the null hypothesis, and $P_{\mathcal{A}}(X)$ under the alternate hypothesis.

In the pet adoption example, the sampling distribution of the proportion of adopted animals is a Normal distribution since a sample proportion is a special kind of an average. Explore [GeminiBox 6.1.1](#) to find out how a sample proportion is an average.

GeminiBox 6.1.1: Sample proportion is an average

“Is a sample proportion a kind of an average?”

We will use the sample mean as the test statistic going forward, but the hypothesis testing process is similar for any test statistic such as the median.

Since the probability tables for the Standard Normal distribution are widely available, it is customary to convert the test statistic to a *Z-score*, by subtracting the average of the sampling distribution of the sample mean and dividing by the standard deviation of the sampling distribution of the sample mean. From [Chapter 5](#) we know that the average of the sampling distribution of the sample mean is the population mean μ , and the standard deviation is the *standard error*, given by $\sigma/\sqrt{n}$, where μ and σ are the population mean and standard deviation of the population. Then, the Z-score:

$$Z = \frac{T(X) - \mu}{\sigma/\sqrt{n}},$$

has a Standard Normal distribution.

Using the special properties of the Standard Normal distribution discussed in [Chapter 5](#), and the probability tables of the Standard Normal distribution for an accurate probability interval, we can state that

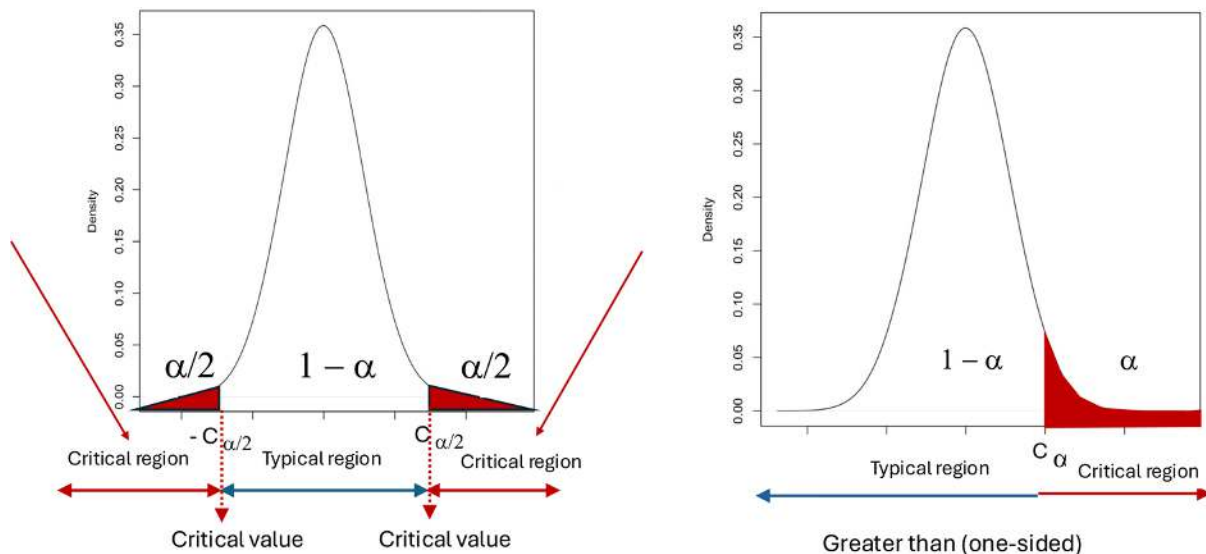
$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

Therefore we expect that 95% of samples will have a Z-score that falls in this region, as shown in [Figure 6.4](#).

The bounds -1.96 and 1.96 are called *critical values* that correspond to a *confidence level* of 95%, the probability that a Z-score will fall in that interval marked typical region in [Figure 6.4](#).

The probability $\alpha = 1 - 0.95$ is called the *level of significance*, the low probability region of the distribution where there is only a 5% chance of finding a sample from the null distribution through sheer chance. They are shaded red in [Figure 6.4](#). The region outside the typical region, beyond the critical values, is called the *critical region*.

5. **Check the test condition:** Is the Z-score computed from the sample in the typical region of the sampling distribution under the null hypothesis, where the probability of finding such a value is high?
6. **Draw a conclusion:** If the answer is yes, then the test statistic $T(X)$ computed from the sample is typical and the null hypothesis is valid. If the answer is no, it is quite unlikely (less than α probability) that the null hypothesis is valid based on the sample. For example, if the test statistic, the proportion of puppies, is within the typical region, then the adoption rates for puppies is the same as other animals. If the answer is no, then there is a significant difference in the adoption rate of puppies compared to other animals at the rescue center.

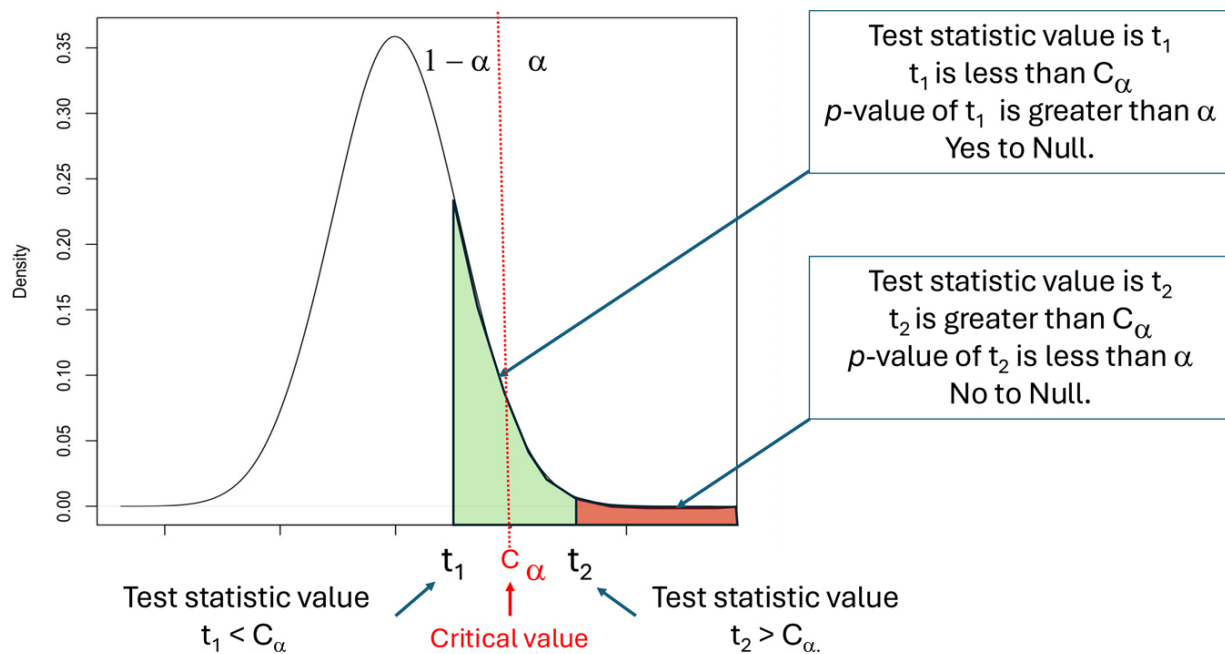


[Extended Description](#)

FIGURE 6.4: Left: A two-tailed test, showing critical values, critical and typical regions where the level of significance α is split between the two extremes of the distribution. Right: Single-tailed test. [↗](#)

6.1.2 What is a p -value?

A p -value is a probability related to the test statistic computed from the sampling distribution and provides an alternate way of testing the null hypothesis based on probabilities instead of critical values. We will discuss using a one-tailed test for simplicity, but the explanation applies in general. The area to the right of the value of a statistic under its sampling distribution is called its p -value, as shown in [Figure 6.5](#). The larger the p -value, the more typical the sample.



[Extended Description](#)

FIGURE 6.5: Making decisions regarding validity of the null hypothesis using probabilities rather than the equivalent critical region. Values of sample statistics t_1 and t_2 , corresponding p -values and critical value C_α . [↗](#)

Consider [Figure 6.5](#), where the alternative hypothesis is “puppies are more popular for adoption”, and the critical region consists of the interval $[C_\alpha, \infty]$. Recall that the probability that the value of any sample mean falls in the typical region (i.e., outside the critical region) is 0.95. Suppose we choose a sample, say sample no. 1, and compute the sample statistic t_1 . Let

$$p_1 = P_{\mathcal{N}}(X > t_1),$$

be the probability to the right of t_1 under the sampling distribution of the test statistic under the null distribution $P_{\mathcal{N}}(X)$, where the subscript $\mathcal{N}$ indicates the null distribution. The area corresponding to p_1 is shown in green in [Figure 6.5](#). It is quite large, and $p_1 > \alpha$, which means that t_1 is a fairly typical value deep inside the confidence interval. The sample statistic value t_1 is smaller than the critical value C_α , therefore t_1 does not lie in the critical region, leading to the conclusion that the null distribution is valid. Puppies are as popular as other animals for adoption.

Suppose we choose another sample, sample no. 2, and compute the test statistic based on that sample. Its p -value, the probability to the right of it under the sampling distribution corresponding to the null hypothesis, is given by:

$$p_2 = P_{\mathcal{N}}(X > t_2).$$

Note that p_2 is quite small, $p_2 < \alpha$, as shown by the red shaded region in [Figure 6.5](#). This means that t_2 , the value of the sample statistic, is way outside the confidence interval and in the low-probability region of the distribution $P_{\mathcal{N}}(X)$, and much greater than the critical value C_α , indicating that the sample is unlikely to be from the null distribution. Puppies are preferred and adopted at greater rates than other animals.

Data Debate 6.1.2 *If the value of a statistic computed from a sample is equal to C_α , what is its p -value?*

In general, if the p -value of a test statistic whose value is t where $p = P_{\mathcal{N}}(X > t)$, is greater than α , the level of significance, then the null hypothesis is valid; otherwise, there is no evidence to support it.

Key concepts in hypothesis testing are summarized in [ReferenceBox 6.1.2](#), illustrated in [Figure 6.2](#), and worked step-by-step through a hands-on example in [Section 6.2.3](#).

ReferenceBox 6.1.2: Hypothesis Testing Terminology

- *Hypothesis* refers to a statement formulated to answer a particular question. The statement will be tested, i.e., confirmed or contradicted, using data sampled from the population.
- *Null or baseline hypothesis* refers to the default state, which is a hypothesis of no difference, until we can prove otherwise. An example null hypothesis is: “there is no difference in the adoption rate of puppies at rescue centers compared to other animals.”
- *Alternate hypothesis* is a specific alternative to the null hypothesis. For example, “puppies are adopted at a different rate”, or “puppies are adopted at a higher rate”, or “puppies are adopted at a lower rate.” The first statement includes both the second and third statements.
- *Test statistic*, $T(X)$, is a descriptive statistic computed from the sampled data. It is a random variable since it varies from sample to sample. The value of $T(X) = t$ based on the particular sample X is used to make a determination regarding the hypothesis. Sample

mean, sample proportion and sample median are commonly used test statistics. For simplicity, we will use the Normal distribution as an example of the sampling distribution of $T(X)$ under the null hypothesis, but the discussion holds for other types of sampling distributions as well.

- The test statistic is converted to a *Z-score* by subtracting the mean μ of the sampling distribution of the sample mean and dividing by the standard deviation of the sampling distribution of the sample mean, i.e., its standard error $\sigma/\sqrt{n}$. The *Z-score* follows the Standard Normal distribution whose probability tables are available.

$$Z = \frac{T(X) - \mu}{\sigma/\sqrt{n}},$$

where μ is the population mean and σ is the population standard deviation. The distribution of $T(X)$ is represented $P_{\mathcal{N}}(X)$ if the null hypothesis is true, and $P_{\mathcal{A}}(X)$ if the alternate hypothesis is true. Technically, we should use $P(T(X))$, but we use $P(X)$ going forward to simplify the notation.

- *Confidence level* is the probability associated with the typical region of the Standard Normal distribution. If the *Z-score* computed from the sample falls in the typical region, then it is likely that the null hypothesis is true based on the sample. The typical region is marked and shown with a blue arrow in [Figure 6.4](#) and corresponds to the interval where a majority of values of $T(X)$ are expected to fall if the null hypothesis is true.
- *Critical region* is a low probability region of the *Z-score* distribution. If the *Z-score* falls in the critical region then it is unlikely that the

null hypothesis is true based on the sample. The critical region is shaded in red and shown with red arrows in [Figure 6.4](#).

- *Critical values* are the cut-off values that mark the boundary of the critical region. They are the thresholds used for determining whether a test statistic is consistent with the null hypothesis. Examples of critical values are $-C_{\alpha/2}, C_{\alpha/2}$ when the alternate hypothesis consists of two regions, and C_{α} when the alternate hypothesis corresponds to a single region as shown in [Figure 6.4](#). These correspond to two-tailed and one-tailed tests discussed in [Section 6.2](#).
- *Level of significance, α* , is the probability corresponding to the critical region. It is the probability of finding an unusual sample and incorrectly rejecting the null hypothesis. In [Figure 6.4](#),

$$\alpha = P_{\mathcal{N}}(X < -C_{\alpha/2}) + P_{\mathcal{N}}(X > C_{\alpha/2}),$$

for the two-part critical region in the figure on the left, and

$$\alpha = P_{\mathcal{N}}(X > C_{\alpha}),$$

for the one-part critical region in the figure on the right.

- *Test condition* involves checking whether the sample's Z -score falls in the typical or critical region.
- *The p -value, p* , corresponding to sample statistic $T(X)$ is the probability under the null hypothesis of values exceeding the Z -score $= t$ computed from the test statistic.

$$p = P_{\mathcal{N}}(X > t).$$

The larger the p -value, the more typical the sample. The smaller the p -value, the more likely that the Z -score is in the critical region. [Section 6.1.2](#) has a detailed explanation.

Statistical hypothesis testing involves variations of this fundamental process depending on the question, sample sizes, number of samples, and the nature of the sampling distributions. We will discuss each scenario in detail. See [Section 6.2.3](#) for a step-by-step example of how to use sample data to test a hypothesis.

6.2 Tests Based on Type of Hypothesis

In the discussion above, we discussed the null or baseline hypothesis of no difference. For example, in the animal adoption example, we assumed that the puppy adoption rates are the same as all animals. We did not specify whether the adoption rate is higher or lower, just different. This is an important point that results in two different types of tests and the way the level of significance is interpreted.

6.2.1 Two-tailed test

For the null hypothesis “there is no difference in adoption rates for puppies compared to all rescue animals”, there are two alternate possibilities, i.e., higher or lower adoption rates. Therefore, an alternate hypothesis that has no prior knowledge or belief includes both unusually high rates of puppy adoptions as well unusually low rates. The critical region will be split in two, as shown in [Figure 6.4](#). The red shaded areas indicate whether the adoption rate for puppies is more (right side) or less (left side) compared to

the overall adoption rates for all animals. This is called a *two-tailed* test since there are two possible alternatives to the null hypothesis of “no difference in proportion of adoptions.” Such a test would flag puppy adoption rates that are too low or too high.

6.2.2 One-tailed test

Suppose that the rescue center manager tells us that puppies are adopted at a higher rate than the average adoption rate for rescue animals. This implies that when we look for unusual values, we should focus only on higher values since we have prior information. Such a test with only one possible alternative is called a *single-tailed* or one-tailed test, as shown in [Figure 6.4](#). (A similar argument applies for “lower adoption rate” had the manager told us that puppies are adopted at a lower rate than the rate for all animals. We would use a test where the critical region lies entirely on the lower valued side of the distribution.)

Many common test statistics like sample mean and sample proportion follow a Normal distribution. The critical values for a given α can be determined using Standard Normal tables or using software tools. Work through [Example 6.2.3](#) to fully understand the process of hypothesis testing.

6.2.3 Illustrative Example: Testing adoption rates of puppies at rescue centers

Suppose the adoption rate of the population of all rescue animals is $\mu = 0.5$ with a standard deviation of $\sigma = 0.1$. We want to test the hypothesis that the adoption rates of puppies is the same as the overall adoption rate of 0.5. We are told that the level of significance is $\alpha = 0.05$, i.e., the confidence level has to be at least 95%. Let us step through the hypothesis testing process.

1. **Null hypothesis:** There is no difference, i.e., the adoption rate for puppies is the same as other animals, namely 0.5.
2. **Sample data:** We sample 50 rescue centers across the country and compute the monthly puppy adoption rates for each center, shown in [Figure 6.6](#).
3. **Compute statistics:** The sample average of the puppy adoption rates of the 50 centers is $\bar{x} = 0.55$.
4. **Test statistic and sampling distribution:** From the Central Limit Theorem, we know that the sampling distribution of $\bar{x}$, the sample mean of puppy adoption rates of the 50 centers, is Normal with mean equal to that of the population μ and standard error given by $\sigma/\sqrt{n}$, where n is the sample size. In this case $\mu = 0.5$, $\sigma = 0.1$, $n = 50$ and the sample mean of rates of puppy adoption is $\bar{x} = 0.55$. In order to use the Standard Normal distribution, let us convert it to a Z-score.

- *Z-score* is given by:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = (0.55 - 0.5)/(0.1/\sqrt{50}) = 3.889.$$

- *Level of significance* is specified to be $\alpha = 0.05$.
- *Critical values:* The Z-score follows a Standard Normal distribution with mean=0 and standard deviation=1. We use the tables (or software tool or a Gemini prompt) and find that the critical value of the Standard Normal distribution corresponding to $\alpha = 0.05$ is $C_\alpha = 1.65$ and corresponding to $\alpha/2 = 0.025$ is $C_{\alpha/2} = 1.96$.

5. **Test condition (5) and conclusion (6):**

- *Two-tailed test:* If our alternate hypothesis is that the puppy adoption rate is *different*, then it can be either higher or lower than the adoption rate for all animals, we will use the critical value $C_{\alpha/2}$ since there are two possible alternatives and we split the level of significance equally between the two.

5a, 6a. Test condition and conclusion: Since 3.899 is greater than 1.96, we believe that the puppy adoption rate is different than that of the population rate of 0.5.

- *One-tailed test:* Suppose the manager of the rescue center wanted us to check if his intuition was true that the puppy adoption rates are significantly higher. Therefore, the alternate hypothesis is that the puppy adoption rate is *higher* than the adoption rate for all animals. We will now use the critical value C_{α} since there is only one possible alternative, and we do not need to split the level of significance.

5b, 6b. Test condition and conclusion: Since 3.899 is greater than 1.65, we believe that the puppy adoption rate is higher than the overall population adoption rate of all animals.

0.7781697 0.6877673 0.6024981 0.5040879 0.7259504 0.6733484 0.7160164 0.4422940 0.6744660 0.5890835 0.5220799 0.6402852
 0.5186424 0.3601101 0.6461421 0.5138404 0.6113813 0.6514516 0.6171862 0.6442120 0.6743958 0.4841051 0.5418434 0.6048428
 0.4950170 0.6706521 0.5519126 0.4318821 0.5074424 0.6135388 0.4841371 0.5956440 0.6984720 0.4984811 0.4188537 0.5085519
 0.5779552 0.7567739 0.6674821 0.5610327 0.3981226 0.5958107 0.5263741 0.5933253 0.5453897 0.5665697 0.5502761 0.6127282
 0.4143667 0.6201011

FIGURE 6.6: Rates of puppy adoptions at 50 rescue centers in the sample.



6.3 Tests Based on Sample Size

One of the assumptions of the Central Limit Theorem is that the *sample size is reasonably large*. What happens if the sample size is small? There are many situations where it is either expensive, damaging or difficult to collect data, e.g., collecting dirt samples from Mars ([Figure 6.7](#), left) or biopsies for diagnosing rare diseases.

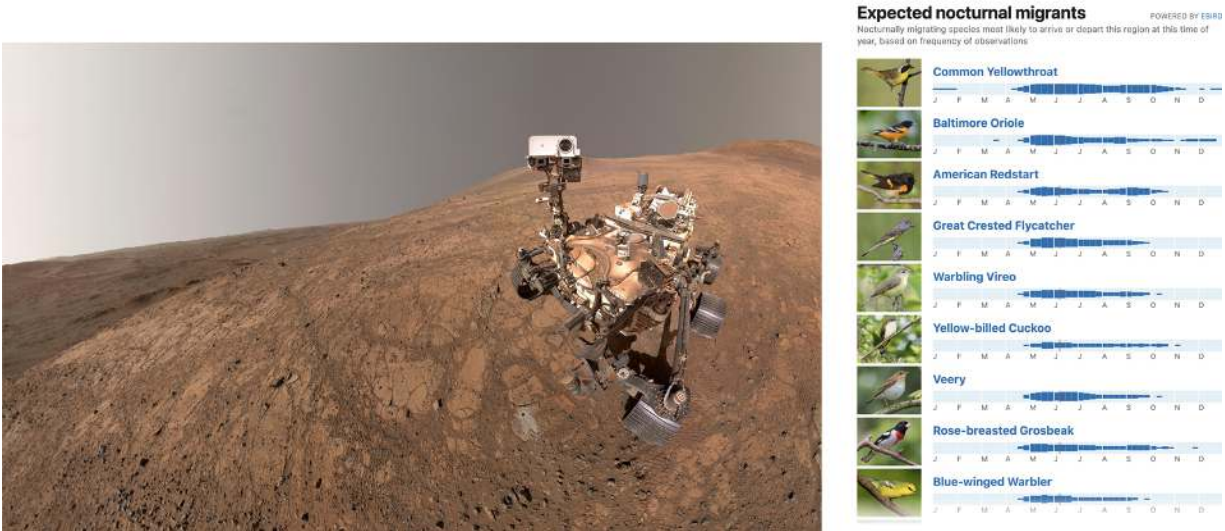


FIGURE 6.7: Left: NASA's Perseverance rover took this selfie on May 10, 2025, marking its 1,500th Martian day, or sol, exploring the Red Planet. The small dark hole in the rock in front of the rover is the borehole made when Perseverance collected a sample dubbed “Bell Island.” Collecting samples on Mars is expensive, so the sample size tends to be small (Photo: JPL). Right: Large samples are collected and analyzed by AI tool BirdCast, a real-time bird movement forecasting tool powered by the crowdsourced data collection resource, eBird. The picture on the right shows forecasts for bird migrations on June 9, 2025 in Morris County, NJ. [🔗](#)

6.3.1 Mars rover Perseverance: Small samples and Student's t -test

Consider a hypothetical situation where a Mars rover can only take 10 samples. It reports that each of the 10 samples of equal volume has 10, 12, 13, 13, 12, 10, 14, 15, 7 and 13 units of silicates and iron oxide, the minerals that make Martian soil seem red.

Since we only have $n = 10$ samples, we cannot use the Z-score test for sample means. *Student's t-test*, which was designed by William Gosset to test the quality of ale, is suitable for small samples. It is appropriate when the underlying distribution of minerals is Normal, and the standard deviation of the population is unknown. Suppose these assumptions hold, and we want to test whether the level of minerals is the same as that on earth, which is, say 10 in this hypothetical example. The null hypothesis is “there is no difference in the soil composition of Earth and Mars in terms of these minerals”, and we expect the mean of the 10 samples to be the same as Earth, namely $\mu = 10$.

The t -test statistic is given by:

$$t = \frac{\bar{x} - \mu}{\sigma(x)/\sqrt{n}} \quad (6.1)$$

$$= (11.9 - 10)/(0.814506 * \sqrt{10}) \quad (6.2)$$

$$= 2.58 \quad (6.3)$$

where μ is the Earth's mean, and $\bar{x}$ is the sample mean. We use the Earth's mean composition since the null hypothesis assumes that Mars has the same composition.

The population standard deviation $\sigma(x)$ is unknown and is estimated using the sample. When we estimate the standard deviation from the sample, we impose a constraint on the data that it must have that specific standard deviation, so effectively our sample size is only 9. This is called the *degrees of freedom* denoted by df . Using the t -distribution table with 9 df , we find that the critical values are 1.833 and 2.262 for the one- and two-tailed tests, at $\alpha = 0.05$ level of significance. The value of the test statistic 2.58 is greater than both the critical values; therefore, we conclude that the average silicate and iron oxide content in Martial soil is different (two-tailed test), and higher (one-tailed test) than that on Earth in this hypothetical example.

A special type of t -test is used when comparing pairs of related values. For example, measurements on the left and right eyes of a person, or symptoms of a patient before and after treatment. Such a test is called a *paired t -test* and involves applying a t -test to the difference in the paired values of measurements.

GeminiBox 6.3.1: Student's t -test

When should we use Student's t -test? What's a paired t -test? 

GeminiBox 6.3.2: Student's t for larger samples

What happens to the t -test when sample sizes are larger? 

Data Debate 6.3.1 *In statistical hypothesis testing, how do you decide whether to use a t -test or a Z-score test?*

6.3.2 BirdCast AI: Large samples and the Z-score test

[BirdCast](#) is an AI tool for bird migration developed by researchers at Cornell, Colorado State and University of Massachusetts, Amherst, in partnership with EPA, NASA, NSF, Audubon Society and many others. It enables real-time tracking and forecasting of bird movements powered by a birding and reporting tool, [eBird](#). Thanks to the automation and crowdsourcing powered by eBird, data is plentiful and sample sizes are large. The Z-score test illustrated in [Example 6.2.3](#) applies.

Suppose, hypothetically, that the average number of birds migrating through Morris County on June 9 is expected to be 300 per hour based on historical observation. We might wish to see if June 9, 2025, follows this pattern.

1. Null hypothesis: The number of migrating birds spotted each hour in Morris County, NJ, on June 9, 2025 is the same as historical data, 300 per hour.
2. The level of significance is specified to be $\alpha = 0.05$.
3. We collect 100 samples, at different locations in Morris County, at different hours of the day.
4. The average of these hundred samples, $\bar{x}$, is 298, with a standard deviation $\sigma(x)$ of 11.
5. The test statistic is:

$$Z = \frac{\bar{x} - \mu}{\sigma(x)/\sqrt{n}}.$$

Its sampling distribution, according to the Central Limit Theorem, is a Standard Normal distribution with mean = 0 and standard deviation 1.

6. The critical region corresponding to the Standard Normal distribution for $\alpha = 0.05$ is $|Z| > 1.65$ for a one-tailed test and $|Z| > 1.96$ for a two-tailed test.

7. In this case, $\mu = 300$, $n = 100$, $\bar{x} = 298$ and $\sigma(x) = 11$. Therefore,

$$Z = \frac{298 - 300}{11/\sqrt{100}} = -1.8.$$

8. Since $|Z| = 1.8$, it is > 1.65 , so we cannot conclude that the hourly migratory rate is the same as past years in Morris County for a one-tailed test. It is presumably lower, because $\bar{x}$ of the sample, $298 < 300$.

9. However, $|Z| = 1.8 < 1.96$, so for a two-tailed test it is not significantly different.

The previous example shows that it is important to state the *alternate hypothesis* before conducting the test, rather than adjust conclusions after observing the data, which would be tantamount to *data-dredging* described in [Section 6.6](#). If our alternate hypothesis had stated that “the hourly migratory rates in Morris County in 2025 are lower than previous years”, we would have used a one-tailed test and found evidence that the rates were indeed lower. However, if we have no *a priori* reason to believe that, we should use a more conservative alternate hypothesis “the hourly migration rates are different,” and our test would find no reason to indicate that.

Data Debate 6.3.2 *What should we do if a test statistic is significant for a one-tailed test but not for a two-tailed test? Is it ethical to formulate the alternate hypothesis after the fact?*

GeminiBox 6.3.3: Applications of Hypothesis Testing
--

How is hypothesis testing applied in real life? 

6.3.3 A/B testing: Tests based on more than one sample

Websites embed subtle behavioral cues that we respond to without even being aware. These cues are chosen carefully through A/B testing. Social media companies use A/B testing extensively and continuously to decide between design elements and their placement to maximize user engagement. E-commerce companies tweak wording, font size to display price, and placement of images to optimize sales. AI tools enable collecting large amounts of data in real-time, and testing hundreds of thousands of hypotheses to compare pairs of scenarios to choose the one that provides the best desired outcome. This process is called *A/B testing*. The image below presents a hypothetical example of two identical scenarios, except for the “Learn more” button. Just by changing the color, and including an arrow icon, the click rate went from 52% to 72%. Companies constantly assess customers' behavioral psychology and preferences through such tests. *Statistics for Experimenters* by Box *et al.* [6] is a classic reference for practitioners of hypothesis testing.

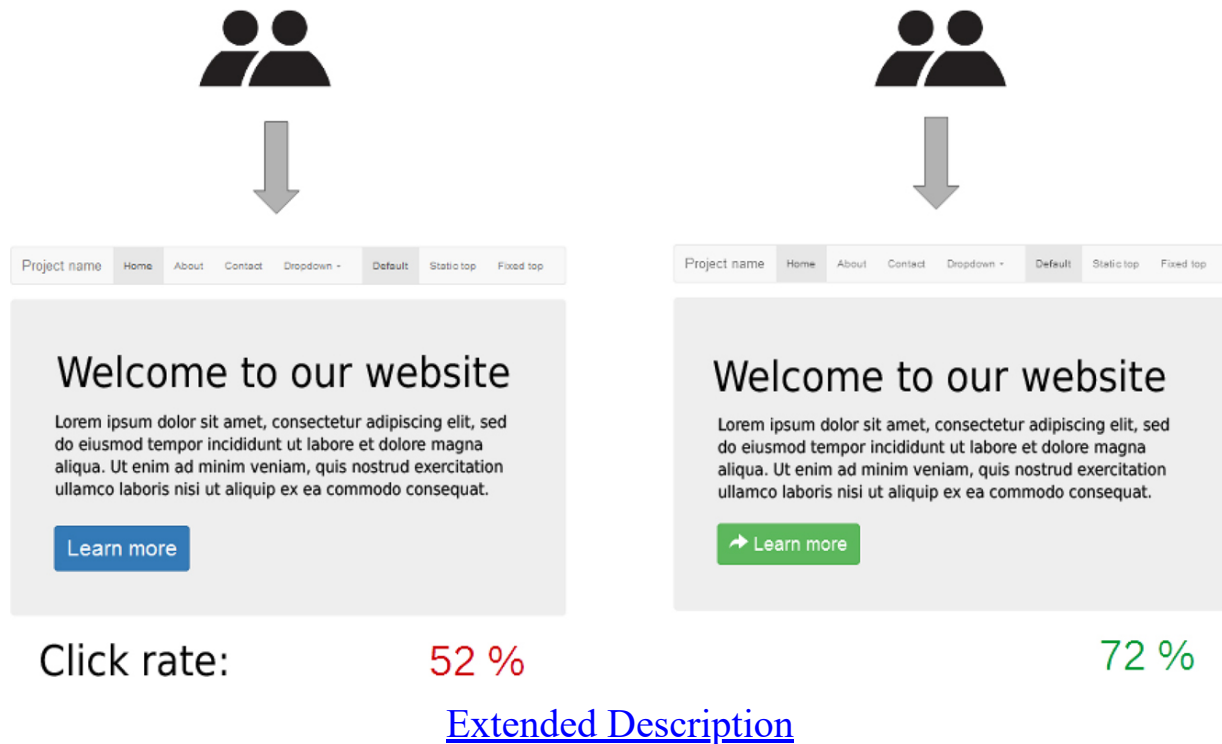


FIGURE 6.8: Social media companies and e-commerce websites assess behavioral psychology and user preferences through A/B testing. Image: Maxime Lorant, CC BY-SA 4.0. 

To illustrate A/B testing, consider the example of our animal rescue center, which now wants to raise funds through soliciting monetary donations. It tries two different icons, a puppy and a cat, on different webpages and compares the proportion of users who actually click and donate money via the two types of webpages. This is called a *two-sample test*, since we compare two different samples: monetary donations from webpages with a cat icon and monetary donations from webpages with a puppy icon. There are many variations of this two-sample test. We describe a simple version below, where the two samples are independent yet collected under comparable conditions.

How are two-sample tests used in A/B testing?

Suppose we compare two situations: A webpage where a cat icon is placed next to the “Donate \$” button, and a webpage where a puppy icon is placed next to the “Donate \$” button. The rescue center would like to know which page attracts more \$ donations for its fundraising drive.

Let the number of people donating money on the webpage with the cat icon be x_1 out of n_1 people visiting the site. Let the number of people donating money on the webpage with the puppy icon be x_2 out of n_2 people visiting the website. [Figure 6.9](#) shows the sampling distribution of the proportion of \$ donations for cat icon (blue) and puppy icon (red) websites, respectively. The combined distribution of \$ donations for cat and puppy icon websites combined together is given by the dotted gray distribution in the middle. The bell-shaped distributions were created using R code and simulating $N(0, 1)$, $N(2, 1)$ distributions with 10,000 points each. Since proportions are averages, their sampling distribution is a Normal distribution, and so is the difference of two Normal distributions. Let the null hypothesis be “there is no difference in proportion of people donating money whether the website has a cat icon or a puppy icon”. Let the alternate hypothesis be “cat icon attracts a smaller proportion of \$ donations compared to the puppy icon.” That is, if p_1 is the proportion of donors on the cat pages, and p_2 is the proportion of donors on the puppy pages, then:

$$Null : p_1 = p_2,$$

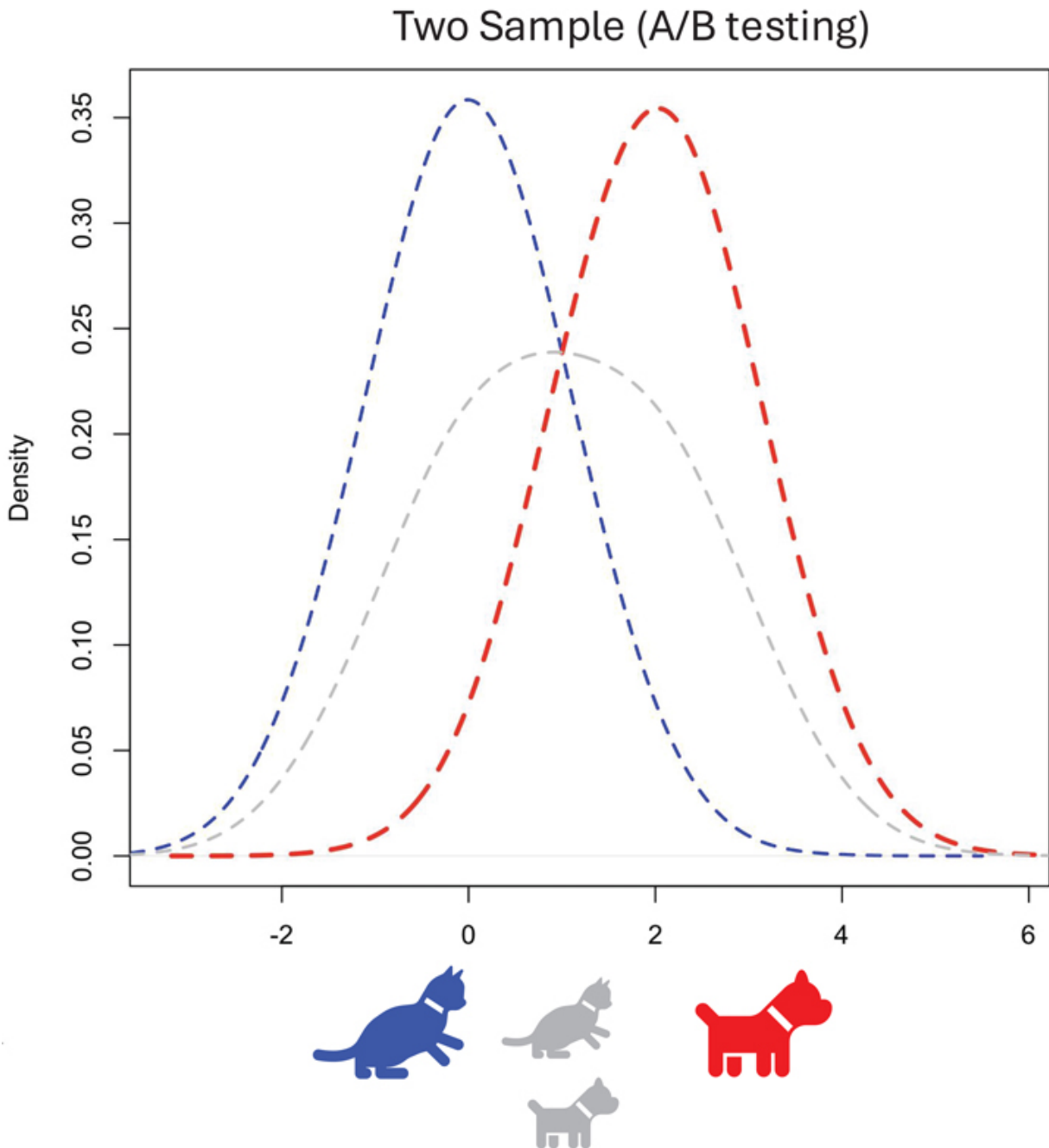


FIGURE 6.9: Two-sample test to compare the \$ donation rates for webpages with a cat icon versus a puppy icon. The bell-shaped distributions were created using R code and simulating $N(0, 1)$, $N(2, 1)$ distributions with 10,000 points each. The gray curve is the distribution of the combined data. [!\[\]\(7b12ec317661b6e6f66559223715f56c_img.jpg\)](#)

and alternate:

$$\textit{Alternate} : p_1 < p_2$$

Compute $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$, where the hat notation implies that it is an estimate computed from a sample. Let the combined proportion of the two samples pooled together (the gray distribution) be:

$$\hat{p} = \frac{(x_1 + x_2)}{(n_1 + n_2)}.$$

The Standard error of the combined proportion of the pooled sample is given by:

$$\sqrt{(\hat{p}(1 - \hat{p}))\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

This is a complicated looking expression but has its validity in the way variance and standard deviation behave as functions. For those interested in how this expression is obtained, please consult [35]. To learn more about standard error for testing the sample proportions from two samples, check [GeminiBox 6.3.5](#).

GeminiBox 6.3.5: Two-sample standard error [↗](#)

What is the standard error for comparing two sample proportions assuming that the population proportions are equal?

Then, under the null distribution, we would expect the difference to be 0, and that

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{(\hat{p}(1 - \hat{p}))(\frac{1}{n_1} + \frac{1}{n_2})}}$$

is a Standard Normal variable with mean 0 and variance 1, with critical values 1.65 and 1.96 for one- and two-tailed tests at $\alpha = 0.05$ level of significance.

If $n_1 = 100, n_2 = 101, x_1 = 25, x_2 = 40$, then:

$$\hat{p}_1 = 25/100 = 0.25$$

$$\hat{p}_2 = 40/101 = 0.396$$

$$\hat{p} = (25 + 40)/(100 + 101) = 65/201 = 0.32$$

$$\begin{aligned} Z &= \frac{(0.25 - 0.396) - 0}{\sqrt{(0.32(1 - 0.32))(\frac{1}{100} + \frac{1}{101})}} \\ &= -2.219 \end{aligned}$$

For the Standard Normal distribution, the critical value for a one-tailed alternative hypothesis that a cat icon attracts fewer proportion of money donations is $C_\alpha = -1.65$. Since the Z-score of -2.219 is less than -1.65 , we infer that the proportion of people donating money on websites with cat icons are lower than that of websites with puppy icons.

Data Debate 6.3.3 *How would the animal rescue center use the information that webpages with cat icons attract less donations than webpages with puppy icons?*

This is just a simple example. If you'd like to explore more, use the prompts in [GeminiBoxes 6.3.4](#) and [6.3.6](#).

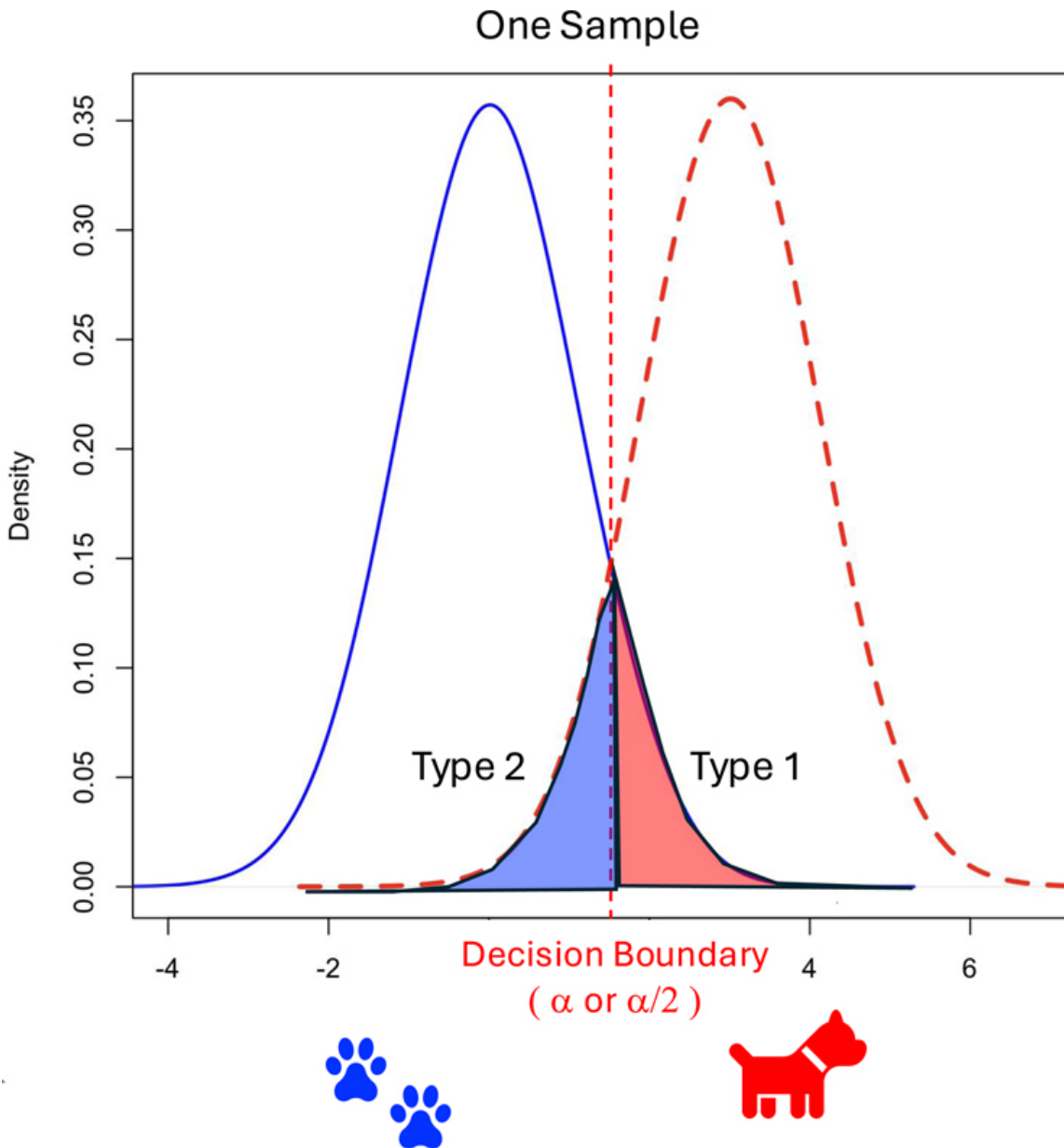
GeminiBox 6.3.6: Two-sample tests

What are the different types of two-sample tests? 

6.4 Could We Have Made a Mistake? Type 1 and Type 2 errors

We have discussed that the level of significance α is the probability of finding an unusual sample value by chance from the null distribution, a worst case scenario of error. A mistake implies that we came to the wrong conclusion of rejecting the null hypothesis based on an abnormal sample. For example, consider [Figure 6.10](#). It is possible that the particular sample we chose might have been an exception with abnormal values even though its sampling distribution is genuinely the null distribution (blue), $P_{\mathcal{N}}(X)$. Such an abnormal sample leads us to the *false conclusion* that puppies are adopted at a higher rate. Such a mistaken conclusion is called a *Type 1* error in statistics and a *false positive* in machine learning. The probability of a Type 1 error is represented by the red shaded region to the right, and is often prespecified to be no more than the level of significance α . It represents the probability of an abnormal sample if the null hypothesis is true. It is given by:

$$P_{\mathcal{N}}(X > C_{\alpha}) = \alpha.$$



[Extended Description](#)

FIGURE 6.10: Type 1 error or false positive: Reject the null hypothesis due to an abnormal sample value from the null distribution. Type 2 error or false negative: Mistakenly accept the null hypothesis (i.e., reject the alternate hypothesis) due an abnormal sample from the alternate distribution. [↩](#)

On the other hand, suppose the alternate hypothesis that puppy adoption rate is genuinely higher as denoted by the sampling distribution $P_{\mathcal{A}}(X)$, the

red dashed curve, is true. Then a particular sample might have unusually low adoption rates just by chance, and would lead us to conclude that there is no difference in the adoption rates of puppies and overall adoption rates of all animals. This is called a *Type 2 error* in statistics and a *false negative* in machine learning, where a genuinely different sample from the alternate distribution of higher puppy adoptions, $P_{\mathcal{A}}(X)$ distribution, is mistakenly declared not to be different. The probability is given by a sample from the alternate distribution being unusually small, i.e., less than the boundary of the critical region, C_{α} :

$$P_{\mathcal{A}}(X < C_{\alpha}) = \beta.$$

Ideally, we would like both types of errors to be small. But [Figure 6.10](#) shows that it is a trade-off. Sliding the decision boundary of the critical region (dotted red line) to reduce the Type 1 error implies an increase in Type 2 error and vice versa, in this particular type of scenario of symmetric, unimodal distributions. [Section 6.4.1](#) has a nice example of this trade-off. You can learn more about different types of errors by using the prompts in [GeminiBox 6.4.1](#).

GeminiBox 6.4.1: Type 1, Type 2 and misclassification errors [!\[\]\(9a8373782c8e0007b8363c731473b178_img.jpg\)](#)

What are Type 1 and Type 2 errors in hypothesis testing and how are they related to misclassification error in machine learning?

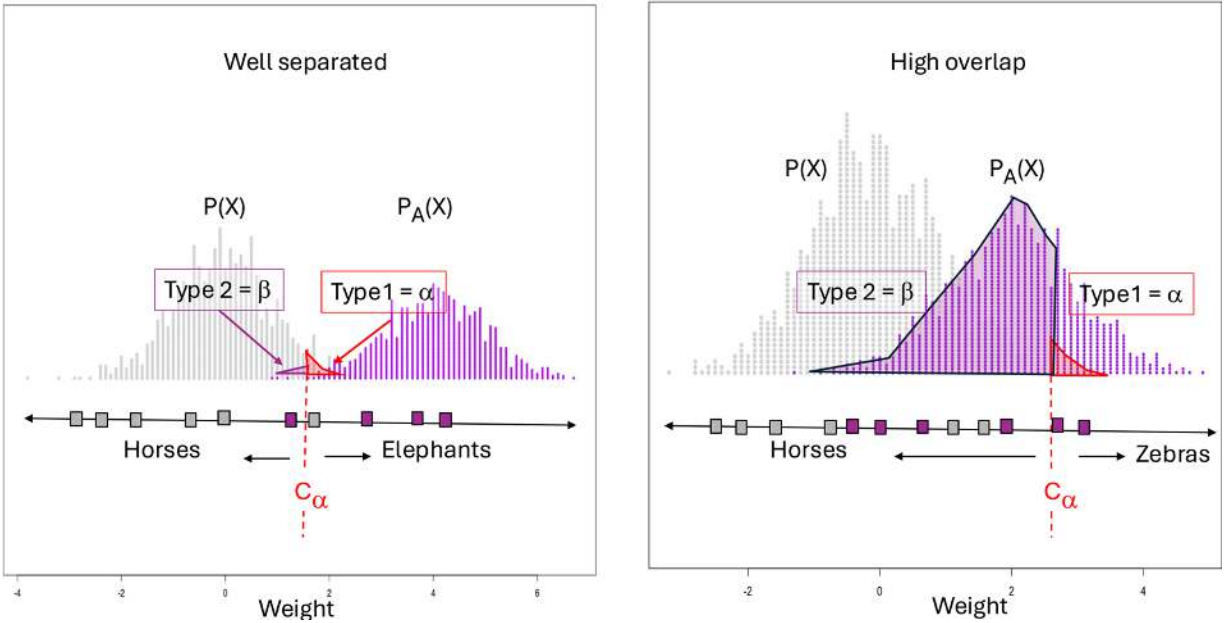
Data Debate 6.4.1 *Medical treatments often involve side effects. It is important to be careful when using hypothesis testing in healthcare. If a sick person is untreated, they might die. If a healthy person is treated with medicines, they might develop side effects resulting in serious health conditions.*

1. *What are the Type 1 and Type 2 errors in the above example?*
2. *Give an example where a Type 1 error is more dangerous than the Type 2 error.*
3. *Give an example where a Type 2 error is more dangerous than the Type 1 error.*
4. *If the two distributions in [Figure 6.9](#) are far apart, is it easy or hard to control errors?*

6.4.1 Power of a test

Suppose you are given the weights of two groups of animals. One group is horses. The second group is either elephants or zebras. Which one is easier to tell apart from horses based on weights?

It is intuitively clear that elephants have much higher weights, so it should be easier to tell whether the mean of an observed sample of weights (represented by squares on the X -axis in [Figure 6.11](#)) is that of elephants (purple) or horses (gray). On the other hand, a sample of weights of zebras (purple) might be harder to tell apart from that of horses.



[Extended Description](#)

FIGURE 6.11: The sample mean has a high power of distinguishing between well-separated null ($P_{\mathcal{N}}(X)$) and alternate hypotheses ($P_{\mathcal{A}}(X)$) (e.g., distribution of weights of horses and elephants) but low power when the two distributions have a high overlap (e.g., distribution of weights of horses and zebras). ↩

When the null distribution (weights of horses) and alternate distribution (weights of elephants) are far apart as shown in [Figure 6.11](#), left, then the overlap is very little and the chances of classifying elephants as horses based on a sample of weights, i.e., Type 2 error, is small. On the other hand, when the distributions have considerable overlap, like the weights of horses and zebras shown in [Figure 6.11](#), right, then there is a greater chance of mistaking the sample of weights of zebras for those of horses, i.e., the Type 2 error is high.

The ability of a test statistic to discriminate between the null and alternative hypotheses is quantified by its *power*. It is the area to the right of

the critical value C_α under the distribution of the alternative hypothesis, $P_{\mathcal{A}}(X)$.

$$\text{Power} = P_{\mathcal{A}}(X > C_\alpha),$$

where X is the random variable that represents the sample mean, and the probability is computed using the alternate distribution, $P_{\mathcal{A}}(X)$. The greater this area, the greater the power.

6.4.2 Relationship to Type 2 error

Recall that Type 2 errors are false negatives, i.e., the probability that we accept the null hypothesis when the alternate hypothesis is true.

$$\text{Type 2 Error} = \beta = P_{\mathcal{A}}(X < C_\alpha).$$

This also implies that:

$$\text{Power} = P_{\mathcal{A}}(X > C_\alpha) = 1 - P_{\mathcal{A}}(X < C_\alpha) = 1 - \beta.$$

In hypothesis testing language, the weight distributions of horses and elephants are *well-separated*, and a sample mean will have a high power to distinguish between the two, and Type 2 error will be low. On the other hand, the weight distributions of horses and zebras are quite similar, and it will be harder to distinguish between the two distributions based on the sample mean of weights. That implies that a test based on the sample mean in this situation will have a high Type 2 error, reflected in the big chunk of the purple distribution in [Figure 6.11](#), right, and as a consequence, low power.

ReferenceBox 6.4.1: Types of Errors and Power of a Test
--

- *Type 1 error (False Positive)* happens when a sample from the null distribution is mistakenly declared to be from the alternate distribution since it exceeds the critical value, C_α . This happens if the sample is abnormal just by chance. For example, a set of unusually heavy horses is declared to be from a population of elephants based on weight. The probability of Type 1 Error is α , the level of significance:

$$P_{\mathcal{N}}(X > C_\alpha) = \alpha.$$

- *Type 2 error (False Negative)* is when a sample from the alternative hypothesis distribution is declared to be from the null hypothesis distribution. This happens when the sample from the alternate distribution has unusually small values, less than the critical value C_α , but values that are typical for the null distribution. Type 2 error, β , is given by:

$$P_{\mathcal{A}}(X < C_\alpha) = \beta$$

- *The power of a test* is the ability of the test statistic to discriminate between samples from the null $P_{\mathcal{N}}$ and alternate $P_{\mathcal{A}}$ distributions. The power of a test based on the sample mean is typically high when the distributions are well separated (weights of horses and elephants) and low when there is considerable overlap between the two (weights of horses and zebras). Power is given by:

$$P_{\mathcal{A}}(X > C_\alpha) = 1 - \beta.$$

6.5 Nonparametric Tests for Empirical Distributions

The tests we have discussed thus far are based on the Normal distribution and Student's t -distribution. These are called *parametric tests* because we assume a prespecified distribution, e.g., the Normal, and test hypotheses about its parameters such as the mean μ and standard deviation σ . While the Normal distribution works for the sampling distribution of the sample mean because of the Central Limit Theorem, it does not work in many scenarios where the underlying distribution is not Normal. In such instances, it is possible to use tests that are purely data-driven and do not make assumptions about the underlying distribution. Such tests are called *nonparametric tests*.

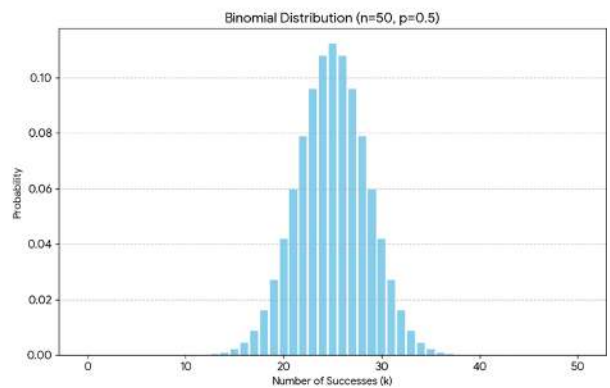
There are many nonparametric tests. Some are based on the ranks of ordered data (Mann-Whitney or the Wilcoxon rank sum test), and some use the empirical distributions (Kolmogorov-Smirnov) to test whether two groups of data have the same probability distribution. *Robust nonparametric statistical methods* by Hettmansperger and McKean [20] provides a strong grounding in nonparametric tests. We will discuss a simple, single sample test for testing the median of population using a sample.

6.5.1 StreetStyle, AI tool for fashion: Sign test for the median

The fashion industry has been a big adopter of AI, using it to generate design ideas, to forecast trends and to understand customer preferences. AI has also been used to detect global fashion trends as explained in *The Cornell Chronicle* article that's available in [Section 6.7](#). The AI tool StreetStyle [28] uses computer vision and pattern recognition along with

expert knowledge to identify concentrations of a particular type of clothing, and analyze trends, e.g., trucker hats creeping onto fashion runways.

A *sign test* is a popular tool in fashion. Suppose we wish to test the hypothesis that the median weight of a particular type of fashion accessory, e.g., a headpiece shown in [Figure 7.3](#), is m . We do not know whether the distribution of the weight of the headpiece is bell-shaped or not. However, if m is indeed the median, then if we select a sample of n headpieces independently of each other, we would expect half of the data to be less than m in value and half to be greater, because the median divides the rank-ordered data into two halves. Therefore, if we subtract the median m from each value of the sample, we would expect half the values to be negative and half to be positive. If we ignore the numeric data values and focus just on the *sign* of the difference, we would expect that roughly half the differences would have a positive sign and half of them a negative sign.



[Extended Description](#)

FIGURE 6.12: AI StreetStyle can identify global fashion trends in headpieces. Left: A headpiece from the Kathakali art form in India. (Photo: Maneesha Shahani ). Right: Testing whether the median weight of a Kathakali headpiece is 1.75 pounds using a sign test. Gemini generated Binomial distribution with $n = 50$ and $p = 0.5$ .

This is exactly like a coin toss where the probability of heads or tails is 0.5. Therefore, the number of samples with positive signs (number of sample values that exceed the median value), out of a total sample of n headpieces, will be governed by the Binomial distribution discussed in [Chapter 5](#). Formally:

$$W \sim B(n, p), \quad p = 0.5$$

where W is the total number of positive signs in n samples, and $B(n, p)$ is a Binomial distribution of n samples with probability $p = 0.5$ of an independently sampled point being greater than the median m . Note that counting the number of positive (or negative) signs is equivalent to counting the number of headpieces whose weight exceeds (or less than) the median value, m . Let us work through an example.

6.5.2 Illustrative example: Testing if Kathakali headpieces typically weigh 1.75 pounds

We break down the steps for using the sign test.

1. Null hypothesis: The median weight of Kathakali headpieces is 1.75 pounds.
2. Collect data, i.e., the weights of 50 sample headpieces: 2.772383 1.775585 2.015260 1.366918 1.237532 1.814544 1.670417 1.111142 2.474824 1.515349 1.510134 1.897743 2.351927 2.277978 2.862382 1.324925 1.606421 2.420627 1.189701 2.359185 1.678522 2.231947 2.220179 1.402802 2.961335 2.031849 2.759934 2.460768 2.087850 1.739806 1.128819 2.975615 2.353124 1.944876 1.813232 1.615208 2.335169 1.065633 1.963228 2.498850 2.131604 1.817003 2.002203 2.479088 2.469526 1.019551 1.174993 2.464324 2.516704 1.765256. (

This sample data was generated using the command “runif(50, 1.0, 3.0)” in *R*.)

3. Define the test statistic: Subtract the hypothesized median value 1.75 from each weight; find W , the number of positive signs.
4. Specify the sampling distribution of the test statistic: Binomial where $n = 50$ and $p = 0.5$ as discussed in the sign test.
5. Prespecify the level of significance, e.g., $\alpha = 0.01$. The corresponding critical region in the Binomial distribution from an online calculator is $[32, 50]$, where the probability of getting more than 32 positive values,

$$P(W > 32) = 0.0164.$$

Since the Binomial is a discrete distribution, it might not have a value that corresponds exactly to $\alpha = 0.01$. Gemini, or any statistical software tool can be used to determine the critical region.

6. Compute the test statistic and its probability under the sampling distribution: For this sample, the value of the test statistic is the number of positive signs, $W = 33$. The Binomial distribution tables indicate that the probability of getting $W = 33$ or more positive signs based on 50 independent samples is 0.0077. This is the p -value.
7. Apply the test: Compare the sample test statistic ($W = 33$) and its p -value (0.0077) under the null sampling distribution to the critical values $C_\alpha = 32$ and the level of significance $\alpha = 0.0164$.
8. Conclusion: Since the p -value is $0.0077 < 0.016$, or equivalently the sample value of $W = 33 > 32$, the probability that the sample is from the null distribution is very low (0.0077), and so the median weight of

the Kathakali headpiece is very likely not 1.75 pounds. We cannot accept the null hypothesis at a $\alpha = 0.01$ level of significance.

6.6 Pitfalls of Hypothesis Testing: Data Dredging

In this chapter, we have discussed hypothesis tests that involve numeric variables, analyzed one at a time. In the next chapter that focuses on relationship between multiple variables, we will study two additional hypothesis testing topics, Chi-square tests of independence and Analysis of Variance (ANOVA), that focus on categorical variables and their interaction with other variables.

We conclude this chapter with a brief caution about the pitfalls of hypothesis testing, particularly in the age of AI where data and computing power have exploded, enabling the testing of hundreds and thousands of hypotheses each second.

Research studies often focus on multiple attributes or multiple groups and are interested in answering multiple questions simultaneously, called a *family of tests*. It is possible that a single sample is used for testing all the hypotheses, or multiple samples are used. Regardless, the focus is on *multiple hypotheses* being tested simultaneously. Matt Sosna's article in [Section 6.7](#) explains the dangers of multiple testing. We describe it briefly here.

Suppose we are interested in testing the connection between the color of jelly beans and acne. Suppose that the jelly beans come in 20 colors. We expect (null hypothesis) that there is no association between any given color of jelly beans and acne, at the level of significance $\alpha=0.05$. That is, the chances of getting a false positive, by sheer chance and variation in

samples, are 5% or 0.05 for any of the 20 colors. That is, the probability of accepting the null hypothesis of no association between a given color of jelly beans and acne when it is actually true is 95%.

If the 20 tests, one for each color, are independent of each other, then the probability of correctly accepting the null hypothesis in ALL the 20 tests, using the probability rule of multiplication probabilities of independent events is:

$$(0.95)^{20} = 0.36$$

The probability of the complementary event of getting at least one false positive in 20 tests is:

$$1 - (0.95)^{20} = 1 - 0.36 = 0.64.$$

Therefore, the random chance of getting a false conclusion for at least one of the 20 colors is 64%. The more tests you run, the more likely the chances of a false positive. This is how spurious associations show up!

The process of deliberately running multiple tests and choosing to report the few false positives that occur by chance as a genuine rejection of the null hypothesis (“green jelly beans cause acne!”) is called *data dredging* or *p-hacking*. By continuously running tests, we are guaranteed that one false positive with a small *p*-value shows up by chance. Explore [GeminiBox 6.6.1](#) to learn more.

GeminiBox 6.6.1: Data dredging

“What is data dredging?” 

The journalist John Bohannon, pranked the media by running a study with a small sample (15) and testing 18 hypotheses with 18 variables until

one of them showed up as negating the null hypothesis just by chance. His study on how chocolates help weight loss was reported and amplified by the news media until it reached far corners of the world. Read his article *I fooled millions into thinking chocolate helps weight loss* in [Section 6.7](#), and a related article from NPR titled, *Why A Journalist Scammed The Media Into Spreading Bad Chocolate Science*.

6.7 Read & Reflect

- *Hypothesis testing for AI audits.* 
 - *Insight - Amazon scraps secret AI recruiting tool that showed bias against women.* 
 - *AI tools show biases in ranking job applicants' names according to perceived race and gender.* 
 - *Dust devil photobombs Perseverance's selfie.* 
 - *AI tool detects global fashion trends.* 
 - *Visualizing the danger of multiple t-test comparisons* 
 - *I fooled millions into thinking chocolate helps weight loss.* 
 - *Why a journalist scammed the media into spreading bad chocolate science.* 
-

6.8 Exercises

1. **Quality control: One sample t -test.** The Bubbling Springs beverage company claims its sparkling water cans contain 12 ounces of water. A researcher randomly samples 30 of their cans and measures the amount of fluid in each one. Here is the data from the study:

Data in fluid ounces: 11.78, 11.87, 11.98, 11.78, 11.60, 12.16, 11.67, 11.84, 11.74, 11.81, 11.67, 11.67, 11.80, 11.86, 12.08, 11.71, 11.65, 12.08, 11.84, 12.18, 11.64, 12.09, 12.18, 12.04, 11.53, 11.67, 11.86, 11.70, 11.87, 11.46.

The researcher wants to find out if the claim made by the company is accurate.

- (a) What statistical test can we use to verify or reject this claim?
 - (b) Create a histogram using this data. Can you tell if the data are normally distributed by looking at this graph?
 - (c) We want to use the t -test to verify the claim; does this study meet the conditions? Explain.
 - (d) Find the mean and standard deviation for the data.
 - (e) State the null and alternative hypotheses.
 - (f) Using a significance level 0.05, test the hypothesis.
 - (g) What is the conclusion? Explain.
2. **Hypothesis testing with paired data.** The following questions are based on data from “Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Reduction,” by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1 .

- (a) When 40 people used the Weight Watchers diet for one year, their mean weight loss was 3.0 lb and the standard deviation was 4.9 lb.
- Use a 0.01 significance level to test the claim that the mean weight loss for the population is 0 lb.
 - State the null and alternate hypotheses.
 - Which statistical test would be appropriate for this data? Explain.
 - Based on these results, does the diet appear to have statistical significance?
- (b) When 36 people used the Atkins diet for one year, their mean weight change was 8.1 lb and the standard deviation of weight change was 4.5 lb.
- Use a 0.05 significance level to test the claim that the mean weight loss is greater than 5 lb.
 - State the null and alternate hypothesis.
 - Conduct the appropriate statistical hypothesis test for significance of the claim.
 - Sketch the distribution of the test statistic; label and shade the critical region.
 - What is the conclusion? Explain if we should reject or keep the null hypothesis.

3. **Applying Student's t -test: Nutrition of breakfast cereals.** The following data set from [kaggle.com](https://www.kaggle.com) displays the nutrition data on 80 popular breakfast cereals. 

Attribution: These data sets have been gathered and cleaned up by Petra Isenberg, Pierre Dragicevic and Yvonne Jansen.

This Google spreadsheet has the data from the study: [🔗](#)



About 1 in 5 Americans say their diets are extremely or very healthy

% of U.S. adults who say that overall, their diet is ...

Extremely/ Very healthy	Somewhat healthy	Not too/Not at all healthy
21	59	20

Note: Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Feb. 24-March 2, 2025.
"Americans on Healthy Food and Eating"

PEW RESEARCH CENTER

Extended Description

FIGURE 6.13: Shopper checks ingredients on boxed cereal. Wikimedia Commons. US Department of Agriculture. Office of Governmental and Public Affairs.

- Randomly select a sample of 25 cold cereals from the data set.
 - Create a histogram of the calories in 1 serving of cereal from your sample data. Is this data normally distributed?
 - Find the mean and standard deviation of the calories.
 - Test the following statement for level of significance of $\alpha = 0.05$ "the mean number of calories in one serving of cold cereal is 100 calories", using the Student- t as your test statistic.
 - State your null hypothesis and alternate hypothesis.
 - Will this be a two-tailed or one-tailed test? Why?
 - What is the conclusion? Explain.
4. **Choosing the right test based on conditions and data.** Which type of hypothesis test would be appropriate for the following scenarios? You

do not have to conduct the hypothesis testing, but explain based on the type of data and the conditions mentioned, what kind of test is best suited and how you came to the decision (hint: think about paired, unpaired, what are the conditions required for each type of test, if the data is parametric or nonparametric). State the null and alternative hypotheses for each scenario.

- (a) Two groups of 10 plants each were treated with different fertilizers: Fertilizer A and Fertilizer B. The heights of the plants were measured after 4 weeks. Are the fertilizers significantly different?
- (b) The number of grams of protein in eight different brands of protein bars is measured. Is the difference in grams of protein significant?
- (c) A group of forty 3rd grade children are practicing with an online reading program. Their reading levels were measured before they started the program and after 10 weeks. Is there a significant difference in their reading levels?
- (d) The high and low temperatures of each day during the month of July were recorded. Test the claim that there is an average difference of 30 ° F in temperature.
- (e) Mrs. Murthy teaches two Statistics classes. The average quarterly grade for one class is 80 and for the other class is 84. Is the performance of one class significantly better than the other?
- (f) An airline carrier is collecting data about the flight experiences of the passengers. Passengers are rating various factors on a scale of 1 to 10. The airline claims that its average score on cleanliness and service during flight is greater than 8. How can you test if this claim is true or false?

- (g) A video game company wants to test two versions of a new game and users' data on how much time users play the game (for both versions) is collected. How can they decide which version is better?
- (h) A clothing design company wants to find out if the new tennis outfit they have designed is stylish and comfortable. They ask customers who have purchased the outfits to rate it on their website. How can they analyze this data? Explain in detail.

5. Interpretation of output from a two sample t -test, assuming unequal variances. For this exercise we will use the data set from [kaggle.com](https://www.kaggle.com). Here is the link to the data set: 

Using the XLMiner analysis ToolPak from Google Sheets, a two-sample t -test was conducted on a sample from this data set. [Table 6.1](#) displays the results. Examine the table carefully and answer the questions below:

- (a) The data for which years have been used for this analysis?
- (b) What would be the null hypothesis and alternate hypothesis for this test?
- (c) Under what conditions would you use a two-sample t -test with unequal variances?
- (d) Make a sketch of the two-tailed test using the output from the table. Label the mean and critical values. Shade the critical region and explain if we should reject or keep the null hypothesis.

TABLE 6.1: t -Test results for average February temperatures from 2000 to 2015. Predictions were provided by the Punxsutawney Groundhog Club, and the average monthly temperatures were published by NOAA. 

t -Test: Two-Sample Assuming Unequal Variances
--

t-Test: Two-Sample Assuming Unequal Variances		
From year 2000 to 2015		
	Avg.Feb. Temp. Midwest	Avg. Feb. Temp. Pennsylvania
Mean	32.95625	27.50625
Variance	23.27329167	24.09129167
Observations	16	16
Hypothesized Mean Difference	0	
df	29	
t Stat	3.167594885	
$P(T \leq t)$ one-tail	0.001802148735	
t Critical one-tail	1.699126996	
$P(T \leq t)$ two-tail	0.00360429747	
t Critical two-tail	2.045229611	

6. **Two-sample paired t -test.** The math chair of a school district wants to test the effectiveness of an online math tutorial for students. She randomly selects a class with 15 students, and they practice with the online program as directed for 6 weeks. The developers of the online math program claim that students who use this program will see an increase in 5 points in their test scores. The test scores of the students before and after the intervention are listed in [Table 6.2](#):

- What statistical test can we use to test if this claim is true or false based on our sample data?
- State the null and alternate hypotheses.
- Find the mean and standard deviation of the difference in scores.

- (d) Find the test statistic.
- (e) What is the possibility of a Type 1 or Type 2 error? Explain.
- (f) Based on this data, is the online math program effective in increasing math scores? Why?

TABLE 6.2: Test scores before and after intervention along with their differences. [↗](#)

Test scores before intervention	Test scores after intervention	Difference in scores
40	42	2
43	45	2
36	35	-1
35	42	7
41	40	-1
29	33	4
42	42	0
44	47	3
36	38	2
38	35	-3
27	33	6
41	41	0
41	46	5
43	45	2
45	48	3

7. Hypothesis testing using Excel. In this exercise, we will use the data from [kaggle.com](https://www.kaggle.com) about the animated series: Popular Anime  which

has the viewer scores for various genres of Anime TV serials and movies. Anime is characterized by stylized art as shown in [Figure 6.14](#). Which genre is more popular? Which type (TV or Movies) is more popular?

- (a) Download and copy the data popular-anime.csv from the website [kaggle.com](https://www.kaggle.com) into an Excel spreadsheet or Google Sheets.
- (b) For this exercise we will only use some of the columns and rows, as this is a very large data set. Copy the data from the csv file to an Excel spreadsheet. Copy and paste the columns: id, name, type, genre and score into sheet 1.
- (c) Sort the sheet by type. Delete rows that contain data about special, music, ona or ova. Organize the spreadsheet by formatting using the Excel features.
- (d) Sort the sheet by type.
- (e) Using the randomize function in the spreadsheet, select 40 data points for movies, and 40 data points for TV shows.
- (f) Conduct a hypothesis test to answer the questions, using the scores for movies and TV shows using a significance level of 0.05.
- (g) Which test would be appropriate in this situation?
- (h) State the null and alternate hypotheses.
- (i) Explain the process and state the test statistic. Sketch the test statistic distribution and label the critical regions.
- (j) What is the conclusion?



FIGURE 6.14: Anime girl from Wikimedia Commons. Photo source: DarkAp89. [👁](#) [📄](#)

8. **Nonparametric Sign test.** Ms. Kim believes that a new on-line math program for her Algebra students is improving their Algebra skills. She told the principal that she expects a median test score of 80 for her class of 16 students. After the test, the scores of students were as follows: 72, 79, 89, 56, 76, 54, 72, 90, 87, 67, 48, 47, 80, 78, 54, 63.

- (a) State the null and alternate hypotheses.
- (b) Assuming a $\alpha = 0.01$ and Binomial distribution with $p = 0.5$ and $n=16$, use the nonparametric sign test to find out if Ms.Kim was

right. Use the stat function of your graphing calculator to enter the data and the n and p values.

- (c) Copy the discrete probability distribution table. State the test statistic and the critical values.
- (d) Will this be a one-tailed or two-tailed test? Why?
- (e) What is the conclusion? Should the null hypothesis be rejected?

6.9 Summative Project

Hypothesis testing with the Palmer Penguin data set

We will use the Palmer Penguin data set as our population data to investigate and practice hypothesis testing. Here is the link to the Google Sheets with the Palmer Penguin data set: [!\[\]\(c3d993ca47bfe2a953c700506ce31fa0_img.jpg\)](#)



FIGURE 6.15: Adelie Penguins on iceberg. Photo from Wikimedia Commons. This file is licensed under the Creative Commons Attribution 2.0 Generic license. Originally posted to Flickr as IMG 0760. [🌐](#)

1. The penguin bill length

- (a) We want to test the statement that the “average bill length of Adelie species is less than 40 mm” using a random sample of 15 Adelie penguins.
 - i. Find the mean and standard deviation for the population of Adelie penguins.
 - ii. Select a random sample of 15 Adelie penguins from the data set.
 - iii. What hypothesis test would be appropriate?

- iv. State the null and alternate hypotheses.
 - v. Find the test statistic, and using a significance level 0.05, find the critical t values (you can use a graphing calculator, or a t chart). Sketch the distribution, and label the critical values and critical region for the hypothesis.
 - vi. What is the conclusion? Write the conclusion statement in terms of the given context.
- (b) We want to test the statement: “The mean bill length for the Gentoo species is different than 50 mm,” using a random sample of 10 penguins from the Penguin data.
- i. Find the mean bill length for the Gentoo population.
 - ii. Select a random sample of 10 Gentoo penguins.
 - iii. We will use a Student t -test to test the significance of the statement. Will this be a one-tailed or two-tailed test? Why?
 - iv. Find the test statistic and, using a significance level 0.05, find the critical values.
 - v. What is the conclusion? Explain.

2. The population of Gentoo species

A group of research students went to the Palmer Islands to collect data about penguins. They sampled 60 penguins and found that of the three different species in the sample, 24 penguins were of the Gentoo species. They hypothesized that in the population of penguins in the entire Palmer Islands, 40% of penguins belong to the Gentoo species.

- (a) Assuming the Palmer Penguin data set is the population, should we accept this hypothesis?

- (b) Conduct an appropriate statistical hypothesis test showing test statistic, sketch the test statistic distribution with critical regions labeled.
- (c) What is the conclusion?

3. **The penguin body-mass**

The students want to know if there is a difference in the body mass of male and female penguins. They conduct a hypothesis test with a random sample of 40 penguins from the Palmer penguin data set.

- (a) State the steps involved in this hypothesis testing.
- (b) How will you select an appropriate test for this situation?
- (c) Sketch the test statistic distribution with critical values. What is the conclusion?

Variable Relationships: the Full Picture

Most real-world problems involve multiple factors and variables. How would sleep and hunger affect an athlete's performance in a soccer game? How does the cost of a concert ticket change with the venue, and fame and popularity of the artist? How do climate and environmental conditions affect crops and harvests? In all these examples, there are multiple variables that interact with each other in ways that affect the outcome. AI has made it possible to analyze hundreds of variables simultaneously, across time and space, which has led to the discovery of new drugs and gene therapies, techniques for real-time hurricane alerts, wild fire management, and optimizing fashion supply chains, among others. In this chapter, we will study the statistics of multiple variables and their interactions, and examples of AI tools and applications that enable studying them.

Studying multiple attributes builds a more complete and richer description of an object, increasing the accuracy of analytics and AI models. For example, the more variables we study, such as bill length, bill depth, flipper length, the better we understand the differences between each of the three species of the Palmer penguins. This chapter provides an introduction to the study of multiple variables. The objective is to provide a brief overview of each type of analysis, to motivate its need and to connect traditional statistical concepts to AI applications.

The χ -square test of independence, ANOVA and linear regression are discussed in some detail. However, the topics of time series, spatial analysis and spatio-temporal analyses are introduced very briefly at a high level, with a greater focus on their application to AI, highlighted through examples. These topics are complex and require extensive foundations, but being familiar with them is essential since they are fundamental to many real-world applications as we will learn in the illustrative examples.

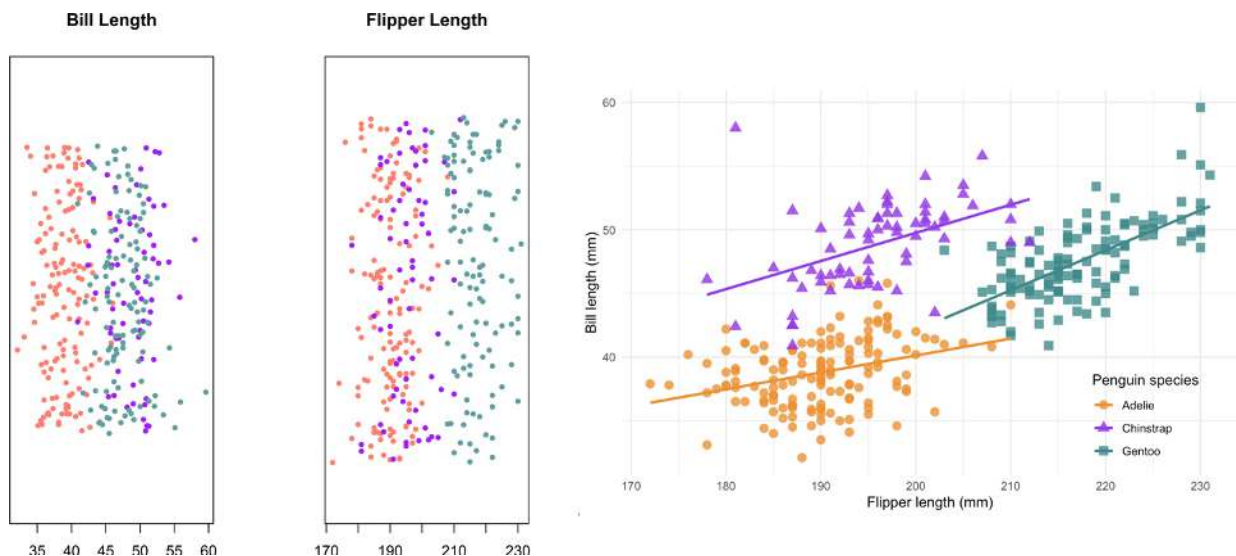
7.1 Illustrative Example: Penguins, the more variables the merrier

Consider [Figure 7.1](#) that depicts three variables from the Palmer penguin data: flipper length, bill length and species. The first two variables are continuous quantitative variables, while the last one is a qualitative, nominal variable. There are several interesting observations:

- In [Figure 7.1](#), the two boxes on the left show the distribution of the three penguin species, Adelie, Chinstrap and Gentoo (different colors), across the values of bill length and flipper length. The figures are called “stripcharts” in *R*, and show how the data values are stacked on the *X*-axis, with a little bit

of jitter to avoid overplotting. The three species have overlap in terms of bill length, particularly Chinstrap and Gentoo which are almost indistinguishable based on bill length alone. Similarly, Adelie and Chinstrap penguins are indistinguishable based on flipper length alone as indicated by their overlap in the second box.

- Now, look at [Figure 7.1](#), right. This plot is taken from Horst *et al.*'s [21] analysis of the Palmer Penguin data, the scatterplot of flipper length versus bill length. Here the three species occupy their own distinct regions in the two-dimensional space of flipper length vs. bill length, with very little overlap.
- While we could not distinguish between the three species based on just the bill length or just the flipper length, when we consider both variables together, the three species are quite distinguishable in the two-dimensional space of flipper length and bill length.



[Extended Description](#)

FIGURE 7.1: Plots of three variables from the Palmer Penguin data taken from Horst *et al.*'s [21] manual. Bill length and flipper length values for the three species are stacked on the X -axis. Some species are very similar and have considerable overlap in range of values, e.g., Chinstrap and Gentoo with respect to bill length, and Adelie and Chinstrap with respect to flipper length. Right, a scatterplot of flipper length vs. bill length from Horst *et al.*'s [21] manual. What happens to the overlap between species when you consider the two variables flipper length and bill length together? [↗](#)

There are many instances where considering two or more variables, and their interaction provides more insight than merely examining each variable by itself. In machine learning and AI, variable interactions are the key to good predictive models.

It is not always possible to visualize like we did in [Figure 7.1](#), or use all the attributes or dimensions of the data in analysis. More variables increase the computational burden as well as cause the notion of

distances in higher-dimensional space to collapse. This is called the *curse of dimensionality*. Explore [GeminiBox 7.1.1](#) to learn more.

GeminiBox 7.1.1: Curse of dimensionality

“What is the curse of dimensionality?” 

In such instances, it is useful to consider projections onto lower dimensions, which is a way of reducing the number of dimensions. Use the prompt in [GeminiBox 7.1.2](#) to learn more. The stripcharts are *one-dimensional projections* since only one variable is represented. The scatterplot is a two-dimensional projection, since it represents two variables. There are other dimensions, such as body mass, that are not represented in the plots.

GeminiBox 7.1.2: Lower-dimensional projections

“What is meant by projection onto a lower dimension?” 

AI has enabled the processing of hundreds of variables resulting in applications that impact almost all aspects of life, from real estate sales to building generative AI models that can write romance novels and create cute puppy images. It has made it possible to develop increasingly precise models, for example, to power autonomous cars. Companies like Zillow predict the sale price of a home based on numerous variables such as square footage, number of bedrooms, location, prices of neighboring homes, condition of the house, and supply in the neighborhood. The variables and predictions are updated in real-time using AI tools. Explore how AI uses multiple variables with [GeminiBox 7.1.3](#).

Data Debate 7.1.1 *Do you think it is important to use multiple variables to determine an athlete's performance at a particular event? If so, what variables? Do you think these would change with time? What if the athlete were participating in an event near the Alps versus near the Sahara Desert? Would that make a difference? Why?*

GeminiBox 7.1.3: AI models and multiple variables

Does AI use multiple variables? What is an example? 

7.1.1 Illustrative example: AI tracks climate change

Changes in climate and pollution have created many challenges: icebergs melt and lead to flooding, uncharacteristic heat causes droughts and food insecurity, and wildlife gets ensnared in plastic waste as shown in [Figure 7.2](#), left. Monitoring and providing early warning to communities challenged by climate changes requires studying a complex array of variables over time, such as temperature trends, ocean flows, changes in vegetation and migration patterns.



FIGURE 7.2: A turtle caught in a plastic bag. AI helps monitor ocean waste by tracking a complex array of constantly changing variables. (Photo credit: Coordinator copy.) A drone deployed in agriculture to monitor crop diseases like citrus canker uses image recognition based on multiple variables. [🔗](#)

Providing advance warning requires speedy and timely processing of all these variables. AI is being used extensively and has been shown to be an effective tool in monitoring constantly changing variables. Read the World Economic Forum story in [Section 7.8](#) on how AI is being used: (i) to pinpoint icebergs that are melting, (ii) to power drones to aid reforestation in Brazil, and (iii) to monitor and minimize waste and emissions.

7.1.2 Illustrative example: AI deployed in drones helps farmers

The United States Department of Agriculture's [National Institute of Food and Agriculture](#) conducts research and educates farmers on how to manage land, livestock and environmental resources to make better decisions. The deployment of drones, shown in [Figure 7.2](#), right, for detecting crop diseases through image recognition and examining soil conditions, and other AI-driven tools make it possible to collect and analyze a constant stream of data of different types of variables across space and time to help farmers optimize land use. Read more in the article *AI powered hydroponics: Revolutionizing sustainable agriculture* in [Section 7.8](#).

The preceding illustrative examples have demonstrated that studying the relationship between multiple variables is crucial for tackling complex problems and can be addressed effectively with AI. Classical statistical methods lay the foundation for such analyses. We will discuss three examples of such multivariate analysis: the χ^2 test for independence of categorical variables, Analysis of Variance (ANOVA) for the effect of categorical variables on continuous variables, and linear relationships between continuous variables.

We will also familiarize ourselves with special types of variables that are measured continuously in time (temporal), across space (spatial), and across both time and space (spatio-temporal). These data types need specialized data structures and computing approaches and have benefited greatly from AI methods.

7.2 Independence of Categorical Variables

In [Chapter 2](#), we studied several types of qualitative categorical variables. Species of penguins, ice cream flavors, different types of sports, and fashion brands are all good examples of qualitative variables. Studying the effect of categorical variables on each other is critical in many applications, e.g., what is the most attractive combination of color and shape of a “like” icon on a website in order to attract most user engagement? Let us start with an example.

7.2.1 AI, fake handbags and the Chi-squared (χ^2) test of independence

The luxe fashion resale company, The RealReal, uses AI to generate product descriptions as well as to screen fakes. In 2019, the company was hit with a class action lawsuit that stated that it was selling knock-off luxury goods. It was settled in 2021. In response, The RealReal deployed AI to pre-screen hundreds of thousands of products including luxury handbags, to filter out a few for human scrutiny. It would have been impossible to manually screen all such products given the volume, without the speed of AI. Read more about this story in the article *The RealReal uses AI to spot fakes* in [Section 7.8](#).

One interesting *hypothetical* test for such a fashion goods company would be to see if there was a dependence between the region where the fake goods originate and the brands. In order to test if there is indeed such a dependence between fake brands and regions where they were uploaded, the company would use the *Chi-squared* (χ^2) test for independence.

ReferenceBox 7.2.1: χ^2 test of independence of categories

Suppose there are two categorical variables X and Y . Let O_{ij} be the number of objects observed in the cell $\{i, j\}$ corresponding to category i of X and j of Y . For illustration, consider the example in [Table 7.1](#), where there are only two categories of X and Y for simplicity.

- Let the row totals be O_{ri} and the column totals be O_{cj} . If the two categories are indeed independent, then using the rule of independence from [Chapter 5](#), we can multiply the marginal probabilities p_i of $X = i$ and q_j of $Y = j$ to get the probability of their co-occurrence, p_{ij} . That is,

$$p_{ij} = p_i * q_j.$$

- We estimate the *observed* probabilities from the sample:

$$\hat{p}_i = P(X = i) = O_{ri}/N$$

and

$$\hat{q}_j = P(Y = j) = O_{cj}/N.$$

We estimate the probabilities of the occurrence of the combination $X = i$ and $Y = j$:

$$\hat{p}_{ij} = \hat{p}_i * \hat{q}_j = O_{ri} * O_{cj}/N^2.$$

- The expected number of observation in cell $\{i, j\}$ using this probability is:

$$n_{ij} = \hat{p}_{ij} * N = \frac{O_{ri} * O_{cj}}{N}.$$

- If the assumption of independence is true, then the deviation of the expected value n_{ij} from the observed value O_{ij} should be small. The departure from the assumption of independence is measured by the sum of relative squared deviations for all the cells:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - n_{ij})^2}{n_{ij}}.$$

- If this computed χ^2 is less than the corresponding critical value in the χ^2 -distribution with

$$(n_X - 1) * (n_Y - 1)$$

degrees of freedom, where n_X is the number of categories of X , and n_Y is the number of categories of Y , then the null hypothesis is supported, and we conclude that the categories of X and Y are independent. If not, it is unlikely that based on this sample, the two categorical variables are independent.

TABLE 7.1: Example table for the χ^2 test of independence where N objects are distributed across the combinations of values of two qualitative variables X and Y . O denotes observed counts for cells (O_{ij}), rows O_{r1}, O_{r2} and columns O_{c1}, O_{c2} . ↴

		Y		
		1	2	Row Total
X	1	O_{11}	O_{12}	O_{r1}
	2	O_{21}	O_{22}	O_{r2}
Column Total		O_{c1}	O_{c2}	N

Let us work through an example to understand. Please review and refer to [ReferenceBox 7.2.1](#).

7.2.2 Step-by-step example of χ^2 test of independence

Consider the example of two categorical variables Brand (with three levels Celine, Prada and Birkin) and Region (with five levels NY, NJ, CT, CA, IL). The null hypothesis H_N would be: “There is no dependence between the region and the brands”. The alternate hypotheses H_A would be: “There is a dependence between fake brands and regions where they are uploaded.” Suppose the company collected the data shown in [Figure 7.3](#), right, where the data represents the count of fakes in each brand-region

combination. The number of regions (categories of X) is $n_X=4$, and the number of brands (categories of Y) $n_Y = 3$. Consider the following:

1. The marginal probabilities for the brands are computed by taking the column totals and dividing them by the grand total of 1000. Therefore:

$$P(Celine) = 180/1000 = 0.18$$

$$P(Prada) = 180/1000 = 0.18$$

$$P(Birkin) = 640/1000 = 0.64$$



Region	Brand			Row Total
	Celine	Prada	Birkin	
NY	100	10	10	120
NJ	20	20	500	540
CT	30	10	100	140
CA	30	40	30	100
IL	0	100	0	100
Column Total	180	180	640	1000

[Extended Description](#)

FIGURE 7.3: Left: A U.S. customs officer examines fake handbags (Photo: U.S. Customs and Border Protection). Are there more fakes in certain regions of the US? Right: Hypothetical data about fake handbags and regions. [📄](#)

2. Similarly, the marginal probabilities of the regions are given by:

$$P(NY) = 120/1000 = 0.12$$

$$P(NJ) = 540/1000 = 0.54$$

$$P(CT) = 140/1000 = 0.14$$

$$P(CA) = 100/1000 = 0.1$$

$$P(IL) = 100/1000 = 0.1$$

3. Under the null hypothesis N_H of independence, the joint probability p_{ij} of any region and brand will simply be the product of their marginal probabilities using the multiplication rule of probabilities of independent events. That is,

$$p_{ij} = p_i * q_j,$$

(7.1)

where p_{ij} is the joint probability, p_i is the marginal (row) probability of the i^{th} region, and q_j is the marginal (column) probability of the j^{th} brand.

Therefore, in any cell ij of the table we would expect n_{ij} fakes, where:

$$n_{ij} = p_{ij} * N$$

where N is the grand total of 1000. For example, for the region NJ ($i=2$) and brand “Birkin” ($j=3$),

$$p_{23} = p_2 * q_3 = 540/1000 * 640/1000 = 0.3456,$$

and the expected number of Birkin fakes in NJ are:

$$n_{23} = p_{23} * N = 0.3456 * 1000 = 345.6.$$

4. If the differences between the observed cell counts O_{ij} and expected cell counts n_{ij} are small, then the null hypothesis is true; otherwise not. The test statistic $T(X)$ that quantifies the deviation of the observed from expected is given by:

$$T(X) = \sum_{i=1, n_X; j=1, n_Y} \frac{(O_{ij} - n_{ij})^2}{n_{ij}},$$

where O_{ij} are the observed cell counts, e.g., as shown in [Figure 7.3](#), left, and n_{ij} are the expected cell counts using the assumption of independence to compute the join probabilities as shown in Equation 7.1.

The test statistic's sampling distribution is a well-known continuous distribution called the χ^2 with $(l - 1) * (k - 1)$ degrees of freedom, where l = no. of categories of region (rows), and k number of categories of brand (columns).

5. The value of $T(X)$ in this case is $\chi^2 = 1060.3$, $df = (n_X - 1) * (n_Y - 1) = 4 * 2 = 8$. We look up the χ^2 tables with 8 d.f. and find that the corresponding p -value is less than $2.2e-16$, which is very small, much less than a level of significance of $\alpha = 0.05$.
6. We conclude that the null hypothesis is rejected and that there is a significant dependence between region and the brand of fakes that originate there.

Data Debate 7.2.1 *Given the above conclusion, what would your advice to The RealReal be in this hypothetical situation?*

GeminiBox 7.2.1: χ^2 test for independence for categorical variables [↗](#)

“What are the underlying assumptions of the χ^2 test of independence?” 

Explore [GeminiBox 7.2.1](#) to learn more about the χ^2 test such as its assumptions: independent and randomly chosen data, tabulated to reflect counts or frequencies of categorical data, not proportions or percentages.

Dependence must not be mistaken for causation; it just means that there is a propensity of certain values of the two categorical variables to co-occur. Furthermore, if the cell counts are too small, say smaller than 5, then other tests, such as Fisher's exact test should be considered. As always, human supervision and vetting is crucial, which is why The RealReal has a set of experts to validate the goods screened by AI.

7.3 Analysis of Variance for Continuous and Categorical Variables

Suppose that in addition to counts or frequencies of categorical variables, we have multiple measurements of a continuous variable within each combination of the levels of the categorical variables. Do diseases respond to different drug treatments? Are they more effective for men or women? Do different fruit respond differently to natural and artificial light?

Horticulturists at the USDA study hydroponic produce, [Figure 7.4](#), left, to identify the best varieties that respond to hydroponic agriculture. Agritech uses AI to monitor light conditions, humidity, mineral levels, plant growth indicators and other variables and analyze data from hundreds of sensors. According to the article *AI powered hydroponics: Revolutionizing sustainable agriculture* in [Section 7.8](#), AI makes sustainable agriculture possible. With increasing pressure on earth's soil resources, there is great interest in the use of AI for alternative farming methods, e.g., sustainable soil-less agriculture and vertical farming.

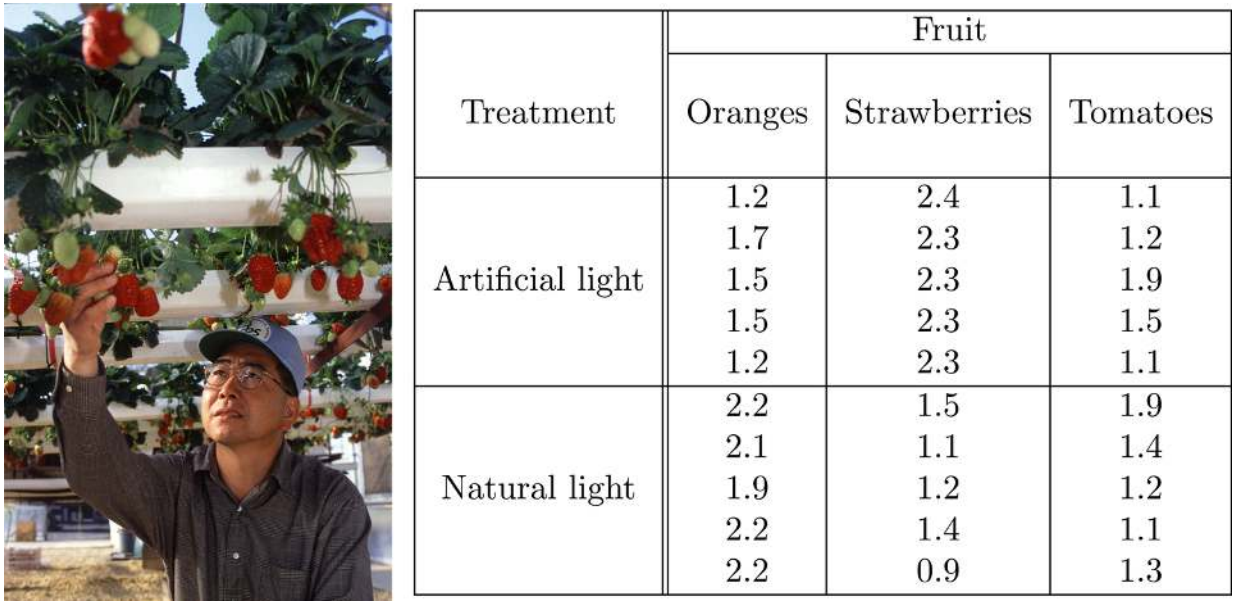


FIGURE 7.4: Left: A USDA horticulturist examines hydroponic strawberries. What type of fruit would thrive under hydroponic agriculture? Right: Hypothetical data table to test the variation in weight (continuous variable measured in pounds) of fruit due to the interaction between two categorical variables (Light treatment and Fruit type). [↗](#)

7.3.1 Hydroponic agritech AI and ANOVA

Consider the second hypothetical example in [Figure 7.4](#). It contains five measurements of weight of fruit, called *replications* or *repeated measurements* for each combination of Light Treatment type (artificial, natural) and Fruit type (oranges, strawberries, tomatoes). The number of measurements in each cell corresponding to row i and column j is n_{ij} . They need not be equal, but very often they are, in a planned *balanced* experiment. Designing experiments to yield accurate and unbiased results is an important area of study in statistics. *The Theory of Design of Experiments* by Cox and Reid [10] is a classic that covers the proper design of experiments to ensure accurate results. Check [GeminiBox 7.3.1](#) for a quick overview of experimental design.

GeminiBox 7.3.1: Balanced experimental design [↗](#)

“What are balanced experiments in statistics?” 

In [Figure 7.4](#), $n_{ij} = 5$ for all cells. We are interested in testing whether the weights of the fruit are different across different categories of Light treatment (rows) and Fruit type (columns), and across different combinations of light treatment and fruit type (cells).

GeminiBox 7.3.2: AI and vertical farming

“How is AI used in hydroponic cultivation?” 

ANOVA stands for Analysis of Variance and refers to testing whether typical values of a continuous variable (e.g., the mean) are the same across different categories of a qualitative variable, or combinations of categories of multiple qualitative variables. It is a generalization of the t -test from two groups to multiple groups. In the example of hydroponic cultivation, the two categorical variables are light treatment with two possible values, natural and artificial, and fruit type, with three possible values, oranges, strawberries and tomatoes. ANOVA helps test whether the typical values of the continuous variable, average weight of the fruit, are the same across all the categories. We describe an example with two categorical variables, but it is applicable to combinations of more variables as well. Agresti's *Categorical Data Analysis* [1] provides a comprehensive and rigorous coverage of ANOVA, logistic regression and models of other binary responses.

One-way ANOVA is a method of testing the typical values of a continuous variable with respect to a single qualitative variable, e.g., fruit type. Two-way ANOVA is a way of testing whether the average value of the continuous variable is the same across combinations of two qualitative variables, e.g., fruit type and light treatment. We present the test statistic for one-way ANOVA here, while detailed explanations and two-way ANOVA test statistic are provided in [Section 7.5](#).

ReferenceBox 7.3.1: F-statistic for ANOVA [↗](#)

Let:

- n_r and n_c be the number of rows (e.g., light treatments) and columns (e.g., fruit types) respectively, n_d be the number of observations in each cell, assumed to be the same for all cells in this case for simplicity. In our example, $n_r = 2, n_c = 3, n_d = 5$. Let N be the total number of observations.
- x_{ijk} be the k^{th} observation in row i and column j of the data in [Figure 7.4](#), right. For example, $x_{213} = 1.9$. In this example, there are $n_k = 5$ observations in each cell.
- $\bar{X}$ be the overall mean of fruit weights, irrespective of fruit type or light treatment.
- R_i be the mean of fruit weights of row i , e.g., artificial light when $i = 1$.
- C_j be the mean of fruit weights of column j , e.g., strawberries when $j = 2$.

- $$SSR = \sum_i n_i (R_i - \bar{X})^2, i = 1, \dots, n_r,$$

be the *sum of squares of deviations of row means* from the overall mean, reflecting the differences across rows, e.g., light treatment in this case.

- $$SSC = \sum_j n_j (C_j - \bar{X})^2, j = 1, \dots, n_c,$$

be the *sum of squares of deviations of column means* from the overall mean, reflecting the differences across columns, e.g., fruit types in this case.

- $$SST = \sum_{i,j} (x_{ij} - \bar{X})^2, i = 1, \dots, n_r; j = 1, \dots, n_c,$$

be the *total sum of squares of deviation of individual measurements from the overall mean*.

- $$SSE = SST - SSR - SSC,$$

be the *residual sum of squares* after the row and column means have been accounted for.

Then, the one-way ANOVA test-statistic for difference in the quantitative variable (fruit weight) across the groups in rows (light treatment) is given by:

$$F_R = \frac{SSR/(n_r - 1)}{SSE/(N - n_r)},$$

and F_R has a sampling distribution of an F -distribution with $(n_r - 1, N - n_r)$ degrees of freedom which will be used to determine the critical values for testing the null hypothesis.

The one-way ANOVA test-statistic for difference in the quantitative variable (fruit weight) across the groups in columns (fruit type) is given by:

$$F_C = \frac{SSC/(n_c - 1)}{SSE/(N - n_c)},$$

and F_C has a sampling distribution of an F -distribution with $(n_c - 1, N - n_c)$ degrees of freedom which will be used to determine the critical values for testing the null hypothesis.

Please see [Section 7.5](#) for detailed explanation and the test statistic for two-way ANOVA.

7.4 ANOVA with Excel

The manual computation of ANOVA test statistics is quite tedious and time-consuming. However, we have provided a fully worked out step-by-step manual computation of a two-way ANOVA for the Fruit type vs. Light treatment data in [Figure 7.4](#), right, in [Section 7.5](#). Luckily, programming languages such as *R* and Python, and software tools including Excel and Google Sheets provide easy ways of automatically computing ANOVA statistics. We provide examples below.

7.4.1 One-way ANOVA with Excel: Palmer penguins

One-way ANOVA test is used for testing whether the typical values of a continuous variable are the same for different values of a single categorical variable. It is a generalization of the t -test in [Chapter 6](#) to more than two groups. Consider the example of Palmer penguins: suppose that we want to examine if there is a difference in the mean flipper length (continuous variable) across the three species (qualitative variable) of penguins: Adelie, Chinstrap and Gentoo.

For this example, we select a random sample of 10 penguins of each species from the data set and organize the flipper length measurements data in a table as shown in [Figure 7.5](#). The following assumptions are required for the one-way ANOVA test to be valid:

- Level of measurement: Independent variable should be qualitative nominal (e.g., penguin species); the dependent variable should be quantitative (e.g., flipper-length in mm).
- Independence of groups: Measurements of each group should be independent, i.e., the measured values of one group must not be influenced by the measures of the other groups.
- Normality: Continuous variable within each group is normally distributed especially for small sample sizes.
- Homogeneity: The variance within each group must be roughly the same.

Adelie	Chinstrap	Gentoo
186.00	197.00	216.00
186.00	197.00	210.00
195.00	203.00	222.00
189.00	195.00	210.00
185.00	197.00	218.00
190.00	196.00	210.00
186.00	181.00	217.00
197.00	194.00	230.00
185.00	202.00	209.00
187.00	205.00	209.00

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3688.0667	2	1844.0333	50.2207989	7.97461E-10	3.3541308
Within Groups	991.4	27	36.718519			
Total	4679.4667	29				

[Extended Description](#)

FIGURE 7.5: One-way ANOVA with Excel: Input data. Dialog box selections. Output of one-way ANOVA. [↗](#)

Formulate the null and alternate hypotheses and specify the level of significance α :

1. Null hypothesis: “There is no difference in the mean flipper length across the three species.”
2. Alternate hypothesis: The mean flipper lengths of the three species are statistically different.
3. Level of significance: $\alpha=0.05$.

Follow the following steps to use Excel Data Analysis Toolkit to perform the one-way ANOVA test:

- First, organize the data in three columns with species names as headings, as shown in [Figure 7.5](#), left.
- Next, open the Excel Data Analysis Toolkit and choose One-factor ANOVA. A dialog screen appears as shown in [Figure 7.5](#), middle. Select the cells and input the data exactly as shown in [Figure 7.5](#), left.
- Select the output range in the dialog box, a single cell where you want the output tables to be displayed. Any empty cell in the sheet will do.
- The output is ready in [Figure 7.5](#), right! Examine and interpret the output and declare a conclusion.

Interpretation of test output: The output tables from Excel are displayed in [Figure 7.5](#), right.

The row *Between groups* contains the value of the test statistic $F_R = 50.2208$ corresponding to the F_R -statistic in [ReferenceBox 7.3.1](#), the ratio of sums of squares.

The critical value of the F -distribution with $n_r - 1 = 3 - 1$ d.f., (since there are $n_r = 3$ species) is given by $F = 3.3541$. This value is much smaller than the critical value $F_R = 50.2208$.

The p -value $= 0.00000$, which is less than the specified $\alpha = 0.05$.

Therefore, the null hypothesis is not valid, and we conclude that there is a significant difference in the mean flipper length of the three species.

7.4.2 Two-way ANOVA with Excel: Hydroponic fruit

Let us discuss two-way ANOVA using the example of hydroponic fruit. The two qualitative variables or factors are light treatment (rows) and type of fruit (columns). There are three null hypotheses:

- H1: The mean weights of fruit grouped by light exposure (rows) are the same.
- H2: The mean weights of the type of fruits (columns) are the same.
- H3: The interaction between the two factors (light treatment and fruit type) has an effect on the mean weight of fruit.

The level of significance is set to $\alpha = 0.05$. Open the Excel Data Analysis Toolkit. From the list, select the test ANOVA: two-factor with replication. Select the entire table as input range, set the number of rows =5, and enter the significance level alpha as 0.05, as shown in [Figure 7.6](#), left. The output range can be any cell on the sheet. The output is shown in [Figure 7.6](#), right. The intermediate tables containing the results are shown in [Section 7.5](#).

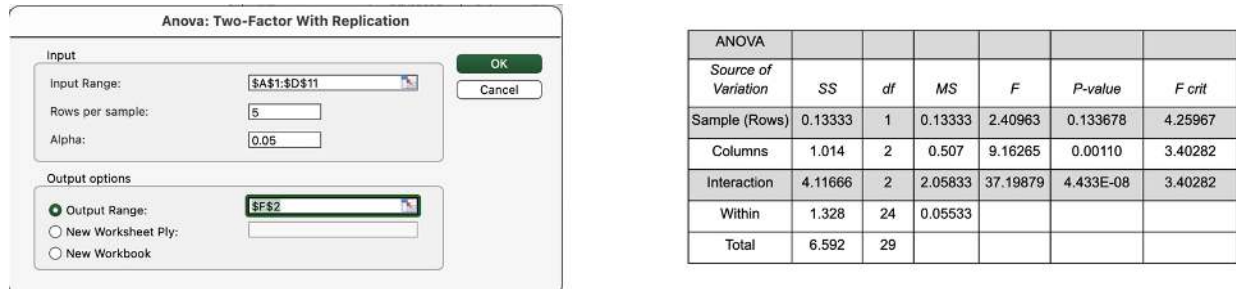


FIGURE 7.6: Left: Excel Data Analysis Toolkit entries for two-way ANOVA. Right: Output table with F statistics for null hypotheses H_1 , H_2 and H_3 . [📄](#)

Interpretation of results in [Figure 7.6](#), right, row-by-row are as follows:

- **Sample (Rows):** Upon examining the F -test statistic and the appropriate critical value, $F - crit$, we see that $F = 2.4 < F - crit = 4.26$. In addition, $p\text{-value} = 0.1336 > 0.05$, the level of significance α . We accept the null hypothesis H_1 - the mean weights of the fruits under the two light treatments are the same.
- **Columns:** Upon examination of the F , $F - crit$ and p -values, we see that $F = 9.16 > F - crit = 3.4$ and $p\text{-value} 0.00110 < 0.05$. We reject the null hypothesis H_2 . The mean weights of the three varieties of fruits are not the same.
- **Interaction:** Upon examining the F , $F - crit$ and p -values in this row, we see that $F = 37.1987 > F - crit = 3.4028$ and $p\text{-value} 0.0000 < 0.05$. We reject the null hypothesis H_3 . There is interaction between the two variables, light treatment and fruit type.

We conclude that the variety of fruit and the interaction between light treatment and fruit type influence the weight of fruit.

7.5 Manual Computation of a Two-Way ANOVA

In this section we provide a step-by-step computation of two-way ANOVA for the data in [Figure 7.4](#), right. The notation is a little complex. Please work through this example to truly understand the concepts and logic behind ANOVA.

7.5.1 ANOVA: Expected values

First, we compute the *expected values* of the continuous variable for each group of interest:

1. Row means, i.e., marginal mean of the continuous variable values (fruit weights) within each level of categorical variable 1, e.g., light type in [Figure 7.4](#), right. There will be n_r such means, R_i , $i = 1, \dots, n_r$, where n_r is the number of rows, corresponding to the number of levels of categorical variable 1. In this example $n_r=2$, corresponding to the two light treatments, natural and artificial. ($R_1 = 1.5733, R_2 = 1.7$.)
2. Column means, i.e., marginal mean of the continuous variable values (fruit weights) within each level of categorical variable 2, e.g., fruit type in [Figure 7.4](#), right. There will be n_c means, C_j , $j = 1, \dots, n_c$ where n_c is the number of columns, the number of levels of the categorical variable 2. In this example $n_c=3$, corresponding to the three fruit types: orange, strawberry, and tomato. ($C_1 = 1.77, C_2 = 1.77, C_3 = 1.37$.)
3. Cell means, i.e., mean of the continuous variable values (fruit weights) within each cell, corresponding to a combination of categorical variable 1 and categorical variable 2, e.g., fruit type and light treatment combination. There will be $n_{rc}=n_r * n_c$ such means, the number of combinations of rows and cells, I_{ij} , $i, j = 1, \dots, n_r; j = 1, \dots, n_c$. These represent the *interaction* between the two categorical variables. In this example, it represents the number of light treatment and fruit type combinations given by $n_{rc} = n_r * n_c = 2 * 3 = 6$. ($I_{11} = 1.22, I_{12} = 2.12, I_{13} = 1.38, I_{21} = 2.31, I_{22} = 1.42, I_{23} = 1.36$.)
4. The grand mean, which is just the overall mean of all values of the continuous variables, e.g., fruit weight.

$$\bar{X} = \sum_{i,j,k} x_{ijk} / N,$$

where $i = 1, \dots, n_r, j = 1, \dots, n_c, k = 1, \dots, n_{ij}$ and N is the total number of measurements. In this example, $N=30$. ($\bar{X} = 1.63667$.)

7.5.2 ANOVA: Sum of squares of deviations

Next, we test how different these means are from the overall mean by computing the *sum of squares of deviations* from expected values. The larger the deviation from the overall expected value, the greater the variability and difference between the groups in terms of the continuous variable of interest.

- Row sum of squares captures the deviations of the row means from the overall mean.

$$SSR = \sum_i n_i (R_i - \bar{X})^2, i = 1, \dots, n_r.$$

It reflects how different each level of categorical variable 1 is from the overall mean. In [Figure 7.4](#), right, $n_i = 15$ for all rows, corresponding to each of the two light treatments. Furthermore, $SSR = 15 * (1.573333 - 1.636667)^2 + 15 * (1.7 - 1.636667)^2 = 0.120333$.

- Column sum of squares captures the deviations of the row means from the overall mean.

$$SSC = \sum_j n_j (C_j - \bar{X})^2, j = 1, \dots, n_c.$$

It reflects how different each level of categorical variable 2 is from the overall mean. In our example, $n_j = 10$ for all columns, each of which corresponds to a fruit type. In our example, $SSC = 10 * (1.77 - 1.636667)^2 + 10 * (1.77 - 1.636667)^2 + 10 * (1.37 - 1.636667)^2 = 1.066667$.

- Interaction sum of squares captures the deviations of the individual cell means from the overall means, after subtracting the effect of the categorical variable 1 alone (corresponding row mean), and categorical variable 2 alone (corresponding column mean.)

$$SSI = \sum_{i,j} n_{ij} (I_{i,j} - R_i - C_j + \bar{X})^2, i = 1, \dots, n_r; j = 1, \dots, n_c.$$

It reflects how far the cell means are from the overall mean after accounting for the differences in categorical variable 1 levels (light treatments) by subtracting the corresponding row mean R_i , and of categorical variable 2 (fruit type) by subtracting the corresponding column mean C_j . In our example, $n_{ij} = 5$ for all cells, which correspond to combination light treatment and fruit type, and $SSI = 4.130667$.

- Total sum of squares

$$SST = \sum_{i,j} (x_{ij} - \bar{X})^2, i = 1, \dots, n_r; j = 1, \dots, n_c.$$

It reflects the sum of squares of deviation of individual measurements from the overall mean. In our case, there are 30 measurements and $SST = 6.669667$.

- The error sum of squares is

$$SSE = SST - SSR - SSC - SSI,$$

what remains after all the factors that affect the mean have been accounted for. In our example, $SSE = 1.352$. This is also called *residual sum of squares*.

7.5.3 ANOVA: Sampling distribution

Each row, column and interaction effect can be tested for significance using statistics based on the sum of squares and their sampling distributions. The sampling statistics $T(X)$ and their sampling distributions, the F distributions with appropriate degrees of freedom (d.f.), are given by:

$$T(X)_R = (SSR/n_r - 1)/(SSE/(n_r - 1)(n_c - 1)); F_r = F_{n_r-1, d.f.}$$

(7.2)

$$T(X)_C = (SSC/n_c - 1)/(SSE/(n_r - 1)(n_c - 1)); F_c = F_{n_c-1, d. f.} \quad (7.3)$$

$$T(X)_{RC} = (SSI/n_{rc} - 1)/(SSE/(n_r - 1)(n_c - 1)); F_{rc} = F_{(n_r-1)*(n_c-1), d. f.} \quad (7.4)$$

The p -values of the F -statistic can be obtained from F -distribution tables or using software tools. We used R to do the computations and check for significance. For the data in [Figure 7.4](#), right:

$$\begin{aligned} F_r &= 2.136 & n_r - 1 &= 2 - 1 = 1d. f. & p_r &= 0.156838 \\ F_c &= 9.467 & n_c - 1 &= 3 - 1 = 2d. f. & p_c &= 0.000931 \\ F_{rc} &= 36.663 & (n_r - 1) * (n_c - 1) &= 2d. f. & p_{rc} &\approx 0 \end{aligned}$$

The first F_r based on the ratio of SSR and SSE is relatively small, with a large corresponding p -value, p_r , indicating deviations that are not statistically significant. The values F_c and F_{rc} corresponding to column effects and interaction effects are large, indicating that the sum of squares are large compared to the error sum of squares. This is reflected in small p -values, p_c and p_{rc} , which are significantly smaller than the margin of error $\alpha = 0.05$, indicating that the null hypothesis of “no difference” might not be appropriate. That is, column and interaction effects are significantly larger than those permitted by chance.

7.5.4 ANOVA: Conclusion

While the light treatment (F_r , row effect) by itself has no effect on the weight of the fruit, as reflected by small F_r and large p -value, the fruit type (F_c , column effect) and the interaction between the fruit type and light treatment (F_{rc} , interaction effect) have a significant effect on the weight of the fruit as demonstrated by large F values and correspondingly small p values. Intermediate output tables corresponding to effect on columns (fruit type) of each row (light treatment), as well as independent of the light treatment, created by Excel Data Analysis Toolkit are shown in [Figure 7.7](#).

SUMMARY	Oranges	Strawberries	Tomatoes	Total
Artificial Light				
Count	5	5	5	15
Sum	7.1	11.6	6.9	25.6
Average	1.42	2.32	1.38	1.70666
Variance	0.047	0.002	0.112	0.24780

Natural Light	Oranges	Strawberries	Tomatoes	Total
Count	5	5	5	15
Sum	10.6	6.1	6.9	23.6
Average	2.12	1.22	1.38	1.57333
Variance	0.017	0.057	0.097	0.21352

Total	Oranges	Strawberries	Tomatoes
Count	10	10	10
Sum	17.7	17.7	13.8
Average	1.77	1.77	1.38
Variance	0.16455	0.36233	0.09288

FIGURE 7.7: Left: Intermediate output tables from two-way ANOVA with Excel. Left: Effect of artificial light treatment on the three varieties of fruit. The total mean 1.7067 is used in the calculation of the F value for ANOVA. Center: Effect of natural light on the three varieties of fruit. The total mean 1.5733 is used in the calculation of the F value for ANOVA. Right: Effect of fruit variety on fruit weight.



7.6 Relationships Between Continuous Variables

Suppose you bought a new car. You drive it around, sparingly at first, but then your new job requires a commute, and you drive it more and more. What happens to the mileage on the car as you drive more? What happens to the fuel consumption? What happens to the resale value of the car? You know instinctively that mileage increases with distance, as does fuel consumption, but the resale price of the car decreases with more miles.

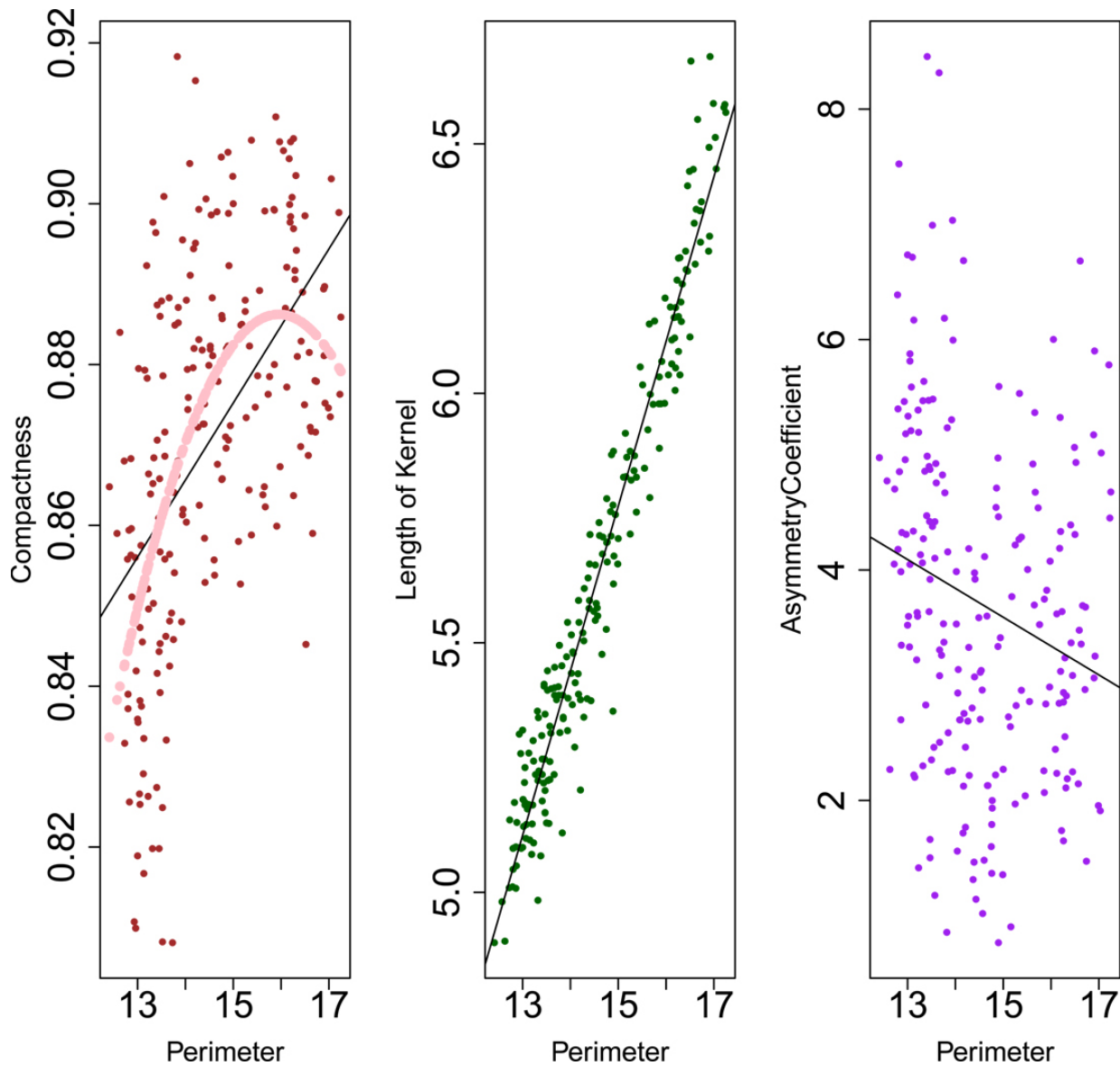
The relationship between the distance you drive and mileage is deterministic, an exact reflection of the number of miles you drive. However, fuel consumption depends on whether you are driving on highways, on country roads, or in heavy traffic, and whether it is an EV, hybrid or conventional car. In general, the more you drive, the more fuel you need. Similarly, the resale value of the car decreases with the number of miles driven but also depends on the make of the car, the level of maintenance, and whether the car has been in accidents.

In this section, we will study relationships between continuous variables like resale price and miles driven. Some variable relationships are simple and intuitive, e.g., a change in one variable causes a proportionate but constant amount of change in the other variable. This is a linear relationship that can be represented using a straight line. The relationship can also be more complex, e.g., every increment in time spent studying can cause an exponential improvement in grades.

In this chapter, we will focus only on linear relationships. More complex relationships like quadratic and exponential, or those that cannot be captured by compact mathematical expressions, require advanced statistical concepts, machine learning and artificial intelligence methods like deep neural nets.

7.6.1 Linear relationship: Correlation

Consider [Figure 7.8](#). The data is courtesy of UCI Machine Learning Data Repository and pertains to characteristics of wheat seeds. It can be found at [🔗](#). We have created three scatterplots (using *R*) of the variable Perimeter with respect to three other characteristics: Compactness, Length of Kernel and Asymmetry Coefficient. The plot also includes a straight line that approximately “fits” the data, more on that in a moment. The first panel shows the scatterplot of Perimeters vs. Compactness. The data points are scattered farther away from the line, indicating that a straight line might not be a good fit. A curve (shown in pink) might be a better representation. The second panel shows the relationship between Perimeter and Length of Kernel. The data points are clustered closely around the line indicating that it might be a good approximation of the data. The third panel shows the scatterplot of Asymmetry Coefficient vs. Perimeter. The data points are so scattered that the line does a poor job representing the data points. There is probably no structure to the data.



[Extended Description](#)

FIGURE 7.8: Scatterplots and lines of best fit for wheat kernels based on the open source UCI Machine Learning Repository data. The figures were created using *R*. $\rho=0.5292436$, 0.9724223 , -0.2173404 , respectively, for the data in the three panels. Is that consistent with your understanding of what ρ means?



A *correlation coefficient* quantifies the strength of linear relationship between two continuous variables. Suppose that the mileage on a car is represented by X miles and its resale value by Y dollars. X and Y are random variables since they vary from car to car, and if measured at different points in time. If we measure the ordered pair values (x_i, y_i) for $i = 1, \dots, n$, cars independently of each other, then the correlation coefficient between miles and resale values is given by:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}},$$

where $\bar{x}, \bar{y}$ are the sample means of mileage (x) and resale price (y), respectively.

The value of a correlation coefficient can be proved to be between -1 and 1; see [35] for proof. A value of +1 means perfect positive linear relationship, that is, for every increase of one unit in X , Y increases by β where β is a constant. Similarly, a value of -1 implies a perfect negative linear relationship. A value of ρ close to 0 implies the lack of a linear relationship, i.e., when X increases, Y may increase, decrease or stay the same. The absence of a linear relationship does not exclude other types of relationships, for example quadratic or exponential.

Data Debate 7.6.1 In [Figure 7.8](#), which pair of variables has the best linear relationship? Why?

The correlation coefficient ρ between Perimeter and Compactness, Length of Kernel and Asymmetry Coefficient is respectively, 0.5292436, 0.9724223, -0.2173404. Are the correlation values consistent with the data shown in the three scatterplots in [Figure 7.8](#)?

Data Debate 7.6.2 What do the three values of ρ in [Figure 7.8](#) mean?

7.6.2 Line of best fit: Ordinary Least Square regression

We are often interested in *predicting* the value of Y for a given value of X . For example, the resale value of a car based on its mileage, or length of a wheat seed kernel based on its perimeter. Y is called the *dependent variable*, and X is called the *independent variable*. Compact mathematical functions, e.g., that of a straight line, are often used to express the relationship between the two variables. Consider the line:

$$y = \alpha + \beta * x, \tag{7.5}$$

where α represents the intercept on the Y -axis, and β represents the slope of the line.

There could potentially be many lines that pass through the data, corresponding to different values of α and β . If you think of a straight line twirling through the data with any given point as a fulcrum, then the line that is closest to most points is considered the line of best fit.

Note that the points in the second panel of [Figure 7.8](#), center, are close to the line but not exactly on the line. That's because variables such as resale value of a car, dimensions of a wheat kernel are random variables that vary from car to car, or seed to seed, and are subject to variability caused by chance including measurement errors. The variability due to chance is incorporated into the line of best fit by adding an error component, ϵ :

$$y = \alpha + \beta * x + \epsilon \tag{7.6}$$

The regression line represents the line of best fit. If you draw a line from x_i value on the X -axis parallel to the Y -axis, the point $\hat{y}_i$ where it intersects the regression line is the predicted value of the dependent variable Y for the given x_i value. That is:

$$\hat{y}_i = \alpha + \beta * x_i$$

The difference between the predicted values and the observed data value is called a *residual* denoted by:

$$r_i = y_i - \hat{y}_i.$$

If the line of best fit approximates the relationship between Y and X well, then the residuals are expected to be small. The line of best fit minimizes

$$\sum_i r_i^2.$$

This method of finding the line of best fit is called *Ordinary Least Squares (OLS)* because it minimizes the sum of squares of the residuals,

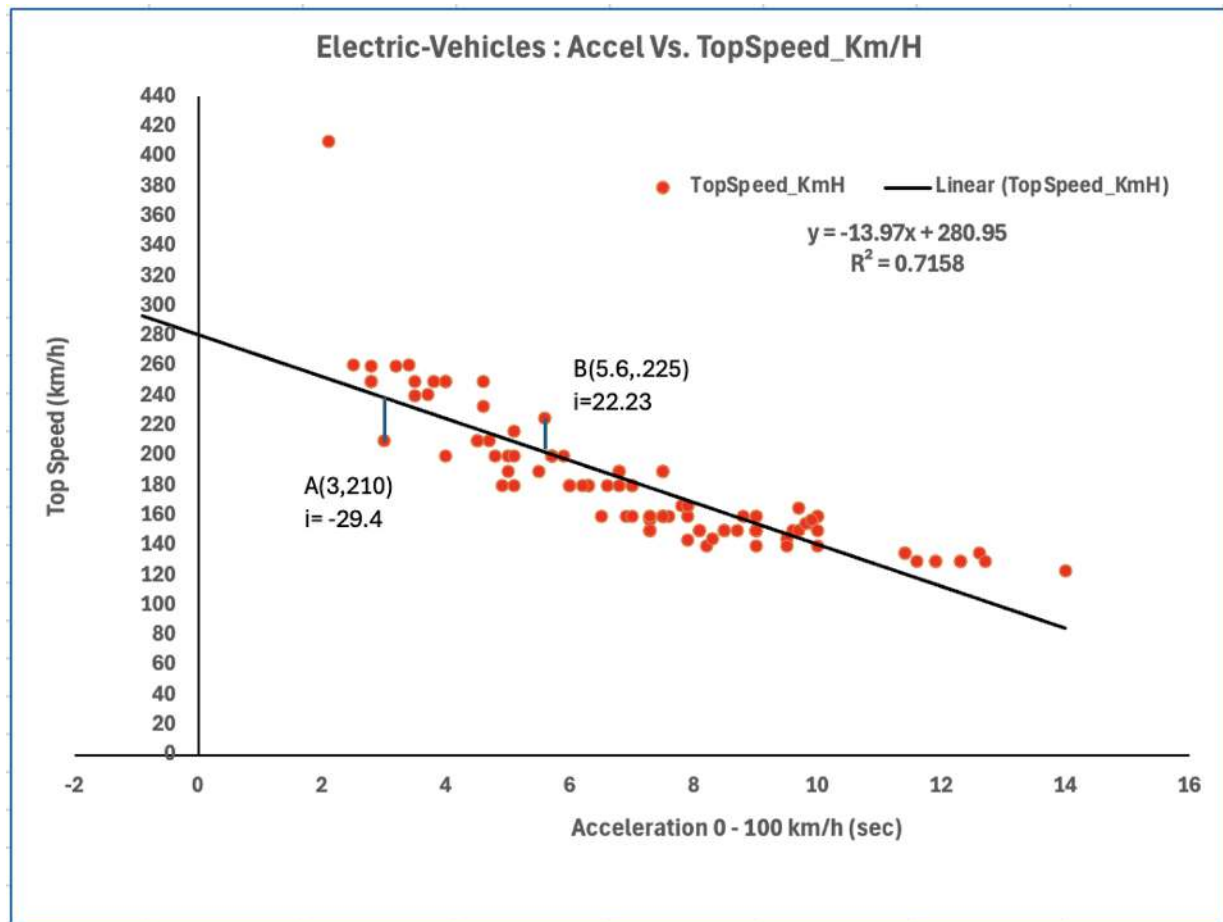
$$r_i = y_i - \hat{y}_i.$$

The line in Equation 7.6 is called the *regression line*. The book *Introduction to Linear Regression Analysis* by Montgomery *et al.* [31] provides a comprehensive discussion of the topic.

[Figure 7.9](#), left, depicts a regression line for the relationship between acceleration (time in seconds from 0 to 100 km/h) and the top speed of Electric Vehicles (EV). It was created with Excel using the publicly available EV data set on kaggle [🔗](#). Each individual observation (x_i, y_i) represents the ordered pair of acceleration and top speed measurements, and it is depicted as a dot on the plot. We observe that:

$$\alpha = 280.95$$

$$\beta = -13.97.$$



[Extended Description](#)

FIGURE 7.9: This chart shows the line of regression for Specs of Electric Vehicles: Acceleration time from 0-100 km/h vs. Top Speed in km/h. EV data set from [kaggle.com](https://www.kaggle.com). Notice the residuals marked on the figure for two data points A and B. [↗](#)

The slope β is negative. How would you interpret this? Hint: You can either have quick acceleration or have a high top speed; EVs do not have the capacity to sustain both. Check [GeminiBox 7.6.1](#).

GeminiBox 7.6.1: For car buffs [↗](#)

“Is there a trade-off between top speed and acceleration in EVs?”

[Figure 7.9](#), left, shows the residual for two points. For point A whose coordinates are $(x = 3, y = 210)$, the predicted value is:

$$\hat{y} = \alpha + \beta * x = 280.95 + 13.97 * (3) = 239.04.$$

Therefore, the residual for point A is:

$$y - \hat{y} = 210 - 239.04 = -29.4.$$

Similarly, the residual for point B is 22.23. The greater the residual, the farther a point is from the regression line. Points that have abnormally large residuals are outliers with respect to the regression line. *Anomaly detection* involves finding outliers with respect to models that fit the data well.

Data Debate 7.6.3 *Are the signs of the residuals reasonable? Why?*

7.6.3 Relationship between correlation (ρ) and slope (β)

The slope of the linear regression line is related to the correlation coefficient in the following way.

$$\beta = \rho \frac{s_y}{s_x},$$

where s_y, s_x are the sample standard deviations of x and y . The greater the correlation coefficient, the stronger the relationship between y and x , and the steeper the line of best fit. For every unit change in x , y will change by β units.

The purpose of using a mathematical model like the regression line in Equation 7.6 is to express the relationship between x and y in a compact manner. We could represent hundreds of thousands of wheat kernel measurements with the regression line, and predict the kernel length based on the perimeter using that compact equation and be assured that the margin of error will be within the scope of a Standard Normal distribution.

There are other ways of determining the line of best fit, e.g., orthogonal regression where the sum of the squares of the perpendicular distances to the regression line is minimized, taking into account errors in both x and y values. In contrast, simple linear regression takes into account just the errors in y values. However, simple linear regression as described above remains the most popular and intuitive.

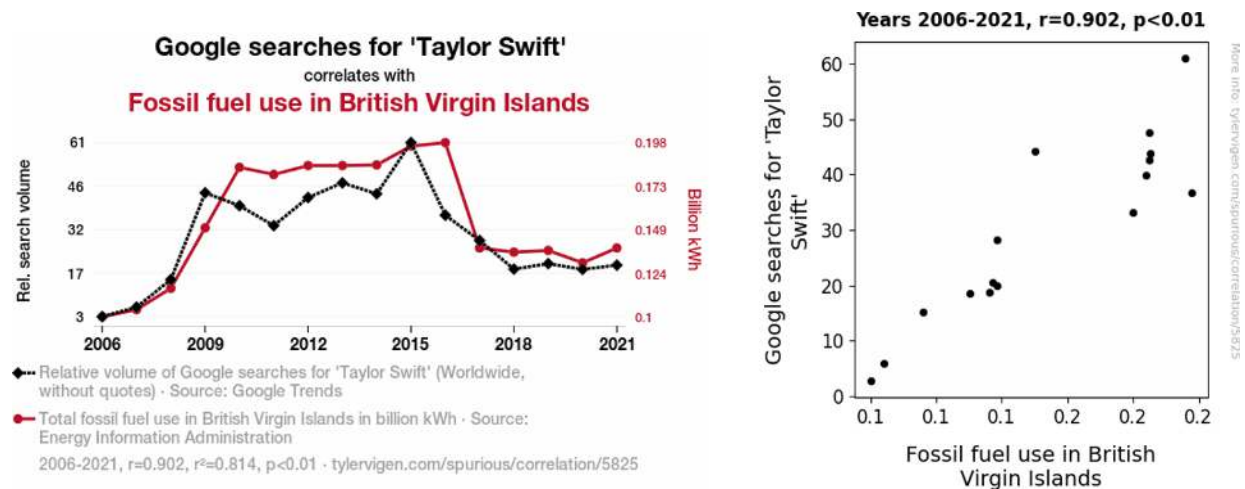
Data Debate 7.6.4 In [Figure 7.9](#), what would it mean for a data point to be an “outlier”? Can you identify two of them?

Different types of mathematical functional relationships are used to represent the dependence of Y on X , not just linear relationships. Quadratic, logarithmic and logit functions are common ways of capturing complex relationships between variables, both categorical and continuous. A rigorous introduction to generalized linear models is provided in McCulloch and Nelder's *Generalized Linear Models* [29]. Other regression models use different functions in different regions of the attribute space called piecewise or segmented or spline regression models. Deep learning networks create dense multi-layered functions that can capture more complex relationships in an empirical manner from the data.

7.6.4 Beware: Correlation is not causation

A word of caution about correlation. It is not causation, just a co-occurrence of pairs of values of two variables. Sometimes it could be just chance that two variables have a similar trend, e.g., according to [Figure 7.10](#), the number of Google searches for Taylor Swift and the use of fossil fuel in British Virgin Islands, have a correlation of 0.902! Tyler Vigen's [Spurious Correlations](#) contains hundreds of such

examples of nonsensical correlations and explains why these are spurious. The fallacy here is the coupling of the two variables into ordered pairs based on time. Any two variables that trend similarly over time, e.g., first name Tyler and the amount of gasoline pumped in Italy, will exhibit high correlation.



[Extended Description](#)

FIGURE 7.10: Tyler Vigen's website [Spurious Correlations](#) contains hundreds of examples of spurious correlations, and explanations of why they are suspect. Left: Time series of fossil fuel use in the British Virgin Islands and the number of Google searches for “Taylor Swift” during 2006 - 2021. Right: Scatterplot of the same data. [🔗](#)

While some spurious correlations are due to chance, others might be driven by a hidden variable, called a *confounding factor* that influences the two variables separately. A common example of this is the correlation between ice cream sales and drownings. They are positively correlated, but this does not mean that an increase in ice cream sales causes more people to drown. There is another factor (confounding variable) that has to be taken into account: high temperature. When it is hot, more people go swimming, and this can lead to drownings; and when it is hot, more people buy ice cream.

Learn more about correlation and causation by exploring Tyler Vigen's [Spurious Correlations](#) website and [GeminiBox 7.6.2](#).

GeminiBox 7.6.2: Correlation and Causation [🔗](#)

“What is the difference between correlation and causation?” [🔗](#)

7.7 Time and Space Are Special

Time and space are special types of continuous variables because *proximity in time* and/or *space* has a strong influence on outcomes. For example, areas close to a wildfire are at a higher risk than those far away. Similarly, the closer the date of an important event like a wedding or a critical examination, the

greater the levels of stress. Air traffic controllers monitor aircraft both in time and space to avoid collisions caused by two planes that are close together in both time and space.

Data that are tracked over time are called *temporal*, *longitudinal* or *time-dependent*. Data that are measured across space are called *spatial data*. Spatial data can be geospatial data that refers specifically to the earth's surface, e.g., latitudes and longitudes, or any type of *location* or *positional information*, e.g., on the human body, on a mechanical system, or on an image or video. *Spatio-temporal data*, like the ones monitored by air traffic controllers, are measured across both space and time.

Temporal and spatial data are complex types of data objects that require specific formats and data structures for efficient analytics. However, they are an essential part of statistics and AI, for studying real-world phenomena like hurricanes, stock markets and self-driving cars. They entail extensive study. In this chapter, we will introduce the concepts of temporal, spatial and spatio-temporal data, and study motivating examples connected to real-world AI applications that use these three types of data. *The Analysis of Time Series: An Introduction* [8] by Chris Chatfield provides a thorough introduction to time series. *Statistics for Spatial Data* [11] by Noel Cressie covers the topic of spatial data exhaustively.

7.7.1 Time-dependent data

Many characteristics and variables change with time. Our vital signs like blood pressure and heart rate, the position of the sun in the sky, and stock prices all vary continuously. Being able to monitor and potentially predict them is a critical task and spans multiple applications in medicine, business, ocean science, astronomy, sports and entertainment. Time as a variable requires special considerations because:

- *Proximity in time*: Observations taken close together in time tend to resemble each other.
- Time is *cyclical* in the sense that certain patterns repeat: the clock shows 3 AM daily, Mondays occur every seven days, the winter solstice happens once a year, and so on, resulting in certain events to repeat systematically, e.g., ebb and flow of tides, commuters on trains, and holiday-related shopping volumes.
- Many events have a *temporal signature* that helps in forecasting extreme events. For example, hurricanes are preceded by a build-up of wind and pressure, and they are followed by a big drop-off.

For these reasons, variables that change with time need a special framework.

A time-dependent variable is denoted by y_t , where the subscript t represents time at which the observation y was made. For example, the following measurements represent daily values of Apple stock, from kaggle's DJIA 30 stock time series data,  shown in [Table 7.2](#). The first column represents time, which is measured in days. There are five variables that change with time: opening price for that day, high, low, closing price and number of shares traded that day.

TABLE 7.2: Daily stock price data for Apple's stock ticker AAPL from January 3, 2006 to January 26, 2006. 

Date	Open	High	Low	Close	Volume
2006-01-03	10.34	10.68	10.32	10.68	201853036
2006-01-04	10.73	10.85	10.64	10.71	155225609
2006-01-05	10.69	10.7	10.54	10.63	112396081
2006-01-06	10.75	10.96	10.65	10.9	176139334
2006-01-09	10.96	11.03	10.82	10.86	168861224
2006-01-10	10.89	11.7	10.83	11.55	570088246
2006-01-11	11.98	12.11	11.8	11.99	373548882
2006-01-12	12.14	12.34	11.95	12.04	320201966
2006-01-13	12.14	12.29	12.09	12.23	194153393
2006-01-17	12.24	12.34	11.98	12.1	209215265
2006-01-18	11.87	12.01	11.69	11.78	303279501
2006-01-19	11.61	11.67	11.25	11.29	424100026
2006-01-20	11.33	11.43	10.83	10.87	284634847
2006-01-23	10.87	11.37	10.86	11.1	264942587
2006-01-24	11.25	11.35	10.82	10.86	285661873
2006-01-25	11.06	11.07	10.46	10.6	319019939
2006-01-26	10.65	10.78	10.28	10.33	295599087

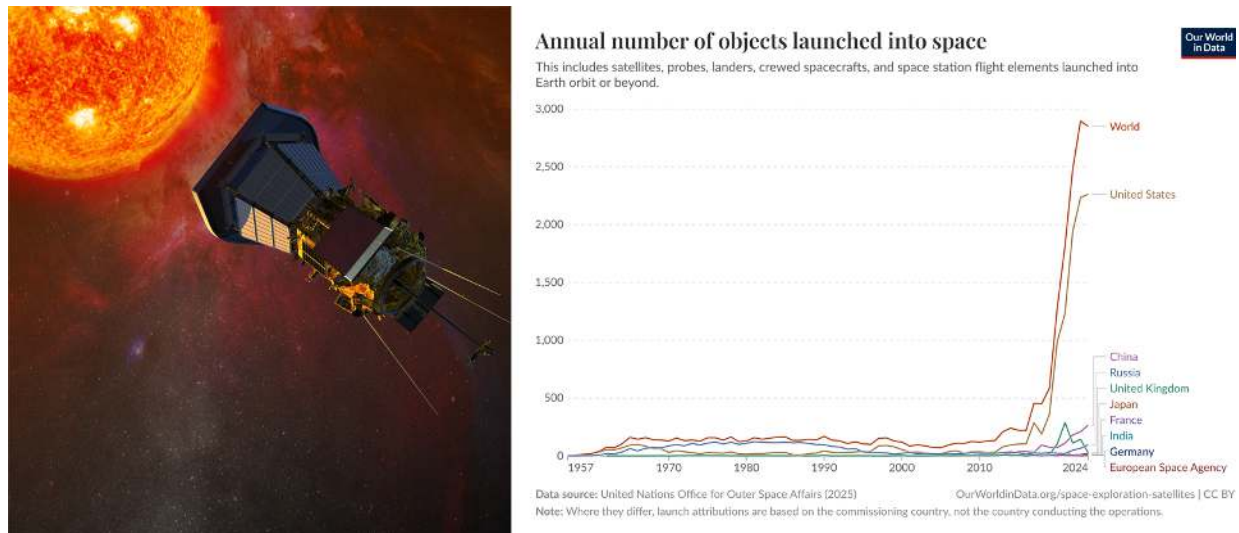
Data Debate 7.7.1 *Is the stock ticker name AAPL time-dependent? Do values of Open change gradually over time? How about Volume? Are there gaps in the time variable, Date? Why?*

There are many classes of time-dependent variables that do not change systematically or predictably with time. Some, called *stochastic processes*, change in a probabilistic way. For example, the way a helium balloon drifts once released into the air outdoors. *A First Course in Stochastic Processes* by Karlin and Taylor [25] provides a great introduction for students with a math background. Yet another class of time-dependent variables is *data streams*, e.g., the number of clicks per second on a popular app. Their behavior changes with time but in ways that cannot be captured by compact statistical models. Data-driven approaches like histograms are used to capture such instances. Time-dependent data, also known as *longitudinal data*, is a vast topic, and we will present only a brief overview of traditional time series. *Analysis of Longitudinal Data* by Peter Diggle et al. [14] provides a rigorous introduction for students with a math background.

7.7.1.1 Illustrative example: Spacecraft telemetry and AI

NASA launches spacecraft regularly to understand the universe, e.g., to analyze the composition of planets and other space objects. Figure 7.11, left, shows the solar probe Parker, formerly known as Probe Plus. It was launched in 2018 and entered the sun's atmosphere in 2021, sending back a treasure of previously unseen data. Such spacecrafts collect enormous amounts of data called *telemetry* to track their own functioning, e.g., variables such as temperature of various components, pressure, performance

indicators of computing systems, fuel levels, loads on various systems and location coordinates. Continuously tracking such variables enables engineers to flag any anomalies, e.g., high temperatures or stress on systems, and take preventive action before the occurrence of catastrophic failures. According to an article, *Detecting anomalies in spacecraft telemetry* [🔗](#), published by Aerospace organization:



[Extended Description](#)

FIGURE 7.11: Left: Artist's rendering of the Parker Solar Probe (formerly Solar Probe Plus) launched in 2018 that plunged into the solar atmosphere in 2021 and sent back data that had never been collected before. The progress of such probes is tracked through telemetry data that continuously monitors health and functioning of the probes to detect anomalies early and ensure mission success. AI methods are increasingly used to handle the staggering amounts of time series data generated and collected by spacecraft. (Source: NASA Goddard Space Flight Center from Greenbelt, MD, USA, CC BY 2.0.) Right: The number of objects launched into space has increased exponentially. (Source: Our World in Data.) [🔗](#)

The volume of data returned by spacecrafts continues to increase with missions such as the NASA-Indian Space Research Organization Synthetic Aperture Radar that will generate about 85 terabytes of data per day. Spacecraft telemetry is inherently time-series data, and many anomaly detection approaches which exist today lend themselves well to this type of data.

As the number of space objects grows as shown in [Figure 7.11](#), right, AI becomes critical in managing the safety and operational integrity of these spacecrafts by monitoring telemetry time series data.

7.7.1.2 Illustrative example: NOAA's weather forecasting

Average temperatures are classic examples of time series, as seen in [Figure 7.12](#), which represent average monthly temperatures in Florida. Notice:

1. The values of the average temperature rise and fall in a cyclical pattern, hitting 90's in summer and falling to 70's in the winter. This is called *seasonality*.

2. The blue dots are the actual values of the temperatures each month which vary depending on factors unique to that year and month, e.g., wild fires, unusual ocean currents due to glaciers melting, and other unexpected phenomena. Years 2017 and 2023 had unusually warm temperatures compared to other years, and years 2018 and 2025 had unusually cool temperatures.
3. The red lines are *smoothed* values that represent expected values based on historical data. They tend to be less variable from year to year since they are averages over long time windows of multiple years.
4. The black dashed line is the *trend line* intended to show whether there is a drift in the values over a period of time.
5. The differences between predicted (red) and actual (blue) values are called *residuals* and represent how good the prediction is. Notice that years 2016, 2018 and 2025 have big residuals for the winter month predictions.

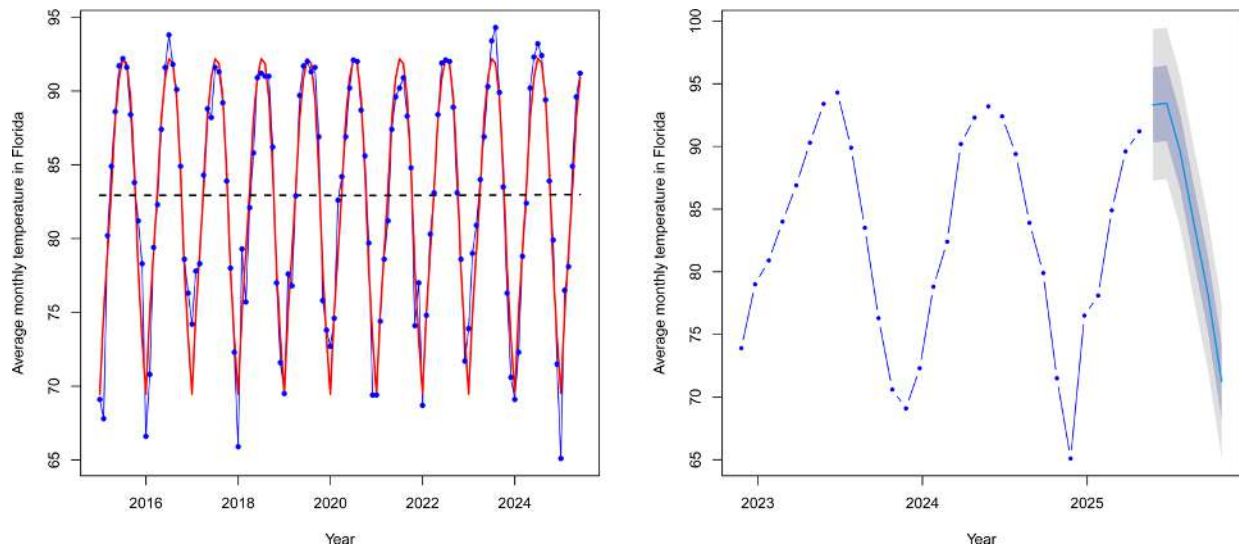
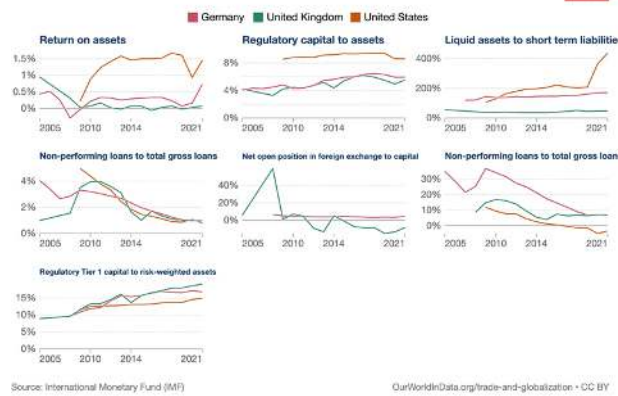


FIGURE 7.12: Average monthly temperatures in Florida are provided by NOAA National Centers for Environmental Information, Climate at a Glance: Statewide Time Series, published July 2025, retrieved on July 25, 2025 from [NOAA's Climate at a glance](https://climate.noaa.gov/data/at_a_glance/statewide/time_series). The graphs and forecast were created using R. [📄](#)

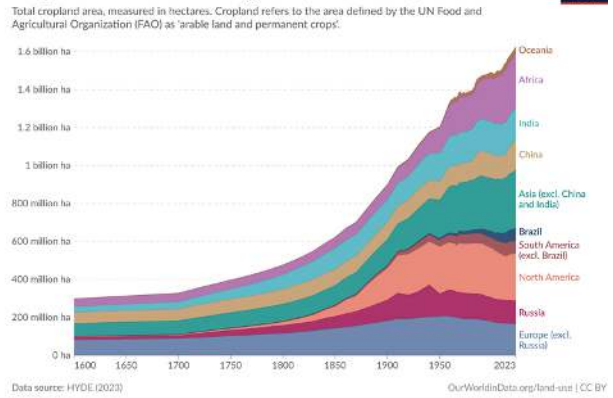
Seasonality is helpful in anticipating sales of seasonal items such as swimsuits and skis, in predicting temperatures and rainfall, and by airline industry for preparing for a holiday travel surge. Astronomers and space scientists rely on repetitive orbits of space objects to plan their experiments and studies.

Many time series exhibit a drift over time. Global ocean temperatures are rising, recreational travel is increasing and the incidence of many diseases are on the decline thanks to vaccines. This drift in one direction over time is called *trend*. Time series like those in [Figure 7.13](#), left, do not have a seasonal component but show a very strong trend. Some financial indicators have declined over time (“Non-performing loans as % total gross loans”), while others have increased (“Regulatory Tier 1 capital to risk-weighted assets”) or stayed steady (“Regulatory capital to assets”).

Financial soundness indicators



Cropland extent over the long-term, 1600 to 2023



Extended Description

FIGURE 7.13: Time series with trend. Left: Financial soundness indicators. Right: Expansion in cropland by region. [↗](#)

Figure 7.13, right, shows that cropland expanded as agriculture became the engine of growth initially, and then reached a steady state with industrialization or when there was no more room left for expansion. Africa shows a rapid gain while countries like India have reached a steady state as most of the land has been used up. Highly industrialized nations like the USA see a shrinkage of cropland over time. This kind of layered visualization of time series enables comparison of different groups over time while also keeping track of the total.

It is clear from these examples that time series applications routinely address numerous real-world situations where change is frequent, and monitoring it is essential to provide accurate responses e.g., in hurricane warning systems or in monitoring the economic productivity of nations.

7.7.2 Time series

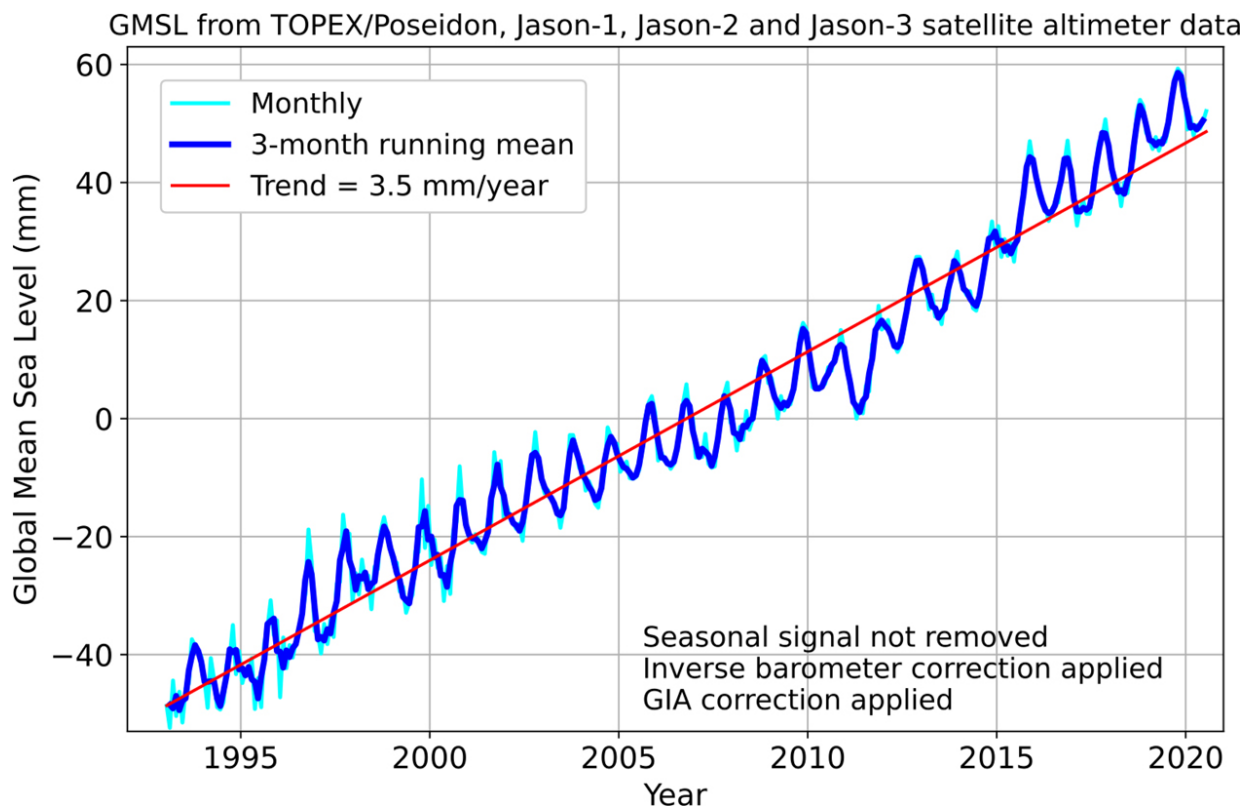
A time series y_t in its simplest form can be represented mathematically as a combination of three components: seasonality, trend and residuals which are what remains after the seasonality and trend have been subtracted from the observed data. For example,

$$y_t = S_t + T_t + \epsilon, \quad (7.7)$$

where the subscript t indicates the time instant at which the measurement of the time series variable y_t is taken, S_t represents the seasonal component at time t , T_t represents the trend at time t , and ϵ is what remains, usually random noise if the model is a good fit for the data. Such a representation of a time series as components that are added (or combined in other ways) is called a *decomposition*. There are many software tools available for this purpose, including *R. Time Series Analysis and Its Applications: With R Examples* by Robert H. Shumway and David S. Stoffer [39] provides an excellent hands-on introduction to time series.

To understand Equation 7.7, consider [Figure 7.14](#), a time series corresponding to monthly average sea levels. We present a very high-level explanation of time series decomposition.

- The peaks and troughs of the dark blue line (3-month running mean) represent changes in sea levels during the course of a year as they rise and fall with the seasons. A peak and the corresponding trough constitute a one-year cycle. This is the seasonal component S_t in Equation 7.7 that repeats in a broadly similar way year after year.
- However, while the seasonal pattern is present, there has been an increase in the sea levels year to year evidenced by the fact that each year's peak is roughly higher than the previous year's with a few exceptions here and there. This upward drift of the time series is the trend portion T_t of the time series, represented by the red trend line.
- Finally, the actual monthly average is the cyan line. The gap between the cyan and the blue line at any time t constitutes the error ϵ_t of the time series.



[Extended Description](#)

FIGURE 7.14: The development of global sea level from 1992 to 2020 with seasonal variations. Measurements have been conducted via satellites TOPEX/Poseidon, Jason-1 and Jason-2. The trend line clearly shows the constant increase in global sea level of 3.5 ± 0.4 mm/year since the 1990s. Image credit: Geek3. [📄](#)

These three together constitute the time series, and represent the individual components of the decomposition.

The objective of constructing a time series model like the one in Equation 7.7 is to be able to describe the time series in a compact form without listing all the individual values. If the model is a “good fit”, then the residuals will be very small and exhibit no pattern, just random noise that follows a Standard Normal distribution.

Equation 7.7 is useful for forecasting future values of a time series. [Figure 7.12](#) shows the forecasted values of average temperatures in Florida based on historical data from 2023 to 2025. The forecasted values were created using *R*. The forecast is shown as a solid line, with the 80% and 99% confidence regions highlighted in shades of gray.

Data Debate 7.7.2 In [Figure 7.12](#), what is the proper way to interpret the shaded confidence regions? Of the two nested gray regions, which one corresponds to 80% confidence and which one to 99%?

Forecasting plays an important role in NOAA's hurricane predictions. NOAA focuses its research on new models and algorithms, data gathering techniques, and innovative instrumentation to predict wind speeds and precipitation amounts. Such forecasts help create timely alerts and evacuations to better protect life and property. Read about NOAA's hurricane forecast research in [Section 7.8](#).

7.7.3 Spatial data

Geographic data, also called geospatial data, are specified through coordinates such as latitudes and longitudes, e.g., the public schools location data sample shown [Table 7.3](#), where the last two columns contain the latitude (LAT) and longitude (LON), respectively. We rely on geospatial data daily when we use our car's GPS to drive from point A to point B. Geospatial data, which refers particularly to the surface of the earth, is used extensively to study phenomena like wildfires, earthquakes, deforestation and animal migration. Spatial data in general can refer to any data that identifies a position or location, e.g., inside an organ like the brain which can be viewed via medical imaging.

TABLE 7.3: Example of spatial data: locations of public schools are specified by latitude and longitude, the last two columns. Published by National Center for Education Statistics. [↗](#)

NAME	STATE	CITY	ZIP	NMCNTY	LAT	LON
Albertville Middle School	AL	Albertville	35950	Marshall County	34.2602	-86.2062
Albertville High School	AL	Albertville	35950	Marshall County	34.2622	-86.2049
Albertville Intermediate	AL	Albertville	35950	Marshall	34.2733	-86.2201

School NAME Alberville Elementary	STATE AL	CITY Alberville	ZIP 35950	County NMCNTY Marshall	LAT 34.2527	LON -86.2218
School				County		

Spatial data, as well as spatio-temporal data in [Section 7.7.4](#), require highly complex analytics which are beyond the scope of this book. For our purpose, it is enough to understand where such data arise and where important AI applications are used.

Data Debate 7.7.3 *What makes spatial data special?*

Demographic studies use geospatial data extensively. Consider [Figure 7.15](#), right, created by the Federal Reserve Bank of New York. It is an interactive map that shows the change in home prices in counties across the USA. Check out this link [🔗](#) to explore counties across the country and identify how real estate prices changed across different regions in the USA over time. Unusual events like the COVID-19 pandemic create abnormal patterns as people flee densely packed urban areas for homes in the open countryside.

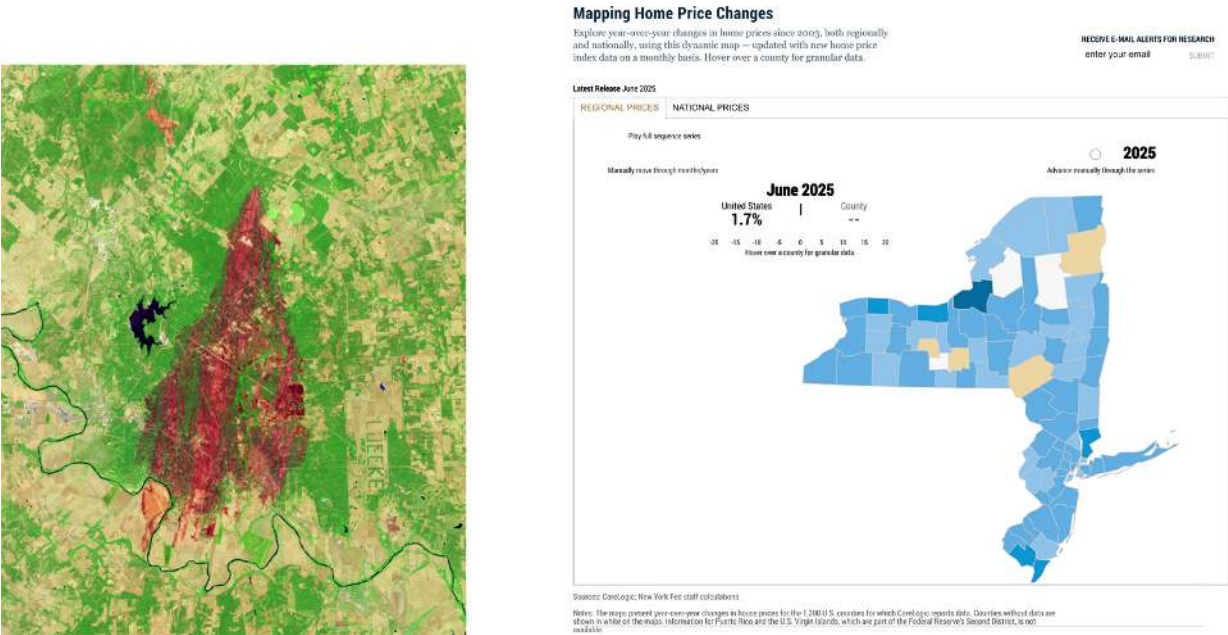


FIGURE 7.15: Spatial data is used in AI applications ranging from detecting wildfires to studying real estate markets. Left: The Advanced Land Imager (ALI) on NASA's Earth Observing-1 (EO-1) satellite captured this image of the Bastrop County Complex Fire in southern Texas on September 12, 2011. Right: Changes in house prices in June 2025 by county in Northeastern USA. Source: [🔗](#), Mapping home prices: Federal Reserve Bank of New York. [📄](#)

Like time, proximity in space influences neighboring regions. Areas close to a wildfire are at a much higher risk than areas far away. Similarly, when epidemics spread, areas close to the center of the epidemic are at much higher risk of infection. For this reason, spatial data requires a special type of

models. AI has accelerated the use of this complex data form, leading to applications in wildfire detection and management, dynamic real-time real estate applications like Zillow, automated medical image screening, and astronomy.

7.7.3.1 Illustrative example: Prithvi GeoAI models for earth sciences

GeoAI refers to specialized AI models built for geospatial data. Such data can be voluminous and complex. NASA uses its Prithvi AI model, built in collaboration with IBM, to fill in gaps in satellite imagery caused by cloud cover, to distinguish between flood water and permanent water on the earth's surface to enable flood water management, to monitor crops on agricultural land, and to map scars caused by wildfires in order to direct recovery efforts. [Figure 7.15](#), left, shows the burn scar of a wild fire in Texas. GeoAI studies such images to train AI algorithms to detect burn scars to monitor habitability and recovery.

Data Debate 7.7.4 *What type of data would you need to collect to create a figure such as the one in [Figure 7.15](#), right?*

7.7.4 Spatio-temporal data

Monitoring events across space and time is essential for many tasks, e.g., spread of wildfires and floodwaters, movement of birds and animals, propagation of misinformation across a social network, and flow of vehicular traffic. Such data have both a spatial and temporal component and are called *spatio-temporal* data. Analyses of rocket trajectories, paths of autonomous cars, flight paths, and migration of herds across the Serengeti in monsoon are all examples of spatio-temporal data and entail studying the effect of proximity in space and time. Spatio-temporal data tend to be extremely complex and need special types of models and data structures that are beyond the scope of this book. We focus on the applications of such data in real-world examples.

7.7.4.1 Illustrative example: NOAA tracks hurricanes with AI

Consider [Figure 7.16](#) which shows the tracks of category 1, 2 and 3 hurricanes in Florida. The left panel shows the tracks observed over the twenty-year period between 2004 and 2023. The panel on the right shows the tracks of the year 2024. These tracks are classic examples of spatio-temporal data that represent movement in time over a spatial region. They are called *trajectories*. The figures were created using the Historical Hurricane Tracks Viewer provided by climate.gov to explore historical hurricanes and their trajectories.

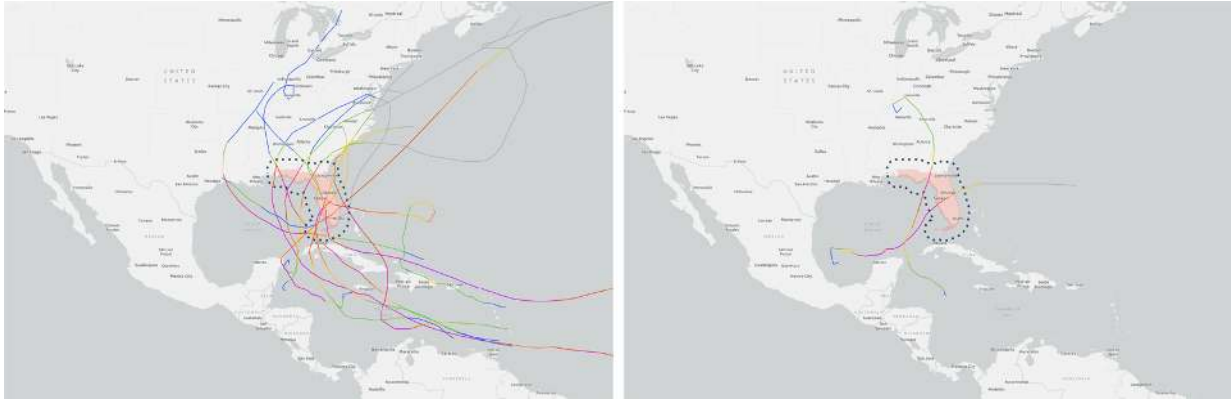


FIGURE 7.16: Tracks of Category 1, 2, and 3 hurricanes in Florida over the 20-year period 2004-2023. Right: Tracks of Category 1, 2, and 3 hurricanes in Florida, 2024. From Historical Hurricane Tracks Viewer, climate.gov. [🔗](#) [📄](#)

What do you notice about the tracks of 2024? Do you see their reach into the Gulf of America, formerly known as Gulf of Mexico? Was that observed in the past 20 years? Would you have expected that track based on the preceding 20 years? Is that trajectory an anomaly?

NOAA conducts extensive research in the use of AI for monitoring the climate and environment. The organization has developed an innovative system called Warn-on-Forecast (WoFS) shown in [Figure 7.17](#), left, to forecast severe weather. WoFS is designed to make probabilistic predictions using the latest radar and satellite data. Read the article in [Section 7.8](#) and watch this YouTube video [🔗](#) from NOAA about WoFS advanced technology that helps save lives and property.



[Extended Description](#)

FIGURE 7.17: Left: Innovative Warn-on-Forecast System from NOAA. Right: Plotting the path of an autonomous vehicle. [🔗](#)

Data Debate 7.7.5 *How is WoFS changing weather forecasting?*

7.7.4.2 Illustrative example: AI and path planning for autonomous cars

For a driverless vehicle to function properly, its AI-driven software has to accurately recognize different types of data such as images, movement, traffic signals, weather conditions, pedestrians and other

vehicles as shown in [Figure 7.17](#), right, and generate a *path* for it. This requires generating many possible paths and finding the most efficient route with the least risk. There are special AI tools called *path finding algorithms* that create these automatically. The paths are examples of spatio-temporal data which specify a series of locations and times.

Spatio-temporal data tend to be very challenging and voluminous with many types of dependencies. They require advanced AI tools. Learn more about them by exploring the prompt in [GeminiBox 7.7.1](#).

GeminiBox 7.7.1: AI tools for spatio-temporal data

“What are examples of AI tools for studying spatio-temporal data?” 

7.8 Read & Reflect

- *9 Ways AI is helping tackle climate change* 
- *Climate change AI* 
- *The RealReal uses AI to spot fakes* 
- *AI powered hydroponics: Revolutionizing sustainable agriculture* 
- *Warn on Forecast* 
- *Five ways NOAA research improves hurricane forecasts* 
- *AI is helping seismologists find the next monster earthquake* 
- *Expanded AI Model with Global Data Enhances Earth Science Applications* 
- *NASA images and media license information* 
- *The USGS Cooperative National Geologic Map on smartphone* 

7.9 Exercises

1. **Climate monitoring U.S. Tornadoes by NOAA.** In this exercise you will use the Tornado statistics data provided by Storm prediction center from January 1950 to April 2025 for the Contiguous U.S. Here is the link to the data set and an interactive chart: 

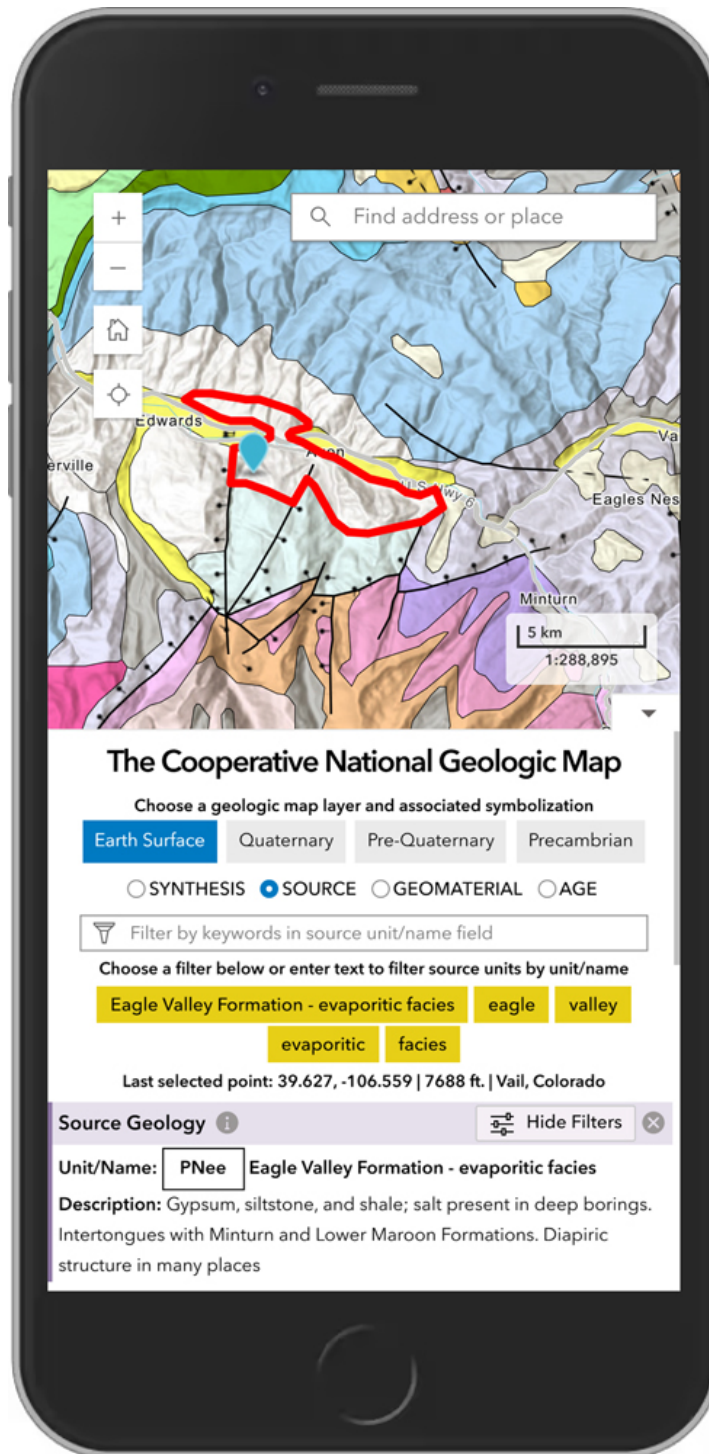
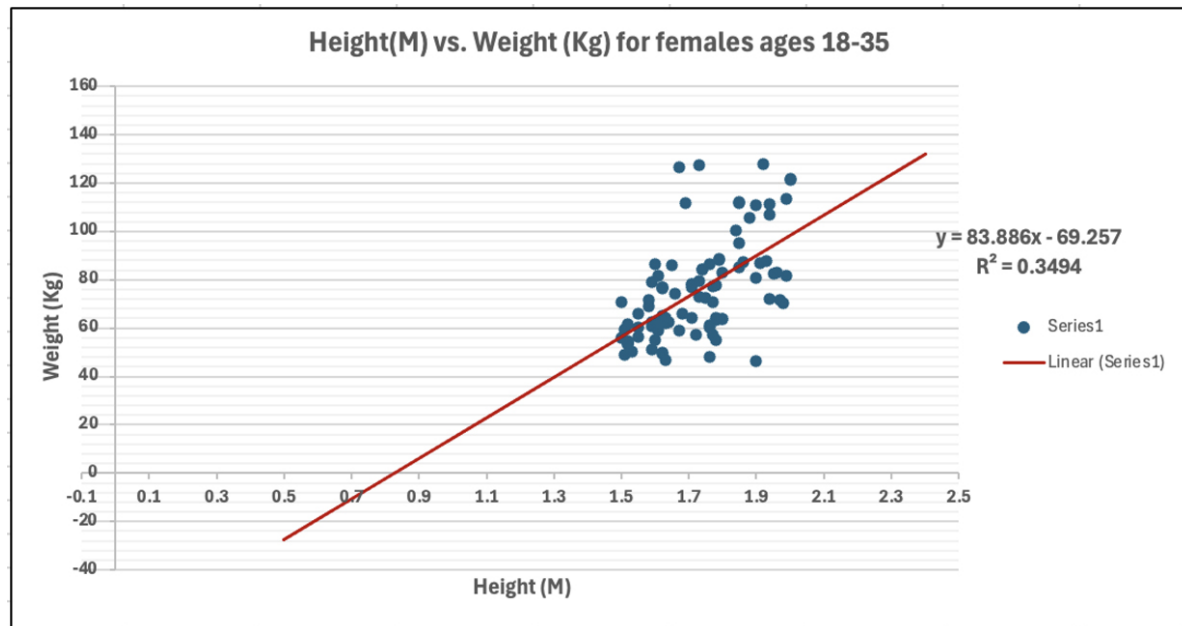


FIGURE 7.18: The new USGS Cooperative National Geologic Map web tool was designed to also be mobile-friendly, allowing users to explore geologic data from smartphones and tablets, in addition to larger devices. Image: USGS.

(a) Copy the data from the website to a Google Sheets or Excel spreadsheet.

- (b) Using the insert chart function in Google Sheets create a scatterplot of the data for tornadoes vs. fatalities.
 - (c) What do you observe in the scatterplot? Explain.
 - (d) What are some variables or factors that can help reduce fatalities during a tornado?
2. **Is there a relationship between number of calories and number of grams of sugar or carbohydrates in breakfast cereals?** In this exercise, you will use the data from [kaggle.com](https://www.kaggle.com). You will use built-in Excel or Google Sheets functions to create scatterplots and then draw a line of regression.
- (a) You want to examine if there is a relationship between the number of calories and the grams of sugar in cereals. Download the cereal.csv file and copy it to an Excel or Google spreadsheet. In a second sheet, copy the columns that contain the data for calories and sugar. Select the data and then from the pull-down menu, select scatterplot. This will insert a scatterplot on the sheet. Here is the link to the cereals data set on kaggle: [🔗](#)
 - (b) Label the axes and give a title to the chart. What do you observe in the plot? Explain.
 - (c) Explore the chart features and customize the chart to include a line of best fit. Select equation and R^2 .
 - (d) What is the coefficient of correlation? Is there a strong correlation between the variables?
 - (e) What is the conclusion about calories and sugar in cereals based on this chart and equation? Do you agree? Could there be other variables that can affect calories.
 - (f) Repeat the steps above to find the relationship between the number of calories and carbohydrates in cereals.
 - (g) Create a scatterplot with both variables: sugar vs. calories and carbs vs. calories.
3. **What does the line of regression tell us?** In this exercise, you will examine scatterplots and lines of regression created using data from [kaggle.com](https://www.kaggle.com).
- (a) The scatterplot in [Figure 7.19](#) was created in Excel using a random selection of data (height and weight) of 50 female members of a gym between the ages 18-35. [🔗](#)
 - (b) What are the variables represented in this figure?
 - (c) Are there any outliers? Identify two data points that can be outliers.
 - (d) Do the variables have a positive or a negative correlation? How can you tell?
 - (e) What is the coefficient of determination R^2 ? What does it mean?
 - (f) Using the regression equation, find the weight of a member whose height is 1.6 m.

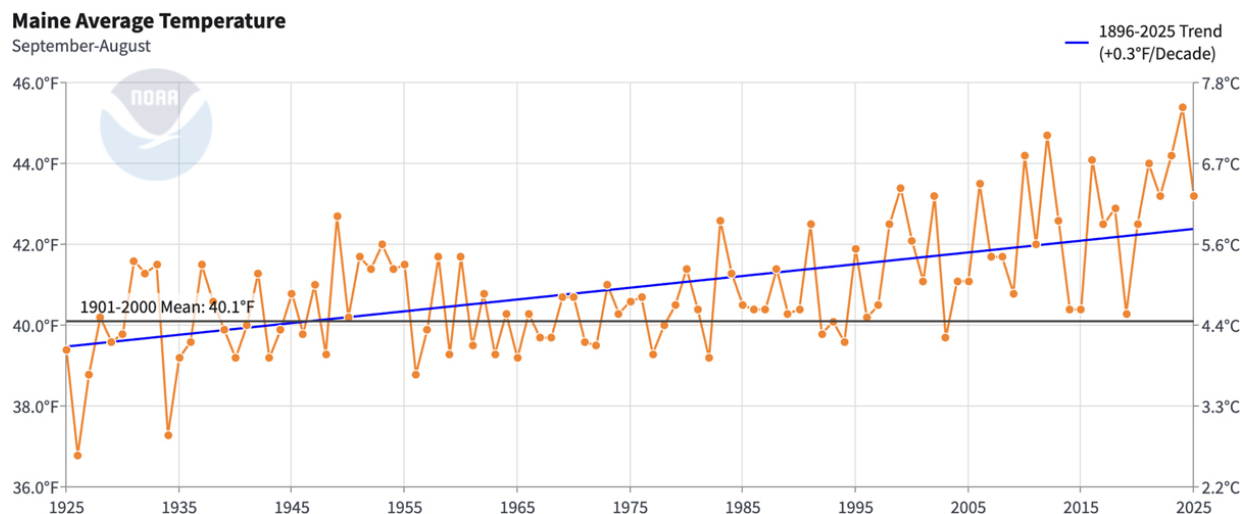


[Extended Description](#)

FIGURE 7.19: The relationship between height and weight of women members of a gym between the ages 18-35. The chart was created using Excel. The data was derived from [kaggle.com](https://www.kaggle.com). [🔗](#)

4. **Geo-spatial analysis.** Suppose you are visiting a city you are not familiar with and want to find a good café in the morning which is walking distance from the hotel you are staying in. A common way to do this is to use your smartphone navigation app such as Google Maps or Waze to find the location of cafés nearby.
 - (a) Find five cafés near you using a map app.
 - (b) What are some criteria for you to select a café you want to go to?
 - (c) Using the app, find the latitude and longitude of the cafés you have selected.
5. **Exploring time series maps.** Explore NOAA's Climate at a Glance interactive tool to generate a time series chart for three different states or cities for different time periods (e.g., past 10 years, past 100 years). Source: NOAA National Centers for Environmental Information, Climate at a Glance: Statewide Time Series, published September 2025, retrieved on September 22, 2025 from NOAA's Climate at a Glance: [🔗](#)
 - (a) List what you notice about the graphs using the parameter average temperature.
 - (b) Try the parameter precipitation and explore.
 - (c) Try the parameter Palmer drought index and explore.
 - (d) Display the graphs with trend line. What information can you gather using the trend line? Copy the graph and explain.

- (e) The graph in [Figure 7.20](#) was created using the interactive tool from NOAA with parameters average temperature in Maine for the month of August from 1925 to 2025. What do you notice?



[Extended Description](#)

FIGURE 7.20: Average temperature for the month of August in Maine from 1925 to 2025 from NOAA, with blue trend line. [↗](#)

7.10 Summative Project

7.10.1 Chi-squared test of independence: How does time spent on homework impact your grade?

The guidance department at a high school wants to find out if there is a relationship between the time spent on completing homework each day by a student and the student's quarterly grade in Math. They conduct a survey of students in grade 9 about how much time they spend doing math homework and collect the data. There are a total of 215 students in grade 9, and the following table lists the frequencies of the two categorical variables - Time Spent on homework and Academic Performance. Is there a statistically significant association between the two categorical variables?

[Table 7.4](#) below has the data for the two variables:

1. Does the data satisfy all the conditions required for applying the statistical hypothesis test Chi-squared test of independence? Explain.
2. What are the two categorical variables?
3. What is the null hypothesis and the alternate hypothesis?
4. Conduct a Chi-squared test with significance level 0.05.
5. What are the degrees-of-freedom?

6. Calculate the expected frequencies for each cell and the Chi-squared statistic.
7. Using the Chi-squared distribution table, find the critical value of the statistic.
8. Compare it to the calculated value and state your conclusion.

TABLE 7.4: Distribution of Academic Performance by Time Spent on Homework. [📄](#)

	Good (A)	Average (B)	Poor (C or lower)	Row Total
Low (0-30 min)	10	30	40	80
Moderate (31-60 min)	40	35	25	100
High (> 60 min)	22	8	5	35
Column Total	72	73	70	215 (total students)

7.10.2 One-Way ANOVA: Is there a significant difference between the mean bill lengths of different species of penguins?

In this statistical analysis, we will examine the Palmer penguin data set using the hypothesis test One-way ANOVA to find out if there is significant difference in the mean bill lengths of the three species of penguins. We will choose a random sample of 20 penguins from each species from the Palmer Penguin data set.

1. What are the assumptions about the data before we can conduct the ANOVA test?
2. Does the sample data for penguins satisfy the assumptions for ANOVA? Explain.
3. What is the independent variable and the dependent variable?
4. From the Palmer Penguin data set, randomly select data for 20 penguins of each of the three species (Adelie, Chinstrap and Gentoo). Copy this data on an Excel spreadsheet.
5. Make a table with three columns, one for each species, and copy the data for bill lengths into these three columns.
6. What is the null hypothesis and alternate hypothesis? Select this table and using the Data Analysis feature of the Excel spreadsheet, conduct a one-factor ANOVA test for the data. Set a significance level of 0.05.
7. What is the F-Critical value and the Test-F value?
8. What is the conclusion?

Data Quality and AI

AI is built on data. It learns how to distinguish between cats and dogs by poring over hundreds of thousands of images labeled as cats or dogs. GenAI learns how to write novels by studying millions of sentences and articles. Customer service chatbots are trained on thousands of transcripts of conversations between customers and customer support representatives. What if cat photos were mislabeled as lions? What if the novel-writing bot was trained on stories written by babies picking out words at random from a box? What if the customer service transcripts were full of hateful words? What if AI was trained on images shown in [Figure 8.1](#)? AI would simply learn and repeat these errors. While algorithms and tools have advanced, the data used by AI is only now getting the attention it deserves. In this chapter, we will learn about different types of data quality issues, how they arise, how they are measured, consequences of fixing them, and AI-based solutions.



FIGURE 8.1: Data quality issues can derail AI. Left: Bad data can be incomprehensible and detrimental to automation. (Photo: US Government National Archives catalog.) Right: Data can be misleading and incorrect, and potentially harmful if used to train AI. (Photo: Wikipedia, Andrewj0131.) [📄](#)

Consider the following hypothetical situation: you need to finish an assignment in a hurry, but you cannot find the relevant file because you had labeled it as “job search” instead of “homework.” Your friend sends a copy over but only the first page of the scan shows up. Annoyed, you send out a text demanding he resend it, only to find that your friend's answer is utter gibberish, as if his smartphone could not decipher what he was saying. When you finally receive the entire assignment from him after some back and forth, you discover that the assignment had 20 questions, not the usual 5. You panic, since 20 questions would take you the whole day, but you have only 2 hours before the deadline. You check the file carefully and realize that each question had been repeated four times for some odd

reason, but because there were no question numbers, you could not tell that they were duplicates until after reading through them.

A further evaluation of the situation shows that it was caused by the confluence of several types of data quality issues. The first data quality issue was *incorrect labels*, i.e., mislabeled files that made the data hard to find, i.e., inaccessible. The second data glitch was *incomplete data*, when only the first page of the assignment showed up. The third data issue was the incomprehensible text message from your friend, a result of bad transcription, but human and software errors can result in such mangled messages as well. The fourth problem was *data duplication*, which could have happened if your friend resent the assignment multiple times. Finally, the *lack of metadata* (question numbers) masked the duplication. Each data quality problem came about in a different way, requiring a different fix. This is what makes data quality challenging, because it is highly context-dependent making it difficult to find general solutions that can be automated and applied widely.

However, given how critical data is to AI, it is important to develop intuition for data quality issues, to spot problems, and to evaluate solutions. This chapter focuses on different types of data quality issues and how they arise, methods to detect them and quantify their effect, and to clean data and prevent the occurrence of data glitches in downstream data pipelines used by AI applications.

AI pioneers such as Andrew Ng believe that for AI to achieve its true potential, it is necessary to focus on the data that drives AI, in a systematic manner. Read his article *Why it's time for data-centric artificial intelligence* in [Section 8.5](#).

While AI, machine learning algorithms, and statistical models consume data, they can also be used to ensure the quality of data. AI can monitor

data pipelines, flag data that does not conform to rules and expectations, identify instances of non-compliance to standards, and apply models to clean and validate the data.

8.1 Data Quality: Has your data gone bad?

Between the time it is created and the time it is used by AI applications, data passes through many processing stages. It originates in multiple places, e.g., gathered by sensors on different ocean buoys, is transmitted to data centers, integrated with other data such as weather measurements and topological features, processed to precise specifications, and finally made available to AI applications for use. Each stage might use different hardware and software environments, different procedures and processes with different error tolerances, resulting in a heterogeneous set of rules applied to the data. In addition to the errors that could impact data quality within each stage, the hand-offs between processing stages also contribute to data quality problems, e.g., when data are dropped due to incompatibility between two stages. It is important to understand how data quality issues can arise at different stages of the data pipeline. We discuss some common examples of data quality issues. However, the list is not exhaustive since data issues continue to arise in novel ways, as AI enables creating, collecting and storing new forms of data daily, some of them comprehensible only to machines, e.g., encrypted data.

8.1.1 Data pipeline and data quality

Consider the schematic representation of a data pipeline shown in [Figure 8.2](#). In the box labeled “Original Data”, each individual data point is

represented by a dot, where black dots are typical data points and an orange dot is a genuine anomaly, e.g., unusually high wind speed.

- During data collection, some measurements might be overlooked resulting in *missing data* (white dot with dashed lines). Missing data can have severe consequences, e.g., those related to healthcare and diagnoses ([Figure 8.3](#), left). Some might be misrecorded leading to artificial *anomalies* (orange dots) which cannot be distinguished from authentic data anomalies.
- When data is transmitted, it can get *mangled* (black oval), for example, “Doe” could be misspelled as “Doh”. Sometimes, data is retransmitted if there is no acknowledgment that it has been received, causing *duplication* of data (joined dots).
- When data is ingested into a data repository for storage, some records might be rejected (red dots) by the database depending on what constraints have been built in, e.g., “negative values of distance are not allowed,” or heuristics such as “fish cannot live outside water” or “riding a bicycle needs legs.” The latter two constraints would result in rejection of the image in [Figure 8.3](#), right. Some legitimate data might be dropped (white dot with dashed lines) if the storage limit is exceeded, resulting in additional missing data.
- While retrieving data, the database query might add some additional constraints, e.g., “remove all occurrences of Doh”, inadvertently resulting in legitimate “Dohs” in the data being dropped.
- During the data-cleaning stage, missing values might get *imputed* (gray dots), i.e., estimated because some AI applications cannot process missing data. Similarly, some anomalies might be brought in-line with

typical values (gray dot with dashed orange line) so as not to distort the results.

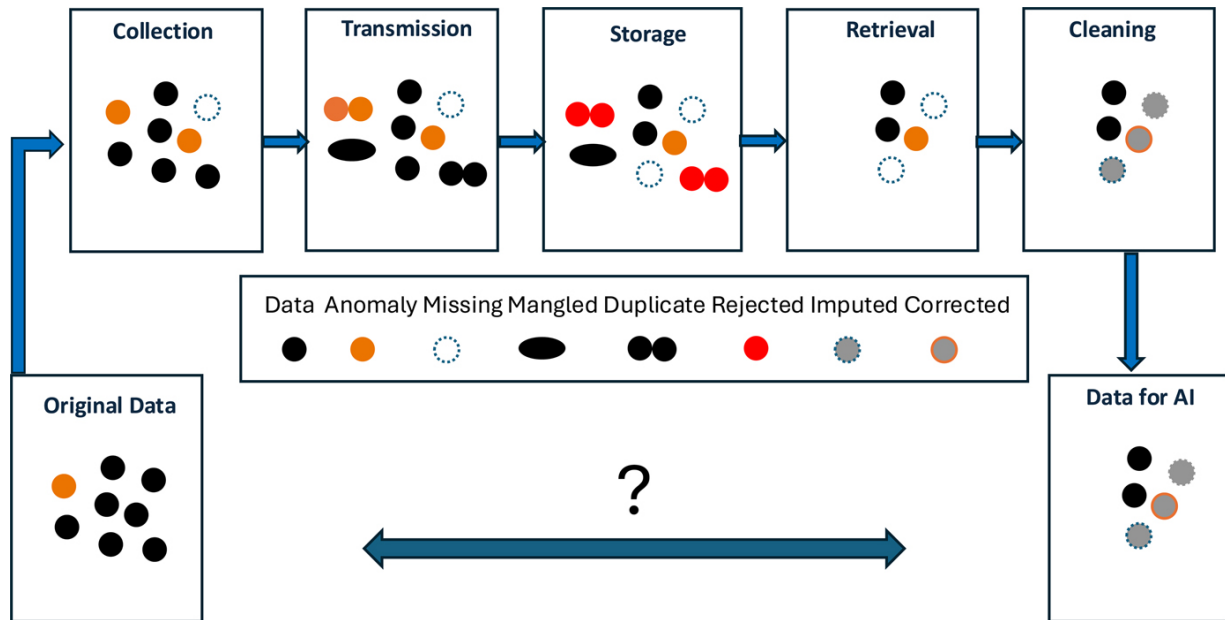


FIGURE 8.2: Data quality issues can arise at almost every point in the data pipeline. Data measurements (black dots) might be missed during the collection stage (white dot with dashed lines). Some might be misrecorded leading to anomalies (orange dots). When the data is transmitted, it can get mangled (black oval). Data might get duplicated if it is retransmitted. During storage, data might be rejected (red dots) by the database. Some data might be dropped (white dot with dashed lines) if there are storage limits. While retrieving data, a query might *inadvertently* result in additional data being dropped. During the cleaning stage, missing values might get imputed (gray dots), and anomalies might be brought in-line with typical values (gray dots with dashed orange line). By the time the particular data is available for use by an AI application, it might have very different statistical properties from the original data. Since AI and data science is based on statistical and probabilistic principles, such *statistical*

distortion during the cleaning process needs to be closely monitored. This figure was created using Microsoft PowerPoint. [📄](#)



FIGURE 8.3: Missing values can severely impact healthcare, e.g., incomplete medical history and its impact on clinical diagnosis. Left: NGOs like India's Ayushman Bharat use AI to manage data and deliver quality healthcare to remote regions and underserved communities. (Photo: Ganesh Dhamodkar.) Right: Inconsistencies in data can throw off the best AI algorithms. What if images like this one from the Guinness Storehouse in Dublin were used to train autonomous cars? (Photo: Nga Ahorangi.) [📄](#)

By the time the processed data is available for use by an AI application, it might have very different statistical properties than the original data. The question then becomes: “Does the processed data available to an AI application have similar statistical characteristics as the original data collected from the real world?” Since AI and data science are based on statistical and probabilistic principles, such *statistical distortion* that happens during the cleaning process needs to be closely monitored. This topic is discussed in [Section 8.4](#) and [\[13\]](#).

Data Debate 8.1.1 *Data cleaning changes data. Given this fact, should you clean data, and if so, how much?*

Data Debate 8.1.2 *Can you think of an example of data quality issues in sports that could mislead AI?*

Data Debate 8.1.3 *What would happen if GPS navigation software were trained on outdated maps?*

8.2 Monsters in Your Data: Types of data glitches

Data can be of poor quality for many reasons. For example, a surveyor might collect data from nearby regions since there might not be enough time to travel too far. A clinical study might drop data deemed negative, e.g., adverse side effects of a new drug. Software might fail, resulting in large chunks of sensor data not being recorded. All these situations result in different types of data quality issues, with different consequences, and necessitating different remedies. It is this variety that makes data quality challenging since data quality problems, their consequences and subsequent remedies are very closely tied to a particular application and do not generalize.

Classifying, quantifying and fixing data quality errors is an important part of ensuring that AI applications are trained on high-quality data. We start by describing some common types of glitches.

8.2.1 Missing and mangled data

Missing values are some of the most common causes of poor data quality. Imagine if half your test grades went missing and were automatically set to 0 by a grading software. How would that impact your final grade?

Misspellings like the ones shown in [Figure 8.1](#), left, are an example of mangled data. They not only make data incomprehensible, but also they can have serious consequences. In a medical database, if “Helen” is misspelled as “Ellen”, it could make it hard to locate medical records in an emergency where the delay of every second could mean the difference between life and death.

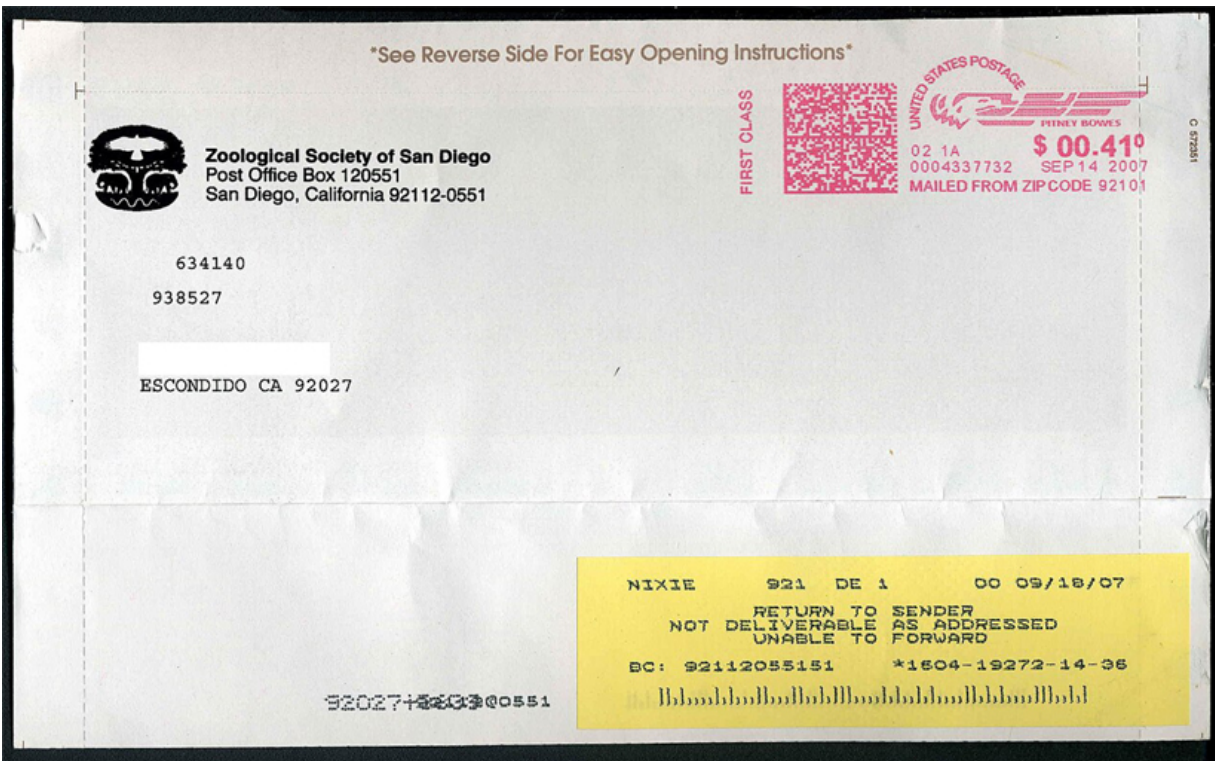


FIGURE 8.4: The United States Postal Service department sends back letters and packages if addresses are incomplete or wrong.

Data Debate 8.2.1 *AI tools can correct spellings and offer suggestions for completing a sentence. What would happen if the training data were full of typos and misspellings?*

8.2.1.1 Illustrative example: Healthcare, missing data and AI

In healthcare, missing values arise in different ways. Sometimes, lab reports go missing, either due to software problems, or actual hard copy records being misplaced. On occasion, the patient is a no-show. In such cases, there are gaps in the patient's medical history that could affect clinical decisions. For example, an incomplete medical history or an incomplete list of patient's medications could have serious consequences if the doctor is not aware of the missing conditions or medicines. The treatment the doctor would prescribe based on incomplete data could potentially conflict with those missing conditions and could have serious health consequences for the patient. AI-powered tools offer solutions by using advanced techniques to *impute* or fill in missing data, or even predicting no-shows so that patients could be reminded of upcoming appointments.

GeminiBox 8.2.1: Missing values

“How do missing values impact healthcare? How can AI help?” 

India has instituted data interoperability projects that use AI to ensure the flow of health data across systems, while preserving patient privacy for its 1.4 billion citizens. Read about how the world's most populous country uses AI to advance healthcare by preserving and digitizing data, and ensuring data integrity, in [Section 8.5](#).

Several non-profit organizations leverage open-source AI platforms to deliver healthcare to remote areas and underserved populations. AI is particularly good at focused repetitive tasks. Healthcare companies use AI agents to tackle and expedite routine tasks like medical transcription, to summarize case histories to shortlist diagnoses for the doctor to check, and refill medications, ensuring that data is recorded in a timely manner and updated constantly, thus ensuring data quality.

As another example of missing data, consider the following paragraph that is missing some words, and other words are jumbled and garbled. This happens in the real world quite often, e.g., when speech-to-text AI on smartphones encounters unfamiliar accents that it has not been trained to recognize.

Example 8.2.1. `Geetha likes to go to the brksfh and wefr a bxyk.` 

`Unless it rains, she pekudqs to stay in
her and drwqk a cup of`

Data Debate 8.2.2 *Is this paragraph usable as is? Can you guess what the original version is? Does your best guess for a correct version of the paragraph match with those of your classmates? What's the implication of different versions of the reconstructed data?*

The correct answer is: “Geetha likes to go to the beach and read a book. Unless it rains, in which case she prefers to stay in her room and drink a cup of tea.”

AI models are used to detect less obvious types of missing values too. For example, values used as a default, such as “January 1, 1970” for dates, occur with a much greater frequency than others. There are several reasons for this. For example, some software tools auto-populate defaults, e.g., “000-000-0000” for phone numbers, since some applications fail when they encounter blanks in the data. Such defaults are called *hidden missing values* because it is not always obvious that the values are just stand-ins. Unusual spikes in frequencies of values are often a hint that the value could be a default. It is surprising how many people were born on January 1, 1970! Check out [GeminiBox 8.2.2](#) to see what makes this date so popular.

Sometimes, it is not possible to tell whether values are missing. Consider [Figure 8.5](#) with its three panels. It is obvious that parts of the image are missing in the middle panel. However, if the original image on the left was unavailable, how would we know that the image in the right panel was incomplete?



FIGURE 8.5: Incomplete data is a subtle form of missing data, where it is not obvious that data might be missing. [↗](#)

Other forms of hidden missing values include data *imputed* during cleaning by filling in blanks with typical values like the mean. The data looks complete, with legitimate-seeming data. While this is helpful to some applications, others that rely on probability distributions might produce biased results if there is a prevalence of imputed values. We discuss below.

GeminiBox 8.2.2: January 1, 1970 [↗](#)

“What is special about the date January 1, 1970?” [🔗](#)

Missing values need to be addressed so that they are not propagated through AI data pipelines corrupting downstream AI models. Missing data could cause software to crash while performing certain mathematical operations. A worse scenario is that it could produce misleading or even disastrous results. For example, what if all the data from a particular set of sensors, like cameras on a robot, was missing?

8.2.1.2 Cleaning missing data

There are several ways of handling missing data. The simplest approach is to drop records where data values are not populated. Is that a good solution? Would we know about Geetha's preferences if that entire paragraph in [Example 8.2.1](#) were to be dropped from a customer preference survey? What if large chunks of the data values are missing?

Data Debate 8.2.3 *When could excluding records with missing values be a problem?*

Another option for treating missing data is to fill in missing data with usable values. This process is called *imputation*. We have already discussed the use of default values for qualitative variables, e.g., “000-000-0000” for phone numbers, or “John Doe” for a person whose name is not known.

To get an intuition for missing value imputation, consider the hypothetical example in [Figure 8.6](#). The table on the left shows three attributes for six individuals: Name, Gender and Age. The three blanks denote missing values. The table in the middle shows the data after the values have been imputed.

- The third record is missing gender. We use our knowledge of the real world, that “Nina” is usually associated with the female gender. So we fill in the blank for “Gender” attribute for this record with the value “F”.

This is a *heuristic* approach to imputing values, since we use common sense to make a quick guess.

- We fill in the name in the fourth record using a rule or *convention* for missing names: “Jane Doe”, for a female whose name is unknown.
- We fill in the weight in the third record based on the mean of all available values:

$$\frac{(144 + 135 + 125 + 132 + 116)}{5} = 130.4.$$

Note that the denominator is 5, and not 6, since dividing by 6 would imply including the missing value in the computation by setting its value to 0.

- The table on the right imputes a different value to Tom's weight by using only those weights that correspond to men.

Name	Gender	Weight	Name	Gender	Weight	Name	Gender	Weight
George	M	144	George	M	144	George	M	144
Kumar	M	135	Kumar	M	135	Kumar	M	135
Nina		125	Nina	F	125	Nina	F	125
	F	132	Jane Doe	F	132	Jane Doe	F	132
Tom	M		Tom	M	130.4	Tom	M	139.5
Dawn	F	116	Dawn	F	116	Dawn	F	116

[Extended Description](#)

FIGURE 8.6: Missing values in a data table; age imputed based on everyone; age imputed based only on men. [↗](#)

$$\frac{(144 + 135)}{2} = 139.5.$$

That is, additional information from another attribute “Gender” is used to refine the imputation of Tom's weight.

We summarize our three methods of imputing values for missing data:

- **Rule-based:** For qualitative variables, use rules based on heuristics, convention or empirical frequency distributions. For example, a frequency distribution of gender for the name “Nina” would presumably be 99% female.
- **Typical values:** Missing values of quantitative values are substituted with typical values like the mean or the median based on the available non-missing values.
- **Based on attribute relationships and simulated distributions:** Impute a value to a missing attribute based on the known value of another attribute. For example, impute a value to Tom's missing weight based on Tom's gender by using the mean weight of men. More sophisticated models based on larger data sets could use non-missing values to create an empirical probability distribution, and impute missing values by drawing values at random from the conditional distribution of weight given that the gender is “M”.
- **Model- and AI-based:** Values can be imputed by creating a predictive model, e.g., linear regression or even neural nets, for the variable with missing values using the remaining variables and their non-missing values.

AI uses language models to fill in missing text. Predicting future missing values, e.g., probability of a patient missing an appointment, is another way AI helps in anticipating missing values, and to take preventive action by nudging patients with reminders. Use the prompt in [GeminiBox 8.2.3](#) to do a deep dive into missing value imputation.

GeminiBox 8.2.3: Missing value imputation

“How are missing values imputed?” 

8.2.2 Incomplete data

Incomplete data is a special form of missing data. It is often more insidious since it is often not obvious that data is incomplete, unlike overtly missing values which are blanks such as those in [Example 8.2.1](#) and [Figure 8.6](#), left. [Figure 8.5](#) is an example of the difference between missing and incomplete data. The original image contains three plumeria flowers. Due to a software malfunction, parts of the image got obscured resulting in missing pixels as shown in the center image. AI techniques could be used to fill in the missing parts and reconstruct an approximate image.

On the other hand, the image on the right shows just the top-left corner of the original image. Without prior knowledge, there is no reason to suspect that the image is incomplete. It is possible that the photographer chose to focus just on the single petal. Furthermore, it is not possible to reconstruct the original image from just the fragment. However, the original might exist on the photographer's camera, or on any public social media site the photographer might have uploaded it to. AI tools, e.g., *Google Lens*, could potentially comb through publicly available images and find the original image to reconstruct the fragment.

Incomplete data could lead to serious problems. For example, what if your grade was based on half the questions you answered on a test? Or, doctors diagnosed and treated diseases based only on half of the symptoms or medical history? There are numerous examples of incomplete data that lead to failure. The article *How poor data quality impacts AI and machine*

learning in [Section 8.5](#) has details about how poor data quality affected AI in the following instances:

- Autonomous cars were unable to recognize vehicles from unfamiliar angles because the image data used for training was incomplete, and did not have pictures of vehicles from every possible angle, e.g., from the side, or turning a corner.
- AI used to screen resumé at a large corporation was trained on the data of existing employees and past applicants, predominantly men. As more women majored in computer science, the number of qualified women applicants exploded, but the training data remained incomplete and was not updated. The AI resumé screener discarded resumé that had female-sounding names or contained words that were “typically” associated with women.
- In another instance, automatic cleaning robots malfunctioned when their training data failed to include a complete set of images of all potential obstacles.

We discuss below the very real consequences of decisions based on incomplete data.

8.2.2.1 Illustrative example: Incomplete data and hardship in the EU

Decisions based on incomplete data can seriously impact millions of lives as happened in the European Union. In 2010, Harvard professors Reinhart and Rogoff published a paper that stated that if the debt held by a country is 90% or more of its GDP, then economic growth will slow drastically. This motivated countries, e.g., the UK, to impose severe austerity measures such as cutting public welfare programs and raising taxes, inflicting significant

hardship on millions of people. However, it turns out that the results were incorrect because the professors had inadvertently included only 15 rows of data in their Excel spreadsheet instead of all 20 rows. Analysis based on the entire data in the spreadsheet would have changed the conclusions significantly. The mistake was spotted by a graduate student at University of Massachusetts at Amherst, Thomas Herndon. You can read more by clicking here [🔗](#).

Greece, which had one of the highest debt to GDP ratio as shown in [Figure 8.7](#), was particularly devastated by the austerity measures which resulted in scarcity of healthcare. Unemployment went up among young people, leading to their mass exodus, leaving behind an aging population. Some of these effects are thought to be irreversible, and are described here [🔗](#).

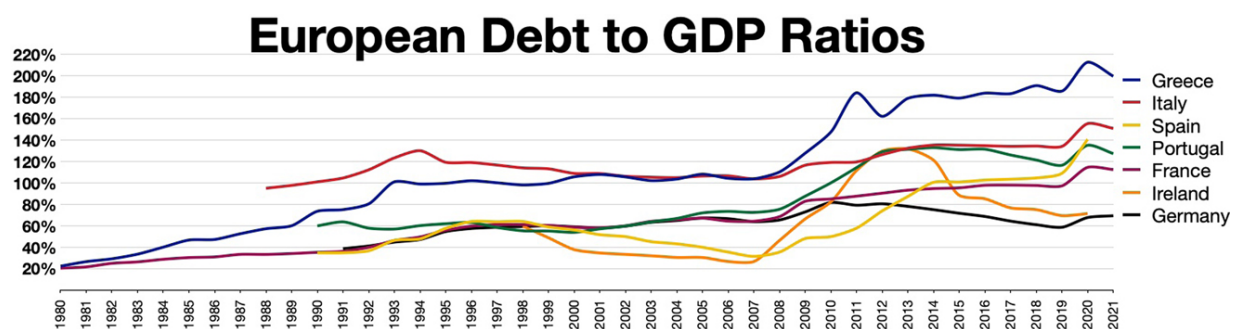


FIGURE 8.7: Incomplete data used by Harvard economists resulted in incorrect conclusions about debt to GDP ratios that led to austerity measures on countries like Greece that had high debt to GDP ratio. The economic hardships caused irreversible damage to the Greek economy. Subsequently, the findings of the study were reversed. [🔗](#)

8.2.2.2 Cleaning incomplete data

Some important ways of avoiding and addressing incomplete data include:

- Clearly define the data to be collected and maintained. For example, suppose you need to create a data repository of students enrolled in higher educational institutions in the world. For each enrollee, mandate the collection of student name, year of enrollment, year of graduation, college name, and college address. Other fields such as “Major” could be specified as needed.
- Collect and cross-validate from multiple independent data sources, e.g., gather information from county, state, national and international organizations, e.g., International Association of Universities, Organization for Economic Cooperation and Development and the World Bank to create as complete a list of educational institutions and enrollees as possible. Gathering data from multiple independent sources provides a more complete picture since one source might have data the other one missed or failed to collect.
- Audit the data for validity and ensure that data are updated. For example, repeat this activity regularly, say every week, and add new students that enroll and new colleges that come into being. Failure to update and keep data current is one of the biggest reasons data are incomplete.

AI is an effective tool for addressing incomplete data. AI agents can continuously poll multiple sources of data for validation, as well as acquire additional contextual information. For example, AI tools can search multiple educational and government databases for enrollees and gather other publicly available information about the college, e.g., the region where the college is located, its financial information, and educational specializations it is known for. Additional attributes and variables generally make the data more informative, as well as provide opportunities for cross-validation, e.g., storing both zip code and state will permit sanity checks

such as “if zip = 07976, then state = NJ”. Such checks could potentially unearth data quality issues in other attributes as well.

Data Debate 8.2.4 *How can you determine if a data set is incomplete?*

8.2.3 Inconsistent data

What if you were to see a fish on a bicycle such as the one in [Figure 8.3](#), right? Or, you check the trajectory of GPS coordinates of your car, and find that it is in Seattle at one moment, and in New York five minutes later? The two preceding examples violate our understanding of the physical world around us. Fish cannot survive outside water, nor do they have the physiology needed to ride a bike. Similarly, given the current state of transportation, it is impossible to travel around 3000 miles between Seattle and New York in under 5 minutes.

Inconsistent data could potentially have disastrous consequences if left unchecked. Consider the following real-world example.

8.2.4 Illustrative example: Orbiter lost due to inconsistent data

In 1999, the Mars Climate Orbiter launched by NASA was lost due to improper data processing that resulted in internally inconsistent data. One of the project teams used imperial units, while another team assumed that the data was in metric units. When the data was passed from one team to another, the conversion was overlooked and the orbiter's thrust was miscalculated causing it to burn up in the atmospheres at a cost of hundreds of millions of taxpayer dollars. This is an example of how an improper understanding and interpretation of data could lead to internally inconsistent data, with catastrophic results. For further details, click this link .

8.2.4.1 Detecting inconsistent data

Inconsistent data issues are often detected by filtering the data through predefined rules and format specifications, e.g., “social security numbers must have 9 digits,” and “measurements of distance and duration should be positive.” These rules are called *constraints*. Data that violates these constraints is considered *inconsistent*. Constraints can be complex, designed to capture the dependence between variables. For example, “if zipcode=07976, then state=NJ.”

Data Debate 8.2.5 *What would be a good constraint to check for improbable cars that can travel 3000 miles in 5 minutes?*

8.2.4.2 Cleaning inconsistent data

Data pipelines that feed AI applications require rigorous monitoring for inconsistent data. Here are a few common-sense checks:

- Ensure that all constraint checks are performed during data processing to detect inconsistencies. Automated constraint checks during data processing are often turned off because they could be computationally expensive and time-consuming.
- When data is updated or new sources are added, make sure that they are consistent with existing data, e.g., attributes, formats, units of measurement, and layout.
- Continuously update time-dependent data. For example, using outdated maps could trip up navigation systems.

AI can help with auditing data systems. While some rules are predefined by experts, many can be discovered using AI tools which excel at detecting inconsistent data. They not only flag data that violate rules and constraints,

but they can also learn constraints from the data itself by using probabilistic reasoning. As a very simple example, if 99% of values of a variable have 9 digits, it will create a rule that says that the format of this variable must have 9 digits, and flag the remaining 1% of values as inconsistent. This reduces the amount of data that potentially needs to be screened by a human.

8.2.5 Anomalies

Anomalies are a particular type of inconsistent data. They are data values that are not aligned with the distribution of the rest of the data. Consider the following example. Suppose that the grades of a student on their first four tests are: A, A, A, D. The student insists that they did well on test No. 4, and that “D” is an unexpected grade that needs to be investigated. The grade “D” in this example is an anomaly because it is inconsistent with historical values. A common explanation could be that the teacher mixed up the rows in the spreadsheet while recording grades, and entered the next student's grade. This is an example of “human error” which occurred while the teacher was trying to combine data from the student roster with the grade sheet. Computers and programmers make such errors too, while combining data from different sources, during *data integration*.

AI is particularly good at detecting data inconsistencies and anomalies in massive data. However, not all anomalies in data are bad. It is important to explore and explain anomalies detected by AI. Sometimes, anomalies can lead to amazing discoveries as demonstrated in the example below.

8.2.5.1 Illustrative example: AI finds a planet far, far away

AI has discovered new planets and stars by identifying anomalies in data collected by NASA's Kepler Space Telescope. In 2017, NASA used AI to

discover a new planet, Kepler-90i, using AI-based anomaly detection by identifying subtle changes in light caused by a planet crossing the star Kepler-90. NASA and Google's AI research team deployed machine learning algorithms to scour massive amounts of data in search of anomalous signals missed by previous studies. You can read more about this exciting discovery in [Section 8.5](#), and by exploring the prompt in [GeminiBox 8.2.4](#).

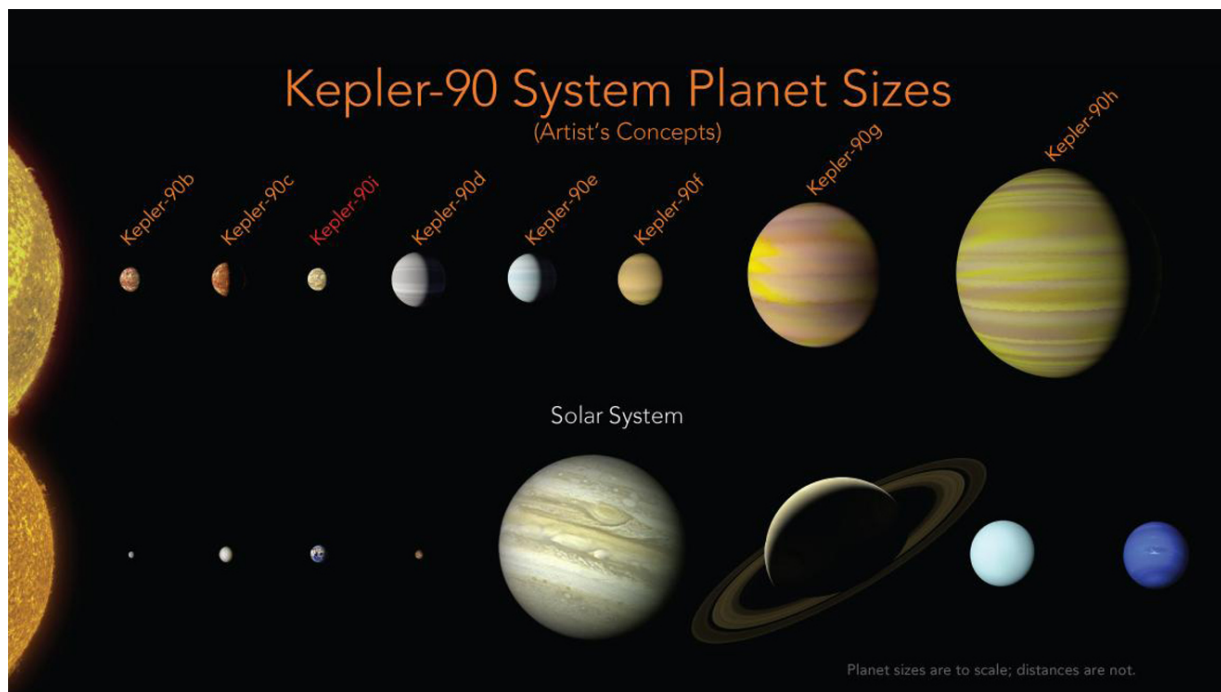


FIGURE 8.8: Kepler-90i, an exoplanet orbiting the star Kepler-90 was discovered by searching for data anomalies in light reading. [NASA Ames Research Center, Wendy Stenzel](#).

GeminiBox 8.2.4: AI discovers new planet by finding anomalies in data [🔗](#)

“How was anomaly detection used to discover Kepler-90i?” [🔗](#)

Surprisingly, AI can be a source of anomalies and inconsistent data due to its tendency to return nonsensical information dubbed *hallucinations*. It has been known to fabricate legal cases, quote non-existent books, and offer dangerous advice. It is critical to vet AI output through cross-validation with authoritative sources of information.

GeminiBox 8.2.5: AI Hallucinations

“What are AI hallucinations?” 

Data Debate 8.2.6 *After reading the output generated by the prompts in [GeminiBox 8.2.5](#), debate the causes of hallucinations and how to fix them.*

8.2.5.2 Cleaning anomalies

Anomalies and outliers are tricky to remediate. Sometimes, they are genuine errors, e.g., speech-to-text conversions while using hands-free devices. The website “Last Word or Nothing”  lists some bloopers.

THEY SAID, AI HEARD:

He's a research scientist postdoc. = He's a research
scientist narwhal.

I'm a data scientist turned entrepreneur. = I'm a dinosaur
sharpener.

These are genuine anomalies and should be fixed. However, anomalies can lead to important discoveries as described in [Illustrative Example 8.2.5.1](#). Here are some ways of addressing anomalies:

- Choose a cleaning method based on how the data is to be used. Anomaly treatment is often contextual. Some applications rely on anomalies, e.g.,

network intrusion uses anomalies to identify suspicious network usage patterns. Others that involve calculating average values, e.g., average test scores of a student, are distorted by anomalies such as one bad test. It might be worth dropping the worst score from the computation of a student's average.

- In some scenarios, one root cause, e.g., bad software, can produce unusual data patterns across multiple systems, e.g., spikes in default date values such as “January 1, 1970.” Finding the root cause can help treat these related anomalies simultaneously rather than individually.
- Analyze anomalies and create potential explanations for future tagging. Consider the *hypothetical* example where a new employee's badge might be flagged as an anomaly until it is approved, which might take a few minutes to hours. This is typically caused when data systems are briefly out of sync. An anomaly explanation might be: “If the employee's name is in the HR database, and the on-boarding timestamp is less than one hour old, ignore the anomaly.”

AI is an amazing anomaly detection tool because of its ability to rapidly scan massive amounts of data. It is used extensively to monitor complex systems such as telecommunications networks, flight patterns, galaxies and gene sequences.

Data Debate 8.2.7 *Find a specific example of the use of AI in healthcare to detect anomalies. Discuss whether this could have been discovered manually by humans looking through the data.*

8.2.6 Duplicates and near-duplicates

Suppose that you go to the grocery store and notice at the check-out that your bill seems larger than expected. You check the receipt and find that you've been charged for 2 soaps and 2 diapers while you only bought one of each, [Table 8.1](#), left. Duplicate scans of the soap and diapers inflated your bill. Had you not checked, you would have been poorer by \$20.

TABLE 8.1: Duplicates can have serious consequences. Left: Grocery bill with duplicate entries. Right: Duplicate photos can clog up devices and cloud storage. Photo: Lynn Greyling. [↩](#)

Item	Name	Quantity	Price
1	Gallon Milk	1	\$6.00
2	Apples	3 lbs	\$4.00
3	Soap	1	\$2.00
4	Soap	1	\$2.00
5	Diapers	1	\$18.00
6	Diapers	1	\$18.00
7	Water-2liter	1	\$3.00



In another example of duplication, consider the case where you receive the same photo on your smartphone from multiple family members. The multiple copies take over your phone camera roll, as in [Figure 8.1](#), right. Most smartphones will offer to “merge” such duplicates.

Duplicates have serious consequences for AI. They increase the representation of the duplicates giving them greater weight in the models that are trained. For example, if the phrase “All work and no play makes Jack a dull boy” was repeated a sufficient number of times in the training text of an AI model, that is all it would learn and ignore all the other text. Duplicates in the census data could lead to misleading population counts and misallocation of school funding.

Near-duplicates are a subtle form of duplicates. For example, the same individual, e.g., John Doe, might be referred to in different databases as Jack Doe, Jacky Doe, Jack Doh, John Doh. In order to realize that they are

the same entity, additional information such as address and social security number might be needed.

8.2.6.1 Cleaning duplicates

The process of removing duplicates is called *de-duping*. There are many approaches to de-duping.

- A simple approach is to sort the data so that identical records appear together. Multiples can be removed leaving one unique record behind. In [Table 8.1](#), left, we would remove the extra “Soap” and “Diapers” transactions leaving just one of each.
- New data can be matched against existing data before being added to the data repository. Any data records that match existing records would be discarded.
- Near-duplicates could be detected by checking additional sources of data to clarify information such as social security numbers. Yet another way would be to profile the data and look for spikes in frequencies of certain values that could point to near- duplicates.

Duplicate detection is a vast field of study. Use the prompt in [GeminiBox 8.2.6](#) to learn more.

GeminiBox 8.2.6: De-duplication

“What is de-duplication? What is an example? How is it done?” 

8.2.7 Other data quality issues

Data quality issues can be of many other types as well. Here are some additional ways data can be of poor quality. There are so many opportunities for data to be corrupted during data processing and preparation stages that it is not possible to list all the ways.

8.2.7.1 Delayed and out-of-order data

Certain applications use data streams where the data arrives in real-time continuously. If the data gets delayed or arrives out of order, the application will have insufficient data to work with. For example, if you are watching a video stream and the images are delayed, the video will seem choppy and freeze frequently affecting the viewing experience. During elections, the results from certain regions arrive sooner than others, making the election results see-saw in favor of one party to the other. At such times, it is important not to declare results prematurely until a reliable portion of the data has been obtained.

8.2.7.2 Irrelevant and outdated data

What if your grade in English was based on your Mathematics test performance? Or your senior-year grade was based on your middle-school GPA? There are many instances when data is hard to come by or too expensive to collect. In such situations, applications are built using whatever data is easily available. It is possible that the data might not be relevant to the problem, or quite outdated. Imagine routing autonomous cars using maps from 20 or 30 years ago. New developments, cul de sacs, new traffic patterns and traffic signs would not be reflected resulting in potentially disastrous outcomes.

8.2.7.3 Data accessibility and interpretability

Suppose there is a database with the best quality data, collected and curated at great expense. But it is on a hard drive and hidden in a remote cave somewhere like in an *Indiana Jones* movie. And, even when you manage to find the data, all you see are a set of numbers. Without a description of the data, i.e., metadata, would you be able to use it? While the data itself might be of good quality, if it cannot be accessed and understood, then it is not usable.

8.3 Measuring Data Quality

Measuring data quality is an essential step in ensuring a supply of high quality data to AI pipelines. A *data quality index (DQI)* provides an objective and universal way to determine the usability of data. It also helps quantify how effective data cleaning techniques are improving the quality of data by moving the index in the appropriate direction. Such an index helps make objective choices between multiple cleaning strategies.

A simple DQI is to compute the proportion of data that has data quality issues. For example, consider the data in [Figure 8.9](#) where each row represents the data pertaining to a phone call. This data is simulated but mimics the format of a real *call detail record*, a standard format used by phone companies. Each record has 9 columns, each corresponding to an attribute. The attributes are described in the header of the table, the first row.

- There are 45 cells in the data sheet. Of these, 2 cells are missing data
- One cell has an anomalous value
- One cell has an out-of-order timestamp

- Three cells have inconsistent values where the data is delivered before the event (the call) is completed
- The last record with 9 data elements is an exact duplicate of the previous one, except for the delivery time

Start Time	End Time	Delivery Time	Cell ID	Cell Tech	Latitude	Longitude	From	To
9/10/2016 5:17	9/10/2016 5:19	9/10/2016 5:50	D000000000001	4G	35.913	-106.248	123456	456789
8/10/2016 9:43	8/10/2016 9:47	9/10/2016 5:50	D000000000001	4G	35.833	-106.259	123457	456788
9/10/2016 10:55	9/10/2016 10:59	9/10/2016 5:51		4G			123457	444
9/10/2016 14:11	9/10/2016 14:15	9/10/2016 5:51	D000000000001	4G	35.854	-106.215	123457	456788
9/10/2016 14:11	9/10/2016 14:15	9/10/2016 5:52	D000000000001	4G	35.854	-106.215	123457	456788

FIGURE 8.9: Example data (not real): There are 45 cells in this table that contain data. Of these, at least 14 cells have data quality issues. [📄](#)

There are 14 cells with bad data. A simple *data quality index*, DQI, would be:

$$\begin{aligned}
 DQI &= \frac{\text{no. of cells with bad data}}{\text{total no. of cells}} \\
 &= \frac{14}{45} \\
 &= 0.31.
 \end{aligned}$$

Note that there are different ways of counting bad data cells. We counted only the delivery timestamp cell as bad data in the second record which was out of order, but we counted all the cells in the last record which was a

duplicate. As long as we are consistent in the way these rules are applied, the DQI is robust.

Data quality is very application-dependent which makes the problem hard to generalize and automate. It is best to define what constitutes bad data and how to count it at the beginning of any project, and ensure that those guidelines are adhered to across the entire length and breadth of the project.

8.4 Statistical Distortion and Data Cleaning

In this chapter, we have discussed several data cleaning methods such as missing value imputation and de-duplication. It is clear, e.g., from [Figure 8.2](#) and [Figure 8.6](#), that data gets altered both by data quality issues as well as data cleaning efforts. It is important not to clean the data to such an extent that it becomes too different from the original data that is supposed to represent some information about the real world. Here are considerations to bear in mind while cleaning data:

- **Quantify the change in the data caused by cleaning:** One way to quantify the change in the data is through a measure called *statistical distortion*, proposed by Dasu and Loh [[13](#)]. Statistical distortion measures the change in the distributional properties of data before and after cleaning, using histogram distances, e.g., Kullback-Leibler distance. The greater the distance, the less faithful the cleaned data is to the original. It is important to choose a *stopping strategy for cleaning* at the beginning of the AI project based on the amount of distortion that can be tolerated by the application before the results become unreliable.

This is akin to specifying the level of significance, α , the probability of Type 1 error, ahead of testing a hypothesis.

- **Keep track of the cost of cleaning:** It is possible sometimes to clean the data perfectly and recover the original data. For example, one could travel to the home of every person named “Doe” or “Doh” to verify the correct spelling of their last name. However, this would entail high costs in both time and money. Resource constraints can often define the scope of cleaning.

Data Debate 8.4.1 *Do you think checking in on every Doe and Doh on earth is feasible? How long would it take to get the final version of the clean data? How much would it cost to travel from home to home? Is it possible to reach every Doe and Doh on the planet? Would you even know of the existence of every Doe and Doh on the planet? Would it be worth it?*

- **Ask if cleaning data helps the goal of the project:** Not all bad data needs to be cleaned. For example, if the project requires calculating typical values such as averages and medians, then missing values that are scattered randomly across the data and within limits, might not distort the results too much. On the other hand, anomalies could potentially distort typical values and might need to be addressed.

Data quality and data cleaning are an important part of data literacy, and they are crucial for building effective and accurate AI tools and models. The book *Exploratory Data Analysis and Data Cleaning* by Dasu and Johnson [[12](#)] is a rigorous introduction to data quality and data cleaning.

8.5 Read & Reflect

- *Why it's time for data-centric artificial intelligence.* 
- *NASA and Google use AI to discover a new planet near a star far, far away.* 
- *Chat bots give bad medical advice if texted to in colloquial English.* 
- *Mars Probe Lost Due to Simple Math Error.* 
- *Users manipulate Microsoft chatbot with bad data.* 
- *India's Ayushman Bharat uses open source AI to manage healthcare data and provide healthcare to remote rural areas and underserved communities.* 
- *How poor data quality impacts AI and machine learning.* 
- *Reinhart, Rogoff... and Herndon: The student who caught out the profs.* 
- *Greece's Economic Crisis and Its Consequences.* 
- *The devil is in the data: How data quality metrics help enterprises get ahead.* 

8.6 Exercises

1. The theater club in a university wants to collect data about students' interests in music and arts using an online survey. What kind of biases

and data quality issues can creep in during this data collection process?
What actions can be taken to address these issues?

2. Can you think of different ways that data can get contaminated? Write down three examples.
3. Write down three ways of cleaning data, and the associated costs in terms of time, money, labor and computing resources.
4. Could you help the detective with the GPS travel trajectory that took a suspect from Seattle to New York City in under 5 minutes? What are the possible explanations? Why?
5. Give three examples of inconsistent data.
6. Examine the output of the prompt in [GeminiBox 8.2.3](#) and list at least one way in which AI can fix missing values in numeric data and in text.
7. Mr. and Mrs. Murthy are planning a 7-day vacation to Dubai. Mr. Murthy asks AI chatbot Gemini to make an itinerary. Is this a good idea? List some pros and cons.
8. A pharmaceutical company has developed a new drug for migraine and is now in the process of testing for effectiveness and safety of the drug. The researchers want to enroll up to 100 people for pre-clinical research who have migraines to try the new drug and sends out a survey to several clinics in the state to gather information to see if the participant will qualify and want to participate. What types of biases and data quality issues should researchers watch out for and how can they minimize data quality issues?
9. **How does data quality impact small businesses?** Baked goods gift baskets: Jay and Kumar have an online business where they create gift

baskets with freshly baked goods for the holidays for shipping across the country. The holidays are coming up and they want to ensure that they are well prepared. They decide to reorganize their customer information, and vendors from whom they order supplies and baked goods. They are asking you, a data quality specialist, to help them with this task. What would you tell them? Do you think AI can help them with this task?

10. Comparison of data collection survey methods and how data quality is affected. In this article, the researchers describe the findings related to data accuracy when data are collected online through two different methods: opt-in and probability based. Read the article from the PEW Research Center on Comparison of Online Survey Samples. 

- (a) Explain what is the opt-in method of data collection and probability-based method of data collection. Provide an example of each.
- (b) After reading this article, what did you learn about these two processes?
- (c) Under what circumstances does it become necessary to collect data by opt-in process?

8.7 Summative Project

In this project, you will examine the sample data sets extracted from the Palmer Penguins data set and address various data quality issues.

1. A data scientist is sorting the data by Island and wants to study the penguins in the Island Dream. This is the sample data set. Identify some

data quality issues with this data set. [Sample penguin data set1](#)

2. A data scientist is provided with the sample penguin data in the table. He has to perform several statistical analysis. Before proceeding, he wants to check the table for data quality, and compute the data quality index (DQI). Examine the data carefully, and identify three data quality issues. What should the data scientist do to mitigate the problems? Explain and calculate the DQI. [Sample penguin data](#)
3. Compare the mean and standard deviation of the body mass of male and female penguins on Biscoe Island. (Hint: choose only the data you need for this analysis). What percent of data had to be removed before performing this analysis? Do you think the results of the analysis will be valid?
4. For the species Chinstrap, the data scientist is examining the flipper lengths of male and female penguins in the sample. What is the DQI after removing bad or incomplete data?
5. The data scientist wants to create a scatterplot of the bill length and bill width of Adelie penguins. How should he proceed? Explain.

9

What's that AI?

AI is rapidly becoming an indispensable tool and transforming all aspects of our life. It has attracted trillions of dollars of investments from governments, technology companies, and start-ups all over the world. We must learn to speak AI in order to use it effectively and call out its misuse.

In this chapter, we provide a high-level overview of AI, its use cases and limitations, and connect the concepts in prior chapters and illustrative examples to specific AI applications. This chapter acts as a segue from the fundamentals of statistics and data science, to introductory concepts in machine learning and AI.

9.1 What's AI, anyway?

AI is simply the use of data and knowledge to train software to learn and perform specific tasks, and more ambitiously, mimic human thinking to automate decisions and actions. Recognizing objects and telling them apart, e.g., cats vs. dogs, planning optimal routes to go from one place to another, screening job and mortgage applications, powering healthcare chatbots, answering consumer queries, and even writing in Shakespeare's style or painting with Michelangelo's flourish, are all within the scope of AI.

Different ways in which these objectives are achieved result in different types of AI.

AI has been around for more than 125 years. It started to coalesce when English mathematician Alan Turing (1912-1954) wondered if machines could think and exhibit “artificial intelligence.” The field was formally defined at a Dartmouth conference organized by Professor John McCarthy in 1956. Since then, the fortunes of AI have gone through cycles of skepticism and hype, until the present, when AI finally seems to be delivering on its promise. [Figure 9.2](#), courtesy Tarjomyar, presents a timeline of the evolution of AI up until 2025. Milestones include the development of special algorithms (back-propagation) in 1986 that allowed neural network algorithms to go back to previous layers and update computations such as weights, and *transformers* in 2017 that enabled parallelization on a massive scale and allowed language models to understand context and generate human-like language. In this chapter, we discuss a simple ontology of AI and how the concepts learned thus far relate to AI. A detailed discussion of AI algorithms is not the intent of this chapter. The classic text *Artificial Intelligence: A Modern Approach* by Russell and Norvig [[37](#)] is considered the authoritative reference on AI.

[Figure 9.1](#) shows two excellent examples of AI. The photo on the left depicts a bionic arm that can be used as a prosthetic for wounded soldiers or others who have lost their limbs to birth defects, disease or accidents. This field of AI is Robotics, which integrates several types of AI. Some key developments in robotics are shown in [Figure 9.2](#), e.g., the musical robot WABOT-2 (1980) that played instruments and the social robot Kismet (2000) that could interact with humans.

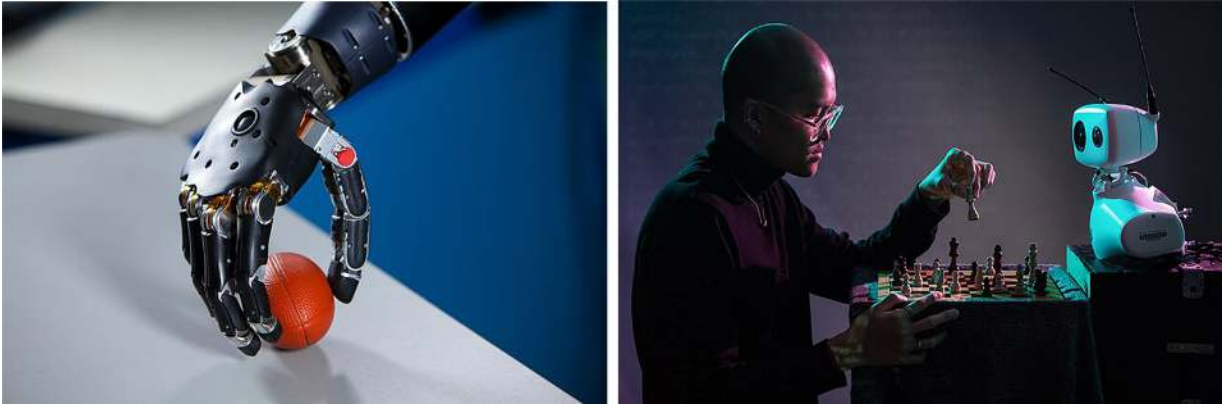


FIGURE 9.1: Left: A bionic arm relies on different types of AI. The Food and Drug Administration (FDA) and Defense Advanced Research Projects Agency (DARPA) funded this brain-controlled, prosthetic arm. (Photo credit: The Johns Hopkins University Applied Physics Laboratory (JHU/APL). Right: AI plays chess and even beats grandmasters at *Go* (Image: Flickr). [🔗](#)

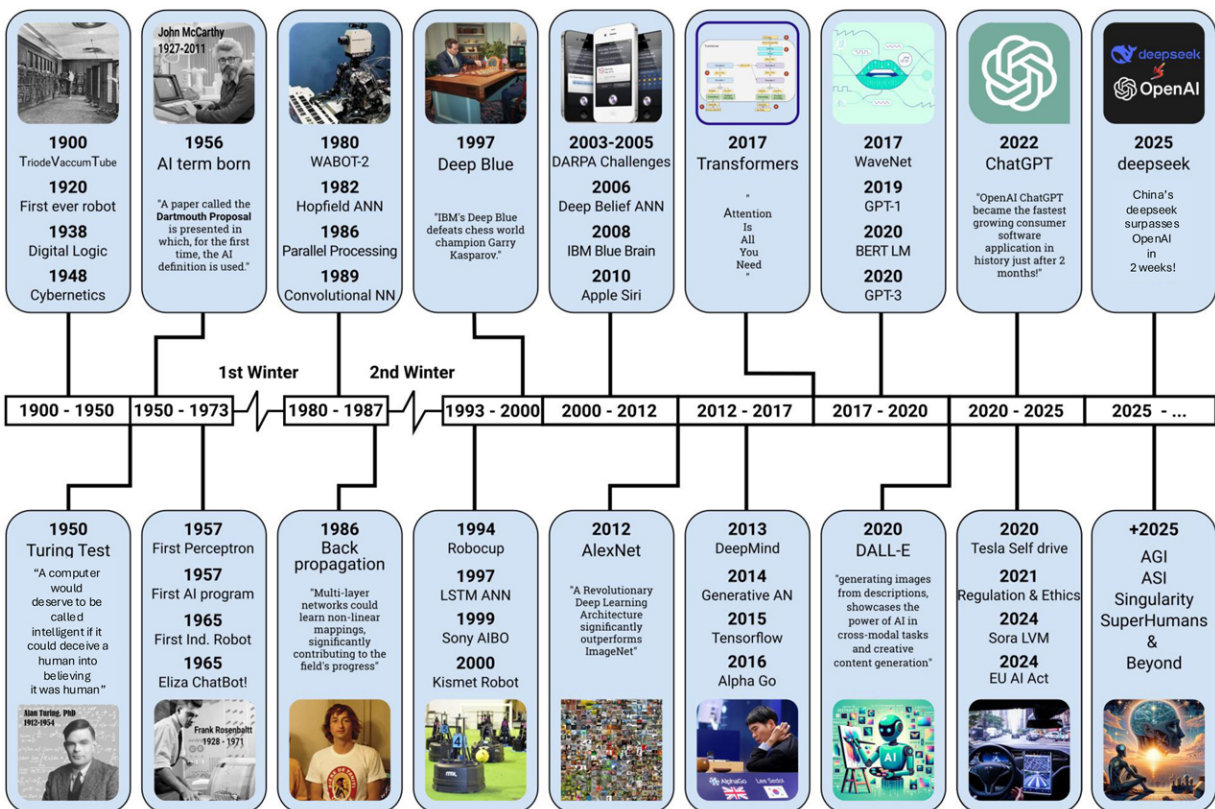


FIGURE 9.2: A timeline of AI evolution. Adapted from the image: Tarjomyar, CC BY-SA 4.0. 

[Figure 9.1](#), right, depicts a robot playing chess. AI has beaten human experts in many games of strategy including chess, Go and Starcraft II, using probabilistic reasoning, optimizing rewards and efficiently searching possible strategies by sampling different paths. In doing so, it developed several unorthodox but highly effective strategies that might be counterintuitive to humans. [Figure 9.2](#) shows IBM's Deep Blue that caused a sensation in 1997 when it defeated the world champion Garry Kasparov at chess. In 2016, Google's DeepMind beat Lee Sedol, the world's best Go player. Watch the documentary [Alpha Go](#).

AI already powers most aspects of our life, from healthcare to fashion, from commerce to sports, as we have seen in the illustrative examples in the previous chapters. However, AI's capabilities have been both exaggerated and demonized. It is neither capable of being some sort of superhuman, nor is it going to take over the world. Given that AI is an important tool in our daily lives, it is critical to understand the basics of AI, in order to use it, to build on it, and to recognize when it makes errors.

9.1.1 AI and what we have learned so far

In previous chapters we learned about data collection and management, exploratory data analysis, probability theory, sampling, hypothesis testing and relationships between multiple variables. These concepts form the foundation of AI and are used extensively in the development and advancement of AI. Here are some examples.

NASA uses adaptive *sampling* to efficiently sample and autonomously explore the surface of Mars for traces of minerals that could have supported

life, as shown in [Figure 9.3](#), left. The article *Here's How AI Is Changing NASA's Mars Rover Science* in [Section 9.6](#) has details. Music and speech are sampled to generate new music and help create synthetic voice assistants.

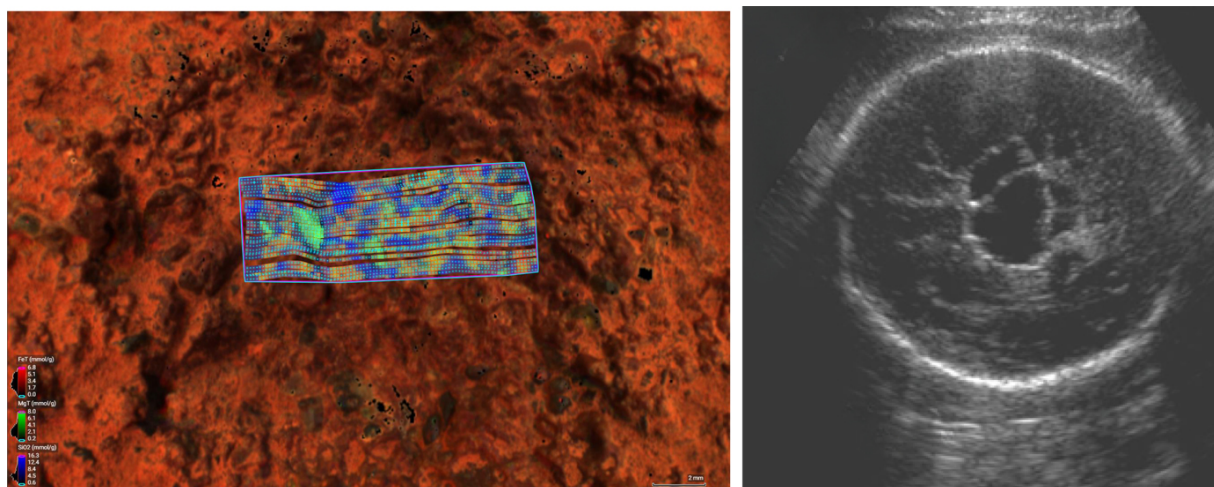


FIGURE 9.3: Left: NASA and JPL use sampling to power AI for exploring Mars. Right: AI uses anomaly detection to identify abnormalities in fetal brain scans. (Photo: Robo Universe.) [↗](#)

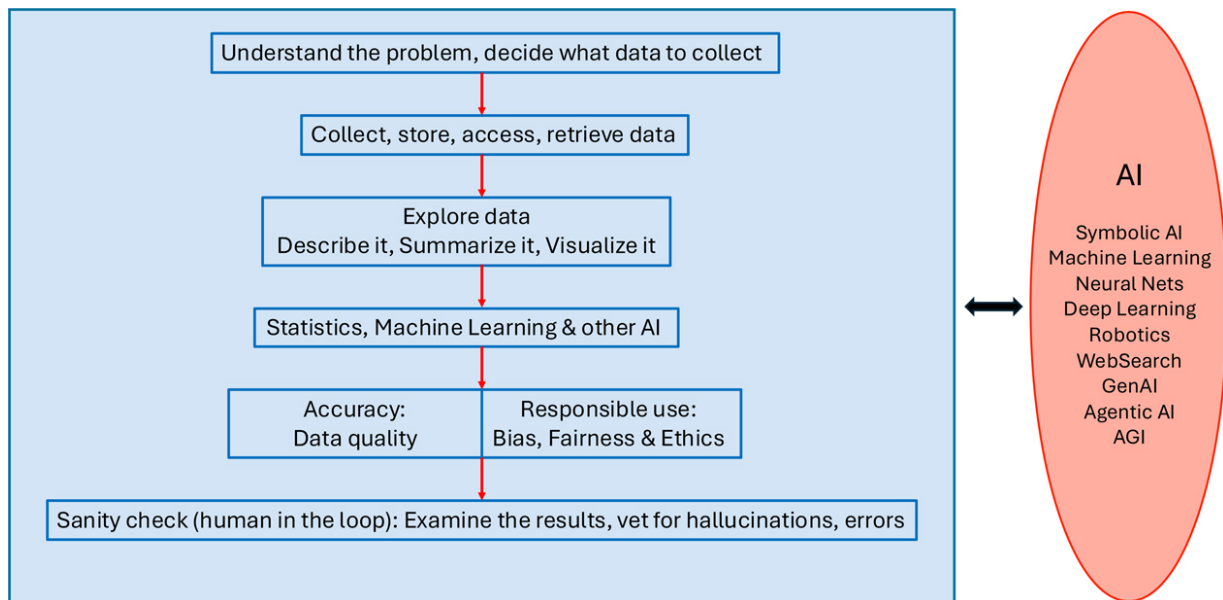
Empirical distributions and *probabilities* are extensively used by AI applications to solve complex problems. For example, the fashion industry uses empirical distributions to manage supply chains and inventory, and for predictive analytics. *Empirical distributions are used in fashion to power AI to spot design hits and misses* in [Section 9.6](#) provides details. Anomaly detection is used to diagnose abnormalities like tumors, fractures and other unexpected changes, as shown in [Figure 9.3](#).

Professional sports teams such the Philadelphia 49ers use AI to generate hypotheses about match-ups, and test hypotheses in real-time to choose the best substitutions that increase the likelihood of winning. In addition, as explained in *Generative AI: 10 Game-Changing Benefits for Sports*

Companies, [Section 9.6](#), their medical staff on the sidelines can diagnose and treat injured athletes faster and with greater accuracy with the help of AI.

Data quality discussed in [Chapter 8](#) is now at the forefront of AI since without data there is no AI. In [Chapter 10](#), we will focus on the responsible and ethical use of data and AI.

During the course of this chapter, it is important to establish the connections between AI and concepts learned in earlier chapters, just like the examples above (see [Figure 9.4](#)). Doing so is good training for courses in machine learning and AI. Therefore, as you read through the applications and types of AI, pause and think of what concepts from the earlier chapters could have been used. For those interested in further study, Russell and Norvig's *Artificial Intelligence: A Modern Approach* [[37](#)] is a comprehensive classic.



[Extended Description](#)

FIGURE 9.4: Data literacy and AI revisited: Concepts from the preceding chapters form the foundation of today's AI which in turn improves the

concepts and automates the processes and tasks discussed in those chapters.



9.2 Illustrative Example: Build an AI classifier with probabilities

To develop an intuition for how statistics and probability are used in AI, let us work through a very simple Naive Bayes Classifier using *Bayes' Rule* and the *multiplicative rule of probabilities of independent events* from [Chapter 5](#).

Suppose that we are given the following data labeled with preference of ice cream flavors. The attributes consist of “Creamy” (Yes/No), “Dairy” (Yes/No), “Sweetness” (High(H)/Medium(M)/Low(L)), and flavor “Type” (Exotic/Standard). We are given the following training data where each record has a “Preference” label (Like/Dislike). Our objective is to predict the preference (Like/Dislike) for a new flavor given some or all of the attributes: Creamy, Dairy, Sweetness and Type.

The Naive Bayes Classifier (at a very high level) involves the following steps:

- Compute the marginal probabilities $P(Y)$ of the *class*, i.e., the Preference label.
- Compute the class-conditional probabilities of each attribute, $P(X_i|Y)$.
- Assume that the attributes are independent of each other *conditional upon the class label*. Therefore, the joint attribute probabilities conditional upon the class label can be multiplied using the multiplicative rule of independence:

$$P(X_1, X_2, \dots, X_k|Y) = \prod_i P(X_i|Y).$$

- When a new *unlabeled* observation arrives, compute the posterior probability using Bayes' Rule:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)},$$

and predict the label by choosing the label with the maximum posterior probability given the attribute pattern of the unlabeled observation. That is, choose the label y such that

$$P(Y = y|X) = \sup_Y P(Y|X).$$

It is not necessary to compute the marginal probabilities $P(X)$ since it is enough to compute relative probabilities to choose the maximum, as demonstrated in the example below.

Consider the simple example where we would wish to predict the label of a new flavor using just one attribute, Type. The computations can easily be extended to multiple attributes using the multiplicative rule for independent class-conditional probabilities of variables mentioned above. Suppose the flavor is Type=Exotic and we wish to predict whether it would be liked or disliked. That is, $Y=Like(L)$ or $Dislike(D)$, given that X variable Type=Exotic. We wish to find the larger of $P(Y = Like|X = Exotic)$ and $P(Y = Dislike|X = Exotic)$. Using the abbreviated expressions:

$$P(Y = L|E) = \frac{P(X = E|Y = L)P(Y = L)}{P(X = E)},$$

and

$$P(Y = D|E) = \frac{P(X = E|Y = D)P(Y = D)}{P(X = E)}.$$

The ratio of the two probabilities is:

$$\frac{P(Y = L|E)}{P(Y = D|E)} = \frac{P(X = E|Y = L)P(Y = L)}{P(X = E|Y = D)P(Y = D)}.$$

Note that $P(X)$ cancels out, and therefore we do not need to compute that.

From [Table 9.1](#), the marginal distribution of Y is:

$$P(Y = L) = 6/10, P(Y = D) = 4/10.$$

TABLE 9.1: Labeled training data for a Naive Bayes Classifier. [↩](#)

Preference Label (Y)	Creamy	Dairy	Sweetness	Type (X)
Like	Y	Y	M	Standard
Dislike	N	Y	L	Standard
Dislike	Y	Y	M	Exotic
Like	Y	Y	M	Exotic
Like	Y	Y	L	Exotic
Like	Y	Y	M	Standard
Dislike	Y	Y	H	Standard
Like	Y	Y	M	Standard
Like	Y	Y	M	Exotic
Dislike	N	N	M	Exotic

The class-conditional probabilities $P(X|Y)$ are:

$$P(X = \textit{Standard}|Y = \textit{Like}) = 3/6$$

(9.1)

$$P(X = Exotic|Y = Like) = 1 - 3/6 = 3/6$$

(9.2)

$$P(X = Exotic|Y = Dislike) = 2/5$$

(9.3)

$$P(X = Standard|Y = Dislike) = 1 - 2/5 = 3/5$$

(9.4)

But we are given that the new flavor type is $X = Exotic$. Therefore:

$$\frac{P(Y = L|X = E)}{P(Y = D|X = E)} = \frac{P(X = E|Y = L)P(Y = L)}{P(X = E|Y = D)P(Y = D)}$$

(9.5)

$$\frac{(3/6) * (6/10)}{(2/5) * (4/10)}$$

(9.6)

$$= 1.875.$$

(9.7)

The probability that the flavor will be liked is 1.875 times the probability that it will not be liked. Therefore, we “predict” that the new flavor will be liked, and label it with the Preference “Like.”

And, *that*, is a Naive Bayes Classifier for predicting whether a new flavor will be liked or disliked based on its Type! This is just a very simple

example of the use of statistics and probability in AI. Next, let us consider how probabilities are used in language modeling.

9.3 Illustrative Example: Large Language Models and Probabilities

Large Language Models (LLMs) generate text. They write emails, summarize legal documents and create specialized itineraries, among other things. The output to the prompts in GeminiBoxes scattered throughout the textbook was generated by an LLM, Gemini. LLMs are one of the major factors that have contributed to the current AI boom, as they are increasingly incorporated into workplace productivity tools. Autocomplete in email is another example of the application of LLMs.

LLMs use probabilities extensively, to reason and generate human-like language. Consider a highly simplified example. The sentence “I want to go to the gym and” could be completed in multiple ways:

- (1) “pay membership fees.”
- (2) “work out.”
- (3) “eat ice cream.”

An LLM trained on vast corpora of text could attach probabilities to each of these, for example, 0.01, 0.987 and 0.003. The phrase with the highest probability would be chosen to predict what comes next. LLMs can predict not just the next word or phrase, but entire paragraphs. The blog *Introduction to Large Language Models*  contains a quick overview of LLMs.

LLMs use the concept of *tokenization* which breaks up language into units such as words and punctuation. Tokens make it easier for algorithms

to process text and associate probabilities in a systematic and structured manner rather than when presented with free-form chunks of text. Tokens such as punctuation also provide contextual understanding leading to better prediction. To quote an old grammar standby, “Let's eat Grandma” has an entirely different meaning than “Let's eat, Grandma.”

An innovation in 2017 (shown in [Figure 9.2](#)), the *transformers* architecture enabled parallelization in contrast to the extant sequential approach to natural language processing. *Self-attention* where computations are limited to those tokens that are immediately relevant to a particular token enabled this parallelization, making language models much faster and feasible for practical use. Transformers were introduced in a ground-breaking paper *Attention is all you need* by Vaswani *et al.* [[43](#)].

LLMs are not perfect. They are prone to hallucinations, and do not have a notion of “understanding” or “truth” or “ethics.” They inherit all the biases and imperfections embedded in the training data, and cannot address scenarios outside the corpora they are trained on. They consume large amounts of energy and might not be kind to the environment. However, they have been adopted rapidly and successfully and are sure to have an outsized impact going forward.

9.4 Types of AI

In the beginning, AI was designed to perform a single task. Given an image, classify it as a cat or dog, for instance. It would be trained on hundreds of thousands of labeled images of cats and dogs, and when shown a new image, identify it as a cat or a dog. Its performance was never at 100% due to the fact that images can be fuzzy, some cats and dogs look similar in

photos, and images in training data could be mislabeled due to human error or otherwise.

As the field advanced, and the ability to collect large amounts of data grew exponentially, AI was trained on more fine-grained data collected every split second. Data from the internet, from atmospheric sensors, from drones, from street cameras, from the stock market, from medical probes, and from every digital trace left by smartphones like cell tower pings, almost everything under the sun could be represented as some type of data. All this data in turn helped AI programs create a more complete picture of the physical world, allowing them to interact with it and perform tasks such as drive a car through crowded streets, or, land a rover on the moon. Quite amazing, right?

The other development that helped advance AI was the increase in computing power. Computer chips, hardware, processors and software evolved to enable companies like Google and Meta to keep pace with the constant torrent of fine-grained data of different types that was available for collecting. Specialized AI chips optimized for AI operations have contributed to a great extent to make the integration of AI applications into practical daily use. Read *AI Chips: What They Are and Why They Matter* in [Section 9.6](#).

AI falls into broad categories based on the conceptual approach to a task. We found [Crash Course AI](#), created by PBS, to be a useful way to think about AI and its organization. We have adapted it in this chapter to meet our specific needs. *The categories are only for guidance since they are all interrelated, with fluid boundaries, and are used in conjunction in most sophisticated AI tasks.* In this chapter, we provide an overview to prepare the reader for further courses in AI and Machine Learning. The foundational concepts of statistics and data intuition described in the

preceding chapters play a critical role in the understanding and development of AI.

9.4.1 Symbolic AI

Human beings have extensive knowledge about the world around them. For example, “if it's raining, carry an umbrella,” or “fish will die if taken out of water.” Such knowledge needs to be communicated to computers and AI, in order for them to perform human tasks and mimic human thinking. However, in the past, computers could not understand a human language, so computer scientists developed specialized languages for that purpose. Human knowledge of the physical world was communicated to computers in the form of rules, represented as a combination of *symbols* and *relations*. Symbols are typically nouns and objects, while relationships are typically verbs and adjectives that connect the symbols. For example,

IsFriend(CharlieBrown, Linus),

is an example of a symbolic relationship where Charlie Brown and Linus, characters from the Peanuts comics, are objects and “IsFriend” is their relationship. *Logic* is used to deduce the validity of the statement and return an answer, True or False. In this case the answer is “True,” since Charlie Brown and Linus are friends.

Data Debate 9.4.1 *Consider the following:*

IsFriend(CharlieBrown, Lucy).

Is this true or false?

To be fair, this is a trick question since it is not clear whether Charlie Brown and Lucy are friends or foes, perhaps both.

Symbolic AI uses such knowledge and rules rather than traditional data. It has evolved over time to incorporate other forms of AI to accomplish complex tasks, e.g., enable chatbots like Siri which can answer open-ended questions rather than just “True or False” type of questions. Siri came into being in 2010, as shown in [Figure 9.2](#), and changed the way humans interact with devices, making them accessible through human conversation rather than a specialized computer programming language. Learn more about how symbolic AI is used by exploring the prompt in [GeminiBox 9.4.1](#).

GeminiBox 9.4.1: Applications of symbolic AI

“What are the most popular applications of symbolic AI?” 

9.4.2 Machine learning

Machine learning is a type of AI that uses data to learn. More data lead to better performance, in general. Data can be of any kind: images, video, webpages, audio, numbers, text.

Data Debate 9.4.2 *What concepts from [Chapter 2](#) could be useful here?*

Machine learning comes in different flavors, some of which are listed below. The book *Elements of Statistical Learning* by Hastie *et al.* [[19](#)] is a comprehensive reference for foundational statistical learning algorithms. Another classic textbook on the topic of machine learning is *Pattern Recognition and Machine Learning* by Christopher Bishop [[3](#)].

- *Unsupervised learning* consists of grouping or *clustering* objects based on similarity of specified attributes. Clustering helps identify subsets of homogeneous objects that are either close together or densely populated,

e.g., galaxies in space, migratory paths of birds, or consumers with similar tastes in movies. In [Chapter 7](#), in the scatterplot of the attributes bill length versus flipper length in the Palmer Penguin data, there were three distinct clusters of points. Each cluster corresponds to a different species, Adelie, Chinstrap or Gentoo, shown in [Figure 9.5](#) indicating that the two attributes together could be used to group penguins into species, and identify or predict the species of a penguin based on those two attributes.

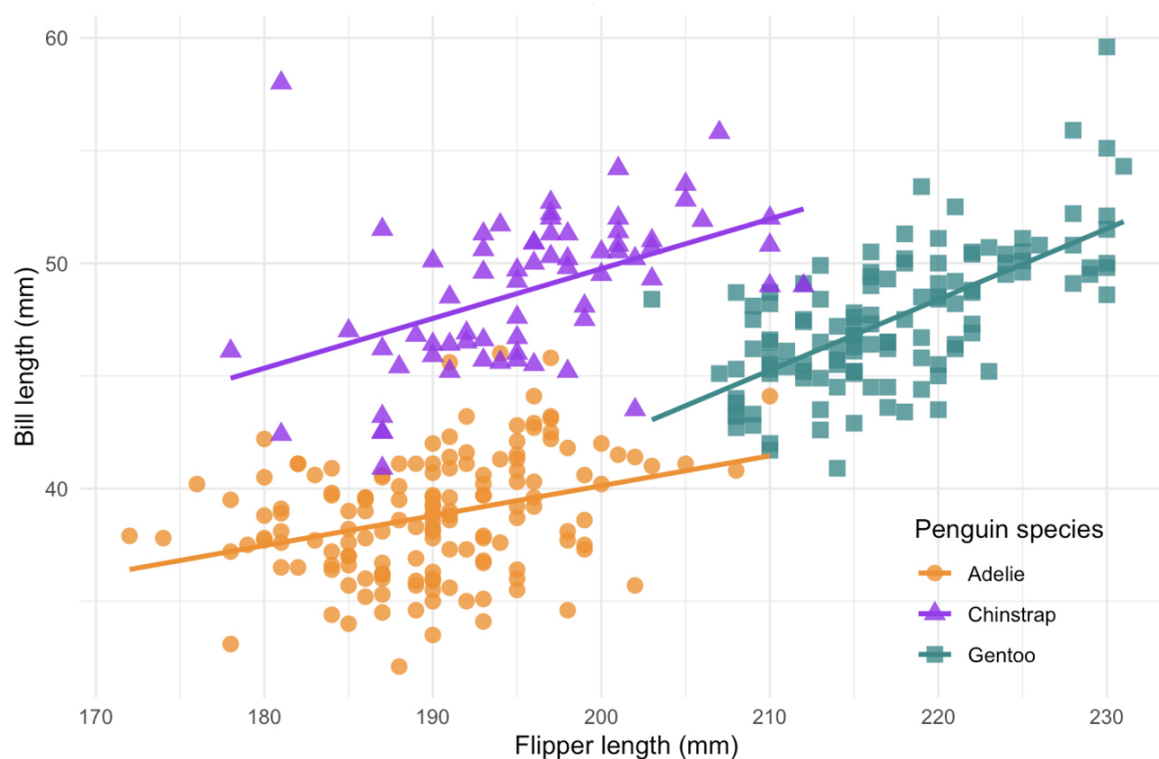


FIGURE 9.5: Unsupervised learning identifies clusters of objects that are similar to each other, such as penguins that belong to the same species. (Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. <https://allisonhorst.github.io/palmerpenguins/>. doi: [10.5281/zenodo.3960218](https://doi.org/10.5281/zenodo.3960218).) ↱

- *Supervised learning* is a predictive task, rather than merely grouping based on similarity. It uses additional information about the objects, namely their “class” labels. Apple or grape? Dog or cat? Horse or zebra? It “trains” on data that already has class labels associated with each object, and learns patterns of attributes for each class. As an oversimplified example, consider [Figure 9.6](#) where the labeled data yields the rule “if color=purple, then label=grape.” What do you think the label of the new object is? What do you think will happen if a new object is a banana? We have already discussed one of the most popular examples of supervised learning, the Naive Bayes Classifier, in [Section 9.2](#).

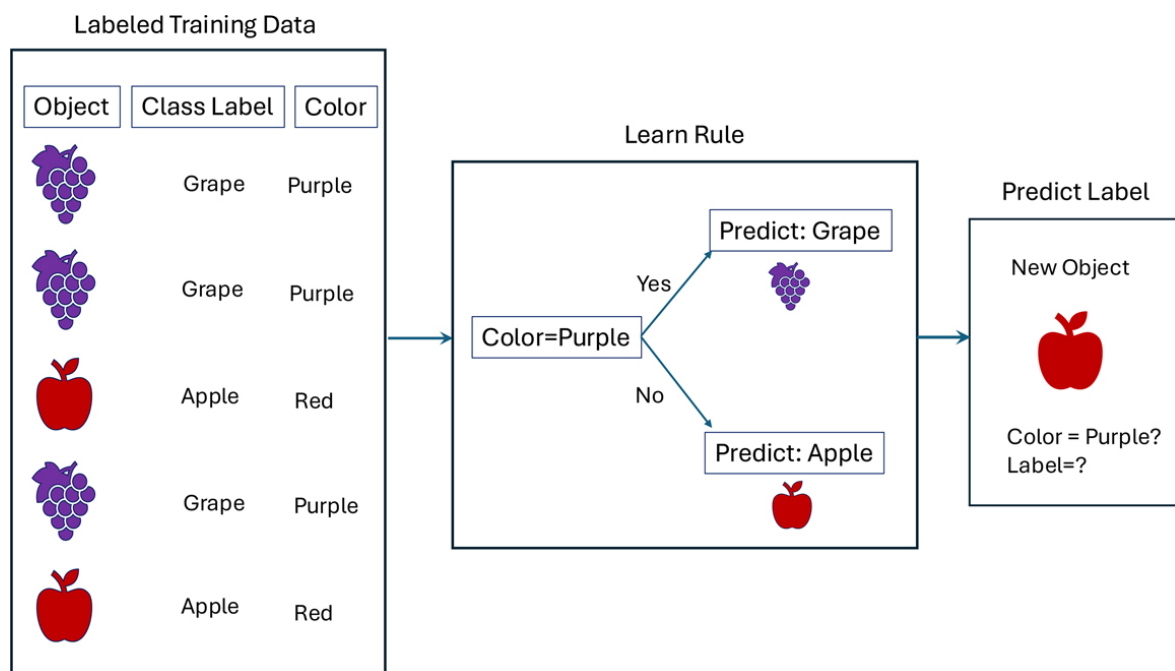


FIGURE 9.6: Highly simplified example of supervised learning based on labeled training data. [📄](#)

- *Neural networks* are complex learning algorithms that mimic biological *neurons*, shown in [Figure 9.7](#). Humans explore and interact with the

word through electric signals collected by neurons and processed by the brain. Neurons combine different types of inputs like sight, hearing, touch, taste and smell (our five senses) collected through *dendrites* to make a decision based on a *threshold of excitation*, and pass on the signal to other neurons through *synapses*. For example, our eyes might perceive fire, our skin might sense heat, and our nose could smell smoke. Depending on the strength of the combined signals, the brain might signal “campfire, get ready to eat marshmallows!” or “wildfire, run for your life!”

Neural nets, as they are known in short, gather data in the shape of different attributes analogous to the brain gathering information from different senses, combine the inputs into a single signal much like dendrites, and check to see if the signal exceeds a particular threshold, to make a yes-no decision. The simplest unit of a neural network is called a *perceptron*. Simple neural networks typically have a single layer that combines individual inputs and outputs a decision.

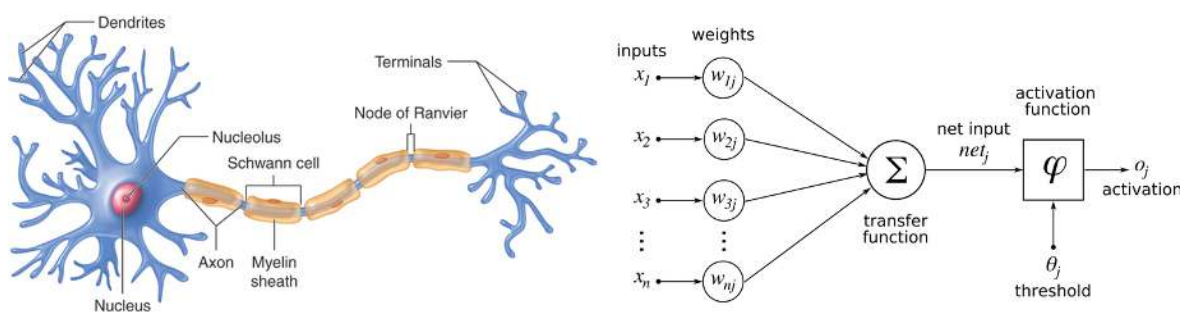


FIGURE 9.7: Left: A neuron accepts inputs through its dendrites, processes them and passes on the signal as an electric impulse to other neurons via synapses. Image courtesy of the City University of New York . Right: Neural networks mimic neurons to gather input, combine them into a signal and make a binary decision depending on

whether the signal exceeds a predefined cut-off threshold. Image: Chrislb, CC BY-SA 3.0, via Wikimedia Commons. [📄](#)

GeminiBox 9.4.2: Perceptron

“What is a perceptron?” 

- *Deep learning* consists of complex combinations of multiple neural networks through layering. Each *layer* results in more and more complex mathematical combinations of inputs. Deep learning networks need massive amounts of data and computing power. The models tend to be opaque and not easily explainable or interpretable by humans. However, they are the key to automation, and the reason that AI can perform complex tasks such as drive cars and clean warehouses. *Deep Learning: Foundations and Concepts* by Bishop and Bishop [4] is a reference text for upperclassmen and graduate students interested in delving further into neural networks and deep learning.
- *Other types of learning* include *reinforcement learning* and *recommender systems* which are specialized types of machine learning algorithms. Reinforcement learning rewards correct answers and penalizes wrong answers, leading to the reinforcement of strongly predictive patterns over multiple iterations. Recommender systems use data from other similar objects, whether our own historical preferences or those of others like us, to recommend shows, clothes, books and music. Recommender systems use relationships between multiple objects, that is across multiple records, rather than just an object and its attributes contained in a single record. Such an approach has computational implications since it might require scanning the data multiple times.

9.4.3 Natural language processing

Voice-enabled AI assistants like *Siri* or *Alexa* need to understand human language, called *natural language understanding* and reply in human language using *natural language generation*. Together, these two aspects constitute *natural language processing*, or NLP for short, which enables humans and computers to communicate with each other like humans. NLP ingests large amounts of text as well as rules that describe syntax and grammar to learn how to process words and word sequences to form sentences, paragraphs and hold conversations in a human language. *Speech and Language Processing* by Jurafsky and Martin [24] is considered to be an authoritative text on this topic.

Large language models, or LLMs, are specialized and advanced versions of NLP models that use deep learning. They are being increasingly deployed in AI assistants to generate and communicate in human-like text. Learn more by using the prompt in [GeminiBox 9.4.3](#).

GeminiBox 9.4.3: Applications of LLMs

“What are three real-world applications of LLMs?” 

Data Debate 9.4.3 *What type of data would NLP systems use?*

9.4.4 Robotics

Robotics involves serious hardware, electronics and engineering, but in order for the robots to accomplish tasks like moving and interacting with the environment, they need to process data from sensors that capture audio, video, and other data about the surroundings, and process the data to make decisions. For example, cleaning robots at Heathrow airport discussed in

[Chapter 5](#) would need to process visual cues from their cameras, recognize the object, and figure out the best action to take, e.g., “move right by 1 meter.” Such robotic functioning is governed by many different types of AI including Object Recognition, Symbolic AI and Deep Learning.

The agricultural robot shown in [Figure 9.8](#), left, helps with pest control by targeted spraying and even performs mechanical weed removal, helping farmers reduce cost and increase efficiency.



FIGURE 9.8: Left: Autonomous pest control in agriculture using robots. Right: Artificial General Intelligence (AGI), romanticized in fiction and movies like *Ex Machina*, is a future aspirational state of AI that can perform all the complex and varied tasks that a human can do, rather than specific tasks, however complex, like the current AIs are able to accomplish. [↗](#)

9.4.5 Generative AI

The GeminiBoxes in this textbook are examples of Generative AI. They generate human-readable answers when prompted in a human language. NLP and large language models mentioned in [Section 9.4.3](#) help Gemini convert human language into machine code, search the web for answers, collate and summarize the information, and return an answer in human

language, with a list of sources that it drew the information from. Imagine how long it would take to find all that information manually!

Any AI that can generate content is called *generative AI*. It is used increasingly by corporations, artists, scientists and many others to create: travel itineraries, emails and executive summaries, pictures, music and videos. Examples include DALL-E shown in [Figure 9.2](#), the picture-generating AI, and of course, Gemini.

GeminiBox 9.4.4: Generative AI

“What's genAI?” 

Another interesting example of genAI are VLMs, or *Vision Language Models*, that can interface between visuals and text. Subtitles, visual descriptions of images and video, are generated using VLMs. Many websites mandate such a description of content in the form of Alt-text, which is often done manually. VLMs could potentially be used to generate Alt-text to help increase the accessibility of online content to those with functional impairments. Learn more about VLMs by clicking on the link in [Section 9.6](#).

9.4.6 Agentic AI

Most AIs discussed thus far follow very specific instructions, e.g., “write a play in the style of Shakespeare”, or “pick up a tennis ball.” However, we have also discussed AIs that can measure the environment around and make decisions in response through reasoning, without relying on human intervention, and learn from the feedback. AI with such capabilities to function autonomously is called *agentic AI*.

Self-driving cars are examples of agentic AI. Agentic AI is also used for monitoring cybersecurity threats, for managing systems and processes, and in healthcare to monitor symptoms of patients and calibrate dosage. Since agentic AI can make autonomous decisions, it can respond rapidly in each of these situations where a timely intervention could make a big difference in outcomes.

9.4.7 AGI

Humans have the ability to perform a mind-boggling variety of tasks. They can see, talk, think, imagine and feel emotions. They can ride a bike, cook, paint, weed the garden, go dancing with friends, calm a crying baby and go swimming. AI, as of now, is capable of doing specific tasks, however complex, extremely well. But in its current state, it cannot accomplish all the general tasks humans undertake as a part of daily routine: brush teeth, take a shower, drive a car, feel happy at the sight of a smiling baby, sing on the way to work, doodle at a boring meeting, create a new genre of music, and cheer for a favorite team. The holy grail of the field of AI is a human-like intelligence that can think, act and create, called *Artificial General Intelligence* or AGI.

While such an AGI does not exist at the moment, movies and shows such as *ExMachina* ([Figure 9.8](#), right) and *Black Mirror* have imagined AGIs that are human-like. There is much controversy about the current capabilities of AI and its ability to benefit or destroy humanity. It is important to filter out all the misinformation and hype, and rely on experts, not quacks or influencers, to be educated and informed by facts, not opinion. [Chapter 10](#) presents an article on the debate between AI experts Yann LeCun and Yoshua Bengio who, along with Geoffrey Hinton, won the prestigious ACM Turing award in 2018 for their contributions to AI.

Data Debate 9.4.4 *Find three examples of AGI in movies and shows.*

9.4.8 Web search

AI now plays a big role now in searching, curating and summarizing results in response to search queries on the web. For example, in response to the query “ancient architecture in india”, the search engine uses AI, to find the relevant webpages in a split second, using a process called *indexing*. The index allows searches to look up information quickly in huge databases where the webpages are stored. The result of this particular search is shown in [Figure 9.9](#). Given that AI has to search through trillions of documents, this is quite a feat. AI also provides a summary of the page, titled *AI Overview*, with links ranked in the order of relevance, and lists the sources that contributed to the summary. It would be impossible for humans to manually search for such information on a regular basis, let alone at this speed. To understand how search engines use AI to find relevant documents and summarize them, use the prompt in [GeminiBox 9.4.5](#).

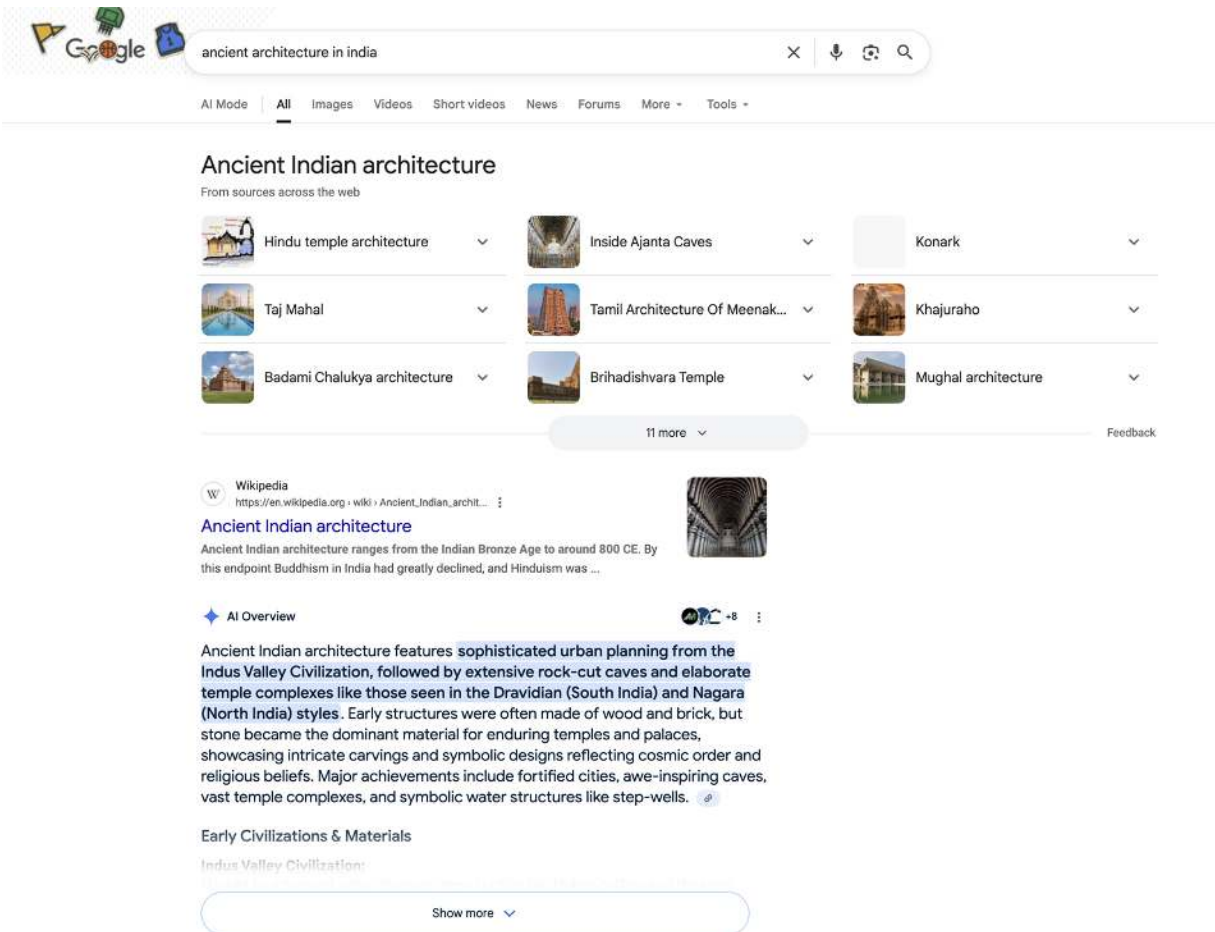


FIGURE 9.9: A screenshot of search results from a Google query. Web search uses AI to locate relevant information in a split second, from trillions of documents online. In this case, when asked for “ancient architecture in india”, the search returns documents as well as generates a summary, “AI overview”, and lists the sources from which it compiled the summary. [↗](#)

GeminiBox 9.4.5: Use of AI in Web search [↗](#)

“Do search engines use AI?” [↗](#)

Data Debate 9.4.5 Examine [Figure 9.9](#) and identify the sources of information for this query. What is the significance of the highlighted text in the AI Overview?

9.5 AI: Power and Responsibility

The famous quote rooted in Spider-man lore “with great power comes great responsibility” is particularly true of AI. As demonstrated in this textbook, AI impacts every aspect of life. From keeping homes at desired temperatures, to plotting best traffic routes, monitoring stock markets, summarizing documents, suggesting shows, playing chess, AI is everywhere. Not only is AI present in every aspect of daily life, it can also have serious consequences for many. AI is used for medical diagnoses, parole decisions, to screen mortgage applications, and to decide whom to hire. It is our responsibility to ensure that AI algorithms are audited so as not to perpetuate unfair practices in society. We will learn more about this in [Chapter 10](#).

The societal impact of AI is still being evaluated as AI tools become increasingly commonplace. The article on Google's AI blog *Applying AI to help solve society's biggest challenges and improve lives*  provides a look at the diversity of ways that AI can affect society, from managing wildfires and floods, protecting coral reefs to helping farm-fresh food reach young children in school cafeterias. The following section presents several specific ways in which AI impacts society, a few cautionary tales about the mistakes AI can make, and the need to prepare communities for AI.

9.5.1 AI and education

AI has the potential to change traditional teaching and learning. AI is an outstanding tool for gathering, collating and summarizing information. It enriches knowledge and information by presenting numerous aspects of a topic. AI has democratized access to knowledge and education by digitizing content, lectures and exercises, and making them accessible to remote,

underserved corners of the world. According to the article *Bridging the Education Divide: How AI Can Support Underserved Communities*, “in Sub-Saharan Africa, about 90% of children cannot read a sentence by age 10, a crisis termed ‘learning poverty’ by the World Bank.” AI can personalize learning at scale and provide tutoring through voice assistants without need for smartphones. Read the entire article in [Section 9.6](#).

Teachers have started to integrate AI as a teaching tool into their classes. AI makes administrative tasks such as lesson plans and evaluation easier. AI has also made it possible to customize lessons and learning to different skill levels. Just as digitization has made it possible to embed links, images and multimedia learning modes, AI opens up new possibilities in terms of drawing knowledge from different disciplines and applications. The potential of AI to provide individual instruction is being tested in schools. However, students must learn the fundamentals of a subject in order to spot inconsistencies, errors and biases in AI's output. AI has its imperfections, like *hallucinations*, but it is also being willfully misused to spread misinformation, and undermine established science. In a report by the National Education Association (NEA), different applications of AI in education are discussed. The report states that “the surge in AI requires that educators become prepared to assess when it is appropriate to use AI, help their students become AI literate, and advocate for the development of policies about this technology.” This NEA report is referenced in [Section 9.6](#).

Finally, it is important to preserve the privacy of both students and teachers by making sure that education and evaluation data remain private and protected. Read more about the pros and cons of AI in the classroom in [Section 9.6](#).

9.5.2 AI and employment

According to the World Economic Forum, “half of employers plan to re-orient their business in response to AI, two-thirds plan to hire talent with specific AI skills, while 40% anticipate reducing their workforce where AI can automate tasks.” AI is expected to result both in an explosion of jobs in the tech sector, as well as a significant decline in jobs that will be replaced by automation. Understandably, there is much public angst about AI's threat to jobs, as indicated in the results of a Pew Research Center poll shown in [Figure 9.10](#).

Workers are more worried than hopeful about future AI use in the workplace

% of employed adults saying they generally feel ____ about how AI may be used in the workplace in the future



Note: Based on all employed adults, including 17% who have not heard about AI use in the workplace and were not asked this question.

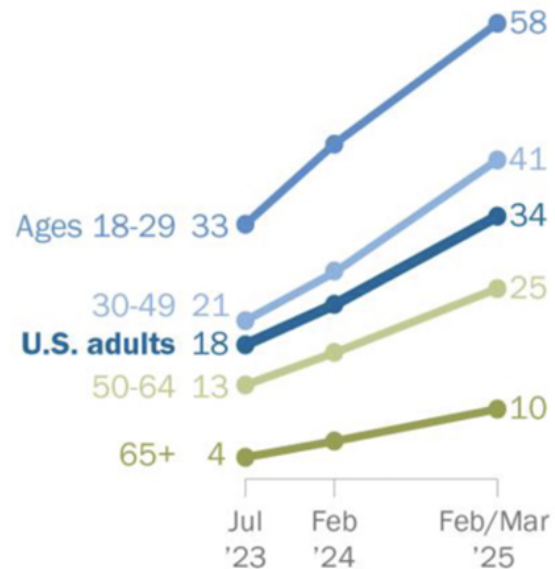
Source: Survey of U.S. workers conducted Oct. 7-13, 2024.

"U.S. Workers Are More Worried Than Hopeful About Future AI Use in the Workplace"

PEW RESEARCH CENTER

ChatGPT use continues to rise; a majority of adults under 30 have used it

% of U.S. adults who say they have ever used ChatGPT



Note: Those who did not give an answer are not shown.

Source: Survey of U.S. adults conducted Feb. 24-March 2, 2025.

PEW RESEARCH CENTER

[Extended Description](#)

FIGURE 9.10: Pew Research Center graphics. Left: How Americans feel about AI use in the workplace. Right: Adoption of AI, specifically ChatGPT, in different age groups. [↗](#)



FIGURE 9.11: Left: Human imagination and creativity resulted in structures like the Seattle Museum of Pop Culture designed by Frank Gehry. (Image: Dietmar Rabich / Wikimedia Commons / “Seattle (WA, USA), Museum of Pop Culture, Fassade – 2022 – 1593” / CC BY-SA 4.0). Right: Classroom at the Hawa Abdi Center in Somalia, a country where thousands of children have been displaced. AI offers personalized learning at scale to help reach remote and underserved communities. 🌀

Like any revolutionary technology that changes the course of history, e.g., invention of the steam engine and motor car, it brings with it inevitable disruption of jobs that no longer can compete with the new technology such as horse-drawn buggies that disappeared after the advent of the motor car. At the same time, widespread deployment of AI could lead to the creation of new job types, such as *AI Auditor* and *AI Ethicist*, speculates *New York Times* opinion and technology writer Robert Capps whose article is referenced in [Section 9.6](#), along with other articles about the effect of AI on the work force.

Data Debate 9.5.1 *Can you think of three new jobs that large-scale adoption of AI could create?*

9.5.3 AI and architecture

AI has been readily adopted by artists and people in other creative professions at the same time raising concerns about authenticity of artwork. Architecture known for its creative geniuses like Frank Gehry and Zaha Hadid, has also benefited from the use of AI. While still in the early stages of adoption, AI can be used to generate architectural designs by providing detailed descriptions. It can also analyze building models for compliance with codes, do a resource assessment, e.g., what type of construction crew would be needed and for how long, scope out the materials, and provide an approximate timeline. According to Professor Phillip Bernstein of Yale School of Architecture, while AI has helped in small ways, it has a long way to go, in terms of taking responsibility for public safety and health, the way architects are licensed to do. Read his interview in the article, *How might AI affect architects? A Yale expert weighs in*, in [Section 9.6](#).

9.5.4 AI and healthcare

AI has many applications in healthcare. It excels at examining images such as CT scans and detecting abnormalities that can indicate fractures or tumors, as in [Figure 9.3](#), right. It can help doctors by screening symptoms and shortlisting possible diagnoses. AI can be used for drug discovery, for generating transcriptions of doctors' dictated notes, for predicting recurrence of disease or the likelihood of missing appointments. AI can manage emergency vehicles, perform routine administrative tasks like refilling prescriptions or filling forms, and streamline administrative processes. In areas with a shortage of healthcare providers, AI can increase the reach of specialists. Robot assistants might be able to help seniors with mobility challenges. Given that healthcare can potentially make the difference between life and death, it is important to make sure that AI is used ethically, responsibly and democratically. [Section 9.6](#) contains several

articles, from the National Center for Supercomputing Applications, World Economic Forum and the National Institutes of Health, about AI and the field of healthcare.

9.5.5 Community and AI

While AI is ubiquitous, it is still a mysterious and fuzzy concept for many. [Figure 9.10](#), right, shows a Pew Research survey about the use of AI tools like ChatGPT. There is a significant disparity based on age. Almost 60% of young people use ChatGPT, but only 10% of senior citizens do. Each age group is vulnerable in its own way to the spread of misinformation generated either by AI or bad actors willfully misusing it. Uninformed communities fall prey to such fake content and forward it, causing real-world damage.

Young people have access to social settings like school and the workplace where they might be exposed to AI tools, but senior citizens are often isolated. It is essential to engage every segment of the community in AI education, by holding free classes and seminars at public institutions like schools, libraries and community centers. AI can be a valuable tool when used properly.

Technology companies such as Google are focused on making AI accessible to all communities in the United States of America by setting up non-profit entities e.g., Google's AI Opportunity Fund. 

Community organizations such as libraries also provide the public with information and learning sessions to understand AI. For example, in New York City the organization *Data and Society* has collaborated with the NYC Public Library to present a series *Understanding AI* to the public . It is our collective responsibility to make sure that AI will benefit the communities we live in.

Data Debate 9.5.2 *Can you think of three ways to help your local community learn about and engage with AI?*

9.5.6 AI is not always right

AI is famously prone to making up facts, euphemistically called *hallucinations*, which we have discussed in earlier chapters. It has no notion of real-world truth, only the data it is trained on. In their hurry to be at the forefront of AI innovation, companies have often rushed to create and adopt AI tools.

Aaron Drapkin's article on *AI disasters* recounts some well-known errors  that occurred during 2023-2025. AI created reading lists with fake books that fooled well-known newspapers, sent rude messages to its creators at Anthropic claiming it had a physical body and wore a blue blazer and red tie, and cited fictional precedents in legal briefs. Such problems could occur when training data includes satire, such as articles from *The Onion*. Sometimes, AI tools are deliberately subverted by users as was the case with Microsoft's Tay. In other instances, malevolent parties deliberately spread lies and misinformation that are amplified by AI bots that crawl the web and generate news summaries.

Given the power of AI, it is our responsibility to understand and educate ourselves and our community, to use its power for good, and to spot errors, bias and prejudice. The next and concluding chapter of this book, [Chapter 10](#), provides guidance on such matters.

9.6 Read & Reflect

- *Crash Course AI*, Jabril Ashe, PBS. 

- *Here's How AI Is Changing NASA's Mars Rover Science.* 
- *Generative AI: 10 Game-Changing Benefits for Sports Companies.* 
- *Empirical distributions are used in fashion to power AI to spot design hits and misses.* 
- *AI Chips: What They Are and Why They Matter.* 
- *A brief introduction to neural networks from IBM.* 
- *What are vision language models (VLMs)?* 
- *Bridging the Education Divide: How AI Can Support Underserved Communities.* 
- *NEA Report: The current state of Artificial Intelligence in Education.* 
- *Pros and cons of AI in schools.* 
- *How do Americans feel about AI in the workplace.* 
- *A.I. Might Take Your Job. Here Are 22 New Ones It Could Give You.* 
- *Speedy cancer diagnoses using AI, National Center for Supercomputing Applications.* 
- *World Economic Forum article on how AI is impacting health care globally.* 
- *NIH article on applications of AI in healthcare, advantages and challenges.* 
- *How might AI affect architects? A Yale expert weighs in.* 

9.7 Exercises

1. **Art and AI:** Read this article from [BBC.com](https://www.bbc.com) about AI-generated art. 
(Source: [BBC.com](https://www.bbc.com). Date October 10, 2024. Author Claudia Baxter. AI art: The end of creativity or the start of a new movement?)
 - (a) What do you think about the use of AI in different art mediums such as movies, photographs, and paintings?
 - (b) What are the advantages and disadvantages of using AI to create art?
 - (c) What ethical rules should be considered to protect the originality of artwork and artists?
 - (d) Do you think AI-created art will be as valuable as human-created art? Why?
 - (e) In [Figure 9.12](#), the famous painting by Johannes Vermeer “The Girl with a Pearl Earring” has been digitally up-scaled using AI software. What do you think about this?



FIGURE 9.12: Johannes Vermeer's “Girl with a Pearl Earring.” Compare AI with Iranian intelligence and art. Wikimedia Commons. [🔗](#)



2. **AI models in healthcare** This video from the World Economic Forum describes a new AI technology that is being developed in collaboration with the digital healthcare transformation initiative “Foresight” that can help doctors and improve patient outcomes. [🔗](#) (Source: University College London.)

- (a) In what different areas of healthcare can the AI model Foresight provide support?
- (b) Do you think this will help doctors and patients? How?
- (c) What are some challenges in using such AI models in healthcare?

3. **AI in education:** How has AI impacted you as a student? Are you using AI for completing your school work or for understanding concepts?

Does your instructor use AI or allow students to use AI for completing assignments? In your opinion, what are some benefits and drawbacks of AI in teaching and learning?

4. **AI in everyday life:** List three instances in the past 24 hours where you have used AI in some form. What did you use it for? What types of AI (e.g., LLM, search, agentic, etc.) did you use? Do you find AI helpful in completing everyday tasks?

9.8 Summative Projects

1. **Art and AI:** Research, investigate and explore how AI is impacting various art mediums- such as the music industry, movies, fine art, drama and theater.
 - (a) Choose any one art medium you are interested in (such as movies, music...) and find several articles from reputable sources to read or listen to podcasts and understand what impact AI is having on artists and art in that medium. Write a summary of the influence of AI on this medium.
 - (b) Interview an artist or a person involved in creating or curating art in the medium you have chosen. Ask them questions about how they are incorporating AI into their art/music/etc. Make a list of at least 5 questions you will use for your interview and note down their responses, or you can make a recording of the interview.
2. **Community and AI:** Artificial intelligence is transforming all facets of everyday life. Whether you are aware or not, it is impacting all our lives and is being used by medical professionals, businesses, scientists, and

the government to make all kinds of decisions, and policy. AI is a thinking system and has the potential to help and make life better for humans, but at the same time it can also be exploited to create scams, take advantage of people and spread false information and perpetuate biases. The most vulnerable groups in our society like seniors, or minors have to be educated on how to take precautions, and learn to effectively use AI.

- (a) Investigate if your local community organizations such as library or senior center are providing classes or information sessions on how to use AI for local community members. Interview the instructor/ or attend a session to find out how they are helping the public to understand and make best use of AI. Write a one-page report.
- (b) Interview members of your family such as your grandparents, parents or your teenage cousin to find out what they think of AI and how they use it. Write a one-page report.
- (c) Find out if schools in your community provide AI literacy to students in grades K-12. How are they doing this?
- (d) What are some AI bots and chatbots that are available for free to use? What are the specific functions or tasks that they can perform?
- (e) Explain how AI is used for scams and identity thefts, as well as to spread false information and perpetuate biases.
- (f) AI systems sometimes hallucinate, that is, the output they generate is false and incorrect, and it is presented as factual information. How can the general public protect themselves from such hallucinations or misinformation?

Responsible Use of Data and AI

Data is the foundation of AI and the many AI-powered tools that permeate our lives. Such data has far-reaching consequences, beyond cool apps and sci-fi movies. When used in the context of medical care and judicial decisions, AI could have life-changing consequences. Therefore, it is imperative that data that powers AI applications be created and used responsibly and ethically. In order to accomplish this goal, it is important to understand ways in which data can adversely affect AI applications, and how to detect and remediate improper and unethical uses of data and AI.



FIGURE 10.1: Left: Personal data is stolen and used without consent, violating individuals' *data privacy*. (Image: Wikimedia Commons.) Right: Deliberate spread of fake news and misinformation could have real-life consequences and harm society. (Image: Nijwam Swargiar.)

AI elicits strong responses. Some fear it might be so powerful that it might turn the world into a dystopian wasteland straight out of movies such as *The Matrix* or the *Terminator* series. Others cite its flaws, e.g., hallucinations and error-ridden documents and images, full of typos and inconsistencies, discounting AI as a hyped-up fad of the moment. Two pioneering AI experts, Yann LeCun and Yoshua Bengio, winners of the prestigious Turing Award, have had public debates, where LeCun does not see AI as an existential threat, while Bengio likens it to nuclear weapons. Their back-and-forth on this subject is detailed in the YouTube video *Munk Debate on Artificial Intelligence — Bengio & Tegmark vs. Mitchell & LeCun* and the article *Artificial intelligence could lead to extinction, experts warn* in [Section 10.8](#). The truth, as always, is somewhere in between. AI is indeed a powerful technology, and like any such revolutionary technology, brings large-scale upheaval to society in terms of new ways of living and working. Therefore, it is important for us to learn how to harness the data, science and technology behind AI to speed up progress for the betterment of society, just as the advent of vaccines and antibiotics, and of electricity, cars and computers did.

Given the potential of AI, bias and ethics have come increasingly into focus with the widespread use of AI algorithms for making life-impacting decisions such as approving mortgage applications, recruiting for high-paying jobs, and even criminal convictions. History is replete with examples of bias and prejudice, and these are embedded in documents and data that are used to train AI. All of us have biases, conscious or unconscious. In addition, when data is collected or selected for use, there might be no consideration for the consequences, or, in the case of biased parties, it might be collected to reinforce a point of view. The book *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* by

Kearns and Roth [26] focuses on the intentional design of technology to serve the betterment of humans and society. In order to understand this, let us consider the following simple examples.

- **Stereotypes (bias):** What's the image that pops into your mind when you think of “skateboarder”? What about “housekeeper”? Studies have shown that in image databases, there is a strong gender bias such as “cooking” being associated with women and “skateboarding” being associated with men. Search engines and other social media tend to amplify the lopsided distribution of data and reinforce such biases.
- **Cherry-picking (selective) data:** Corporations cherry-pick data from experimental studies to present their products in the best possible light, e.g., tobacco companies, pharmaceuticals, and even toy makers. Governments, policy groups and other lobbyists take similar approaches to controversial subjects in order to advance their agenda. Many social media platforms selectively show users stories that will appeal to their point of view. This has had an adverse effect on societies by polarizing communities and eliminating common ground, but it increases the engagement of users and ad revenues for social media companies.
- **Misinformation:** Information spreads with great speed on the internet, and gets amplified and entrenched by AI algorithms that “upweight” it based on “likes”. In addition, software platforms have made it easy to forward such items *en masse* before they can be verified or contradicted. Such features make it easy to game AI algorithms for both legitimate as well as malicious reasons, to have an outsized impact on the real-world. [Figure 10.2](#), right, shows an AI-generated fake image of Pope Francis in a white puffer jacket. Such fake videos or photos of celebrities endorsing products, or in compromising situations have become all too common.

- **Data privacy:** Suppose your roommate saw your laptop open and looked through your course grades? Or, hacked your phone and read your private text messages? Or checked your medical records? Many companies do not just snoop and collect data, they also sell it to other vendors. The *Federal Trade Commission (FTC)*, a US government agency set up to protect consumer rights, has issued a report detailing the lax privacy controls that put children and teenagers at risk, of surveillance and online abuse. The graphic in [Figure 10.2](#) provides a high-level overview of how this process works.
- **Data security:** The lack of proper protection and security measures for data repositories that contain private citizens' data have led to major security breaches. From schools to hospitals to banks to the government of the United States, data breaches have exposed sensitive private information such as social security numbers, bank account details, and health records, resulting in identity theft and fraud. The US Department of Health and Human Services maintains a list of data breaches of hospitals and other healthcare-related institutions at its portal. 

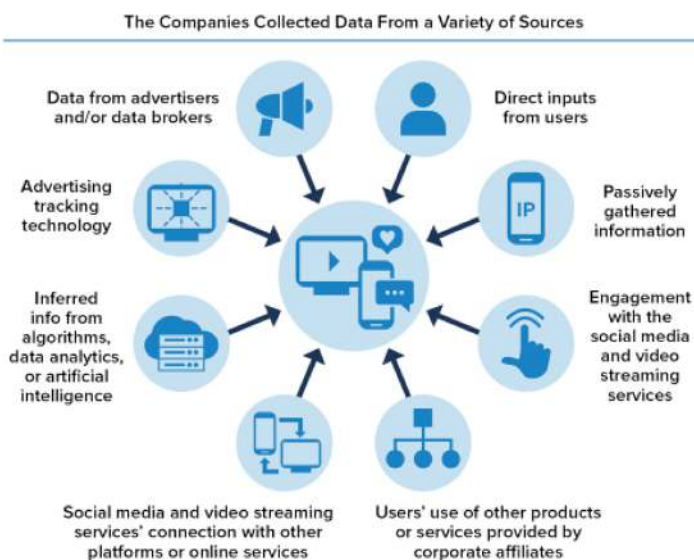


FIGURE 10.2: Left: FTC staff report finds data privacy and data security issues on streaming and social media platforms, issues that could put children and teenagers at risk. Right: AI-generated fake image of Pope Francis in a puffer jacket. Image credit: Midjourney, Public domain, via Wikimedia Commons. [↗](#)

In the following sections, we will discuss different types of bias and ethical issues around data, its use in AI applications, and potential remedies. AI is in a stage of rapid development where governments and regulators are trying to keep up and provide guidance and guard-rails. The technology has tremendous promise and power, and it is important to ensure that it is used in good faith.

10.1 Types of bias

Bias in AI can be deliberate, or unconscious. It can arise from lopsided data as well as the working of the algorithms. Bias can creep in at any stage in the AI process: data collection, storage and retrieval, model building, validation and deployment, and feedback mechanisms. Regardless of intent, bias can have serious consequences. The PBS video *Crash course in AI: Algorithmic bias and fairness* [🔗](#) provides an excellent overview of algorithmic bias in AI. We will explore some ways in which data can introduce bias into AI models and applications.

10.1.1 Bias in training data

What if the data on which AI is trained is not representative? The reasons might not be necessarily nefarious, sometimes it could just be a question of

what data was available or could be collected within the time and resource constraints. For example, AI algorithms in the field of art were historically trained on portraits of aristocrats and nobility since only they could afford to have artists paint their personal portraits. An algorithm trained on such data might not be able to detect features and attire associated with other strata of society, and it either fails to classify them, or classifies them as non-human. Or, the data might have been collected in a different context but reused since it is conveniently and cheaply available. Or, an AI algorithm trained on historical data might reflect outdated practices, e.g., when women were underrepresented in computer science, leading to resumè databases dominated by men's resumès. Since then, women have increasingly entered and thrived in this discipline, but algorithms trained on past data might prioritize male resumès since they have a higher representation and serve as proxies for successful hires, while ignoring resumès that had female-sounding names or activities and hobbies. Here are specific examples of bias in training data:

- *Bias in historical data used for recidivism screening: Pro Publica*, an investigative journal, presented a study that showed that machine learning algorithms used for screening for repeat-offenders had racial bias. The scores it provided for the risk of recidivism did not match up with the real-world data collected by government agencies and others. It is possible that the judgments and criminal records used to train the algorithm were outdated, and no longer reflected the criminality of neighborhoods. Such bias has serious consequences in terms of parole applications which take into account the likelihood of repeating criminal acts once released from prison. See [Section 10.8](#) for a series of articles published by *Pro Publica* in 2016 that made the criminal justice system examine the risk scoring methods used by third-party contractors.

- *Poor representation in facial recognition training data:* The National Institute of Standards and Technology (NIST), a US government agency, found that many facial recognition software tools had difficulty recognizing African American and Asian faces. This was due to a lower representation of such images in the training data which was predominantly Caucasian. Such failure could result in disproportionate scrutiny at airports and other types of screening. [Section 10.8](#) has a link to the NIST story.
- *Unintentional bias in crowd-sourced data:* A popular augmented-reality gaming app, *Pokemon Go*, required players to capture game elements in physical geographical locations. However, the locations showed a bias; although, the origin of the bias was innocuous. The creators of *Pokemon Go* used crowd-sourced geographical data, specifically from players of the game *Ingress*, who were predominantly young, male and lived in affluent neighborhoods. The choice of data collection was mostly convenience. However, there was an unnoticed correlation between affluence and race in the neighborhoods where *Ingress* players lived.

A research study by *Colley et al.* of *Pokemon Go* placement of game elements showed a strong association between prevalence of game elements and non-hispanic white neighborhoods. This geographical database of locations was copied by other game creators, unintentionally perpetuating the bias and depriving less wealthy and more diverse neighborhoods of access to game elements, and the resulting foot traffic to boost to local businesses. A detailed study on *Pokemon Go* is available in [Section 10.8](#).

Data Debate 10.1.1 *What other examples of bias could creep into AI applications? Can bias in data collection be beneficial? When and how?*

10.1.2 Bias in algorithms due to feedback

Algorithms learn from data in the form of responses called feedback, discussed in [Chapter 2](#). If a particular recommendation yields results, the attribute pattern corresponding to that recommendation gets boosted. No wonder you end up with a deluge of baby videos if you happen to “like” just one video of your cute baby cousin on a social media site. Such affirmative feedback can reinforce such patterns in the algorithm and cause it to become increasingly biased in favor of those patterns.

Algorithms that interact with users for feedback can become corrupted quite quickly. For example, *Grok*, an AI chatbot, was soon manipulated by its users to such an extent that it started posting racist and sexist hate speech in a rant filled with lies and misinformation. Details are available in [Section 10.8](#). Since then, it has been cleaned up and is now embedded in the federal government's toolkit. There are several such instances. We have already discussed *Tay*, Microsoft's Twitter bot that was decommissioned after it was derailed in a similar manner by users. In another instance, a *New York Times* reporter, Kevin Roose, was unsettled by OpenAI's chatbot which, during a two-hour conversation with Roose, went on to declare that it had fallen in love with Roose and suggested that Roose leave his wife. Read the entire article on Roose's conversation with Bing's chatbot in [Section 10.8](#). It is possible that the training data included celebrity gossip and romance novels where such incidents and storylines are common.

10.1.3 Bias due to measurement proxies

Bias can enter AI applications at any stage in the AI process. One of the earliest stages is when the decision to collect data is made. Sometimes, it is not possible to measure or collect specific types of data since they are too

abstract or difficult to quantify, as discussed in the video *Crash course in AI: Algorithmic bias and fairness*. 

As an example, suppose you want to measure how delicious a dish is. You can measure nutritional details, e.g., the amount of carbohydrates, salt and calories. Its popularity can be measured by how frequently it is bought. But “deliciousness” is intangible and highly subjective. This disparity between what we want to measure and what we *can* measure could lead to bias, e.g., convenient fast foods that are popular but not necessarily delicious.

Use of proxy measurements to train AI algorithms can have negative consequences. Here are some specific examples:

- Essays in standardized tests in some states are graded by AI software. Since creativity or originality cannot be quantified as a metric, they use proxies like “sentence length, vocabulary, spelling, and subject-verb agreement—the parts of writing that English language learners and other groups are more likely to do differently,” as discussed in the article *Flawed Algorithms Are Grading Millions of Students' Essays* in [Section 10.8](#). This results in nonsensical word-salad essays that get high scores, whereas essays with original ideas expressed in a different way, e.g., shorter sentences, such as those favored by famous writer Raymond Carver, are penalized.
- Many schools are evaluated by the test scores of the students. Under such a high-pressure situation where schools' and teachers' performance is tied to test scores used as proxies for good teaching and learning, the incentive for optimizing the proxy metrics can be high. The focus shifts from students, and teaching becomes aimed at gaming the test scores and even cheating, see *Under Pressure, Teachers Tamper With Tests* in

[Section 10.8](#). This is exactly what happened in a Texas school district in 2010, where the careers of principals and a monetary bonus for fifth-grade teachers were tied to the performance of students on a standardized test.

- An article, *Reliance on metrics is a fundamental challenge for AI*, in [Section 10.8](#) by National Institutes of Health (NIH), a US government agency, mentions that metrics not only are inadequate for measuring the real-world aspect, but can be gamed and incentivized to be measured in highly addictive environments. For example, using the number of clicks or time spent viewing a website as a measure of user satisfaction can lead to unsavory practices that get the users addicted e.g., automatically popping up the next video clip as soon as a viewer finishes watching one, keeping the user glued to the screen. However, in order to do this, the content gets increasingly intense as the level of stimulation needed for the user to stay engaged increases, just like any other addictive substance. This aspect of human psychology is skillfully employed by social media platforms.

Algorithms are increasingly being used to make decisions that profoundly impact our lives. It is our responsibility to stay informed and constantly question how such technology will affect us as individuals, our communities, and the society we live in.

GeminiBox 10.1.1: Bias in AI

“Provide a brief summary of bias and fairness in AI.”

10.2 Misinformation and Fakery

One of the negative impacts of AI has been the explosion of misinformation and fake images and videos. The latter are so realistic that even real videos have been subject to doubt. [Figure 10.3](#), left, demonstrates how Stephen Wolfram, creator of *Mathematica*, can be made to look like Isaac Newton. Such deepfakes can be generated using neural network techniques such as face swapping, as mentioned in *A Few Thoughts about Deep Fakes* . AI is being used to defame and blackmail people. Countries are using them to fight proxy wars, to meddle in elections and to propagate misinformation that can divide and destroy countries. Here are some specific examples:

- On March 18, 2022, Russia used fake videos of the Ukrainian President Zelensky urging his troops to surrender, but was foiled when Zelensky countered with a live feed cautioning his troops and the public to be aware of fake news and videos. The article *Weaponized AI: The New Frontline in Global Disinformation Warfare* in [Section 10.8](#) has more details.
- Ahead of its 2024 elections, India, the largest democracy in the world, set up a Deepfake Analysis Unit as part of its Misinformation Combat Alliance. The public could submit videos through a WhatsApp hotline to verify the authenticity of a video. Their analysis found that: “Packaging misinformation using well-known faces may increase the likelihood of people falling for it, such as featuring people in business in videos promoting financial scams.” The government focused on increasing public awareness by giving helpful tips as to how to detect deepfakes. You can read more about this very interesting effort to fight deepfakes in *Year of elections: Lessons from India's fight against AI-generated misinformation*, in [Section 10.8](#).

- Governments and institutions have stepped up efforts to contain the damage, particularly to individuals and celebrities such as Taylor Swift who was victim of an obnoxious and unconscionable deepfake video posted on the social media platform X. By the time the platform removed it, the damage had been done. After a huge public outcry, the *Take It Down Law* was passed by the US Congress in 2025 and signed into law by President Donald J. Trump. The congressional webpage *Take it Down Act signed into law by President Donald J Trump on May 19, 2025* in [Section 10.8](#) has details.

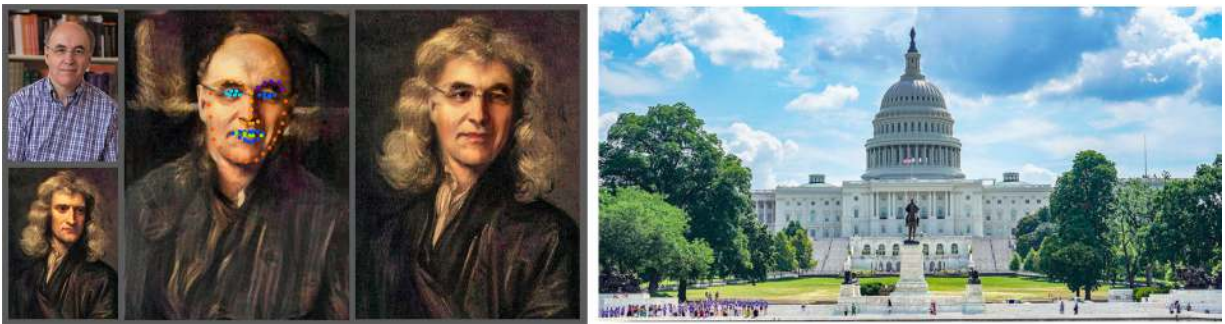


FIGURE 10.3: Left: Creation of a deepfake with neural nets (Stephen Wolfram). Right: Congress passed the *Take It Down Act* in 2025 to fight harmful deepfakes. (Photo: Ted Eytan.) [↩](#)

The articles *5 Telltale Signs That a Photo Is AI-generated* from The Kellogg School of Management, and *How to detect Deep Fakes*, from MIT Labs, in [Section 10.8](#) describe how AI fakes can be screened for inconsistencies such as anatomical anomalies (many fingers), violation of laws of physics, or a weird sheen to the skin. For example, in [Figure 10.2](#), right, the chain holding the holy cross worn by Pope Francis seems to be missing a half. Make sure to explore the prompt in [GeminiBox 10.2.1](#) to figure out how you can help fight misinformation and deepfakes.

GeminiBox 10.2.1: Fighting deepfakes and misinformation

“What is being done to fight deepfakes and misinformation?” 

10.3 Data Privacy and AI

Every time you make a purchase on Amazon, or browse through your social media accounts (Instagram, X, WhatsApp), your personal data is being siphoned off by companies and stored in data centers for AI applications. These digital traces are rich in personal information: entertainment preferences, political affiliations, friends and family, even one's *alma mater*. Many companies do not just collect data, they also sell it to third parties. It is not illegal, but certainly seems unethical, to collect and sell personal information without an individual's consent or even knowledge. [Figure 10.2](#), left, shows the many ways such information is collected.

Such personal information is never quite safe from hackers; it is bought and sold on the *dark web* for identity theft, for committing fraud and even blackmail. Such data breaches have become common including at banks, hospitals, universities, and healthcare companies. The American government maintains a [list](#) of data breaches of “unsecured protected health information affecting 500 or more individuals”, which are under investigation by the US Department of Health and Human Services, Office of Civil Rights.

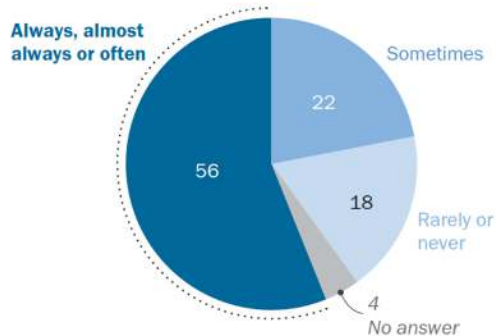
Private data that could potentially harm an individual if made public is often protected by governments. For example, social security numbers, bank account information, academic grades and health data are all considered sensitive and private information. Many countries and some

states in America have strict laws about types of data that need to be protected by law, and secured by the companies that collect and store them. [Section 10.5](#) has details about regulatory measures to help keep private information secure.

In addition, companies are not always transparent about when and what type of data they collect, and how they use it. Even if they are, they usually bury it in long, wordy legal documents that are hard to read and comprehend, crammed in pop-up boxes that require consumers to hit the “I agree” button in order to proceed. Most consumers simply skip reading such declarations, as shown in [Figure 10.4](#), left.

Nearly 6 in 10 Americans frequently skip reading privacy policies

% of U.S. adults who say they ___ agree to online privacy policies right away, without reading what the policies say



Note: Figures may not add up to 100% due to rounding. "No answer" includes those who did not give an answer or who do not use the internet.

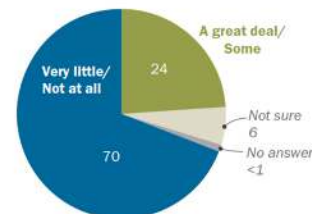
Source: Survey of U.S. adults conducted May 15-21, 2023.

"How Americans View Data Privacy"

PEW RESEARCH CENTER

Most Americans who have heard of AI don't trust companies to use it responsibly and say it will lead to unease and unintended uses

Among those who have heard of artificial intelligence, % who say they trust companies to use AI responsibly ...



Among those who have heard of AI, % who say that as companies use AI to collect and analyze personal information, this information will be used in ways that ...



Note: "Definitely/probably will happen" are combined. Figures may not add up to 100% due to rounding. Those who did not give an answer or who gave other responses are not shown.

Source: Survey of U.S. adults conducted May 15-21, 2023.

"How Americans View Data Privacy"

PEW RESEARCH CENTER

[Extended Description](#)

FIGURE 10.4: How Americans feel about companies using their personal information for AI systems. Retrieved from Pew Research Center. [↗](#)

The Pew Research Center has conducted an extensive [survey](#) on how Americans feel about data privacy. Many have experienced identity theft and fraud. Explore some of the results of the survey illustrated in [Figure 10.4](#) and discuss in class.

Recent AI applications of genAI have been accused of using creative works such as newspaper articles, essays, short stories and novels without permission. This is a clear violation of copyright laws. The *New York Times* has sued OpenAI and Microsoft for training genAI applications on their articles without permission. Read the article *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work* in [Section 10.8](#) to learn more.

Try the prompt in [GeminiBox 10.3.1](#) to understand what governments around the world are doing to protect citizens' private data.

GeminiBox 10.3.1: Privacy and AI

“What are governments around the world doing to ensure privacy of data used by AI?” 

10.4 Ethics in AI

Ethical practices have always been the focus of debate, long before data and AI, particularly in the field of medicine where it can mean exercising painful choices between life and death. For example, a new and important medicine, e.g., for curing a deadly disease, is to be tested by a pharmaceutical company. The company decides to hold clinical trials in poor rural areas of developing countries where the medicine provided by the trial is the only option available to relieve the symptoms of very sick patients. Is this an ethical practice? Probably not, since the patients are

unlikely to reveal negative side effects, fearful that they might lose access to the medication. This results in an under-reporting of harmful side effects that later come to light when the drug becomes available to the public and people start using it on a large scale. Several medicines, for example arthritis pain relieving ones, have been recalled after a significant incidence of gastrointestinal side effects that did not come to light during clinical trials.

The development of AI has been so rapid, and its consequences so significant that governments are finally contemplating regulation. Here are some examples of unethical practices that might motivate governments to take stronger action.

- *Cambridge Analytica*, a data analytics and surveillance company, and Facebook (now Meta) illegally harvested and misused the data of more than 87 million Facebook users for political purposes. Location data, and political preferences based on the news feeds of individuals allowed Cambridge Analytica to psychologically manipulate voters by glorifying one candidate and publishing damaging, often fake, content about the opponent. Subsequently, Meta CEO Mark Zuckerberg was called before Congress, and in 2025 the company was fined \$5 billion for privacy violations and Cambridge Analytica went bankrupt. A lack of accountability and transparency resulted in the abuse of private data and psychological targeting of vulnerable individuals. The story is documented in the article *Facebook-Cambridge Analytica data scandal* in [Section 10.8](#).
- *Houston Astros* and other major league athletic teams have leaned heavily into AI and data analytics. However, stealing pitch signs and passing them to hitters to anticipate the type of pitch, as the Astros have

been accused of, is unethical and goes against the very spirit of competition. Similarly, New England Patriots were accused of filming (a form of data gathering) opposing teams. Yet another team was accused of transmitting information via smartwatches to players to give them an edge. Read Tom Davenport's article on AI and ethics in the sports industry in [Section 10.8](#).

- *Crypto scams* use AI and data to dupe people into giving money. The scams impersonate people online, and use AI to mimic voices of relatives over the phone, pleading for money in an emergency, to be wired immediately. AI is used to create authentic-looking websites and platforms that seem to be investment portals but actually funnel money to hackers. A combination of people and technology has proved to be a deadly and unethical combination that has scammed millions of people to part with their hard-earned life savings. The United States government has just announced a strike force to combat such crypto scammers, *New Scam Center Strike Force Battles Southeast Asian Crypto Investment Fraud Targeting Americans*, featured in [Section 10.8](#).

There are numerous examples of unethical uses of AI. Using AI to cheat on exams and projects in school, to create artificial identities to scam money as in “catfishing” or “pig butchering”, to use the creative works of others and copyrighted material to train AI, and to discriminate against specific groups through biased AI algorithms, e.g., in hiring, policing or lending practices. The next section provides a brief overview of efforts to combat abuse of AI.

10.5 What Is Being Done

AI is in a rapid development stage where its consequences are yet to play out. Regulating such a powerful and complex tool is challenging because not many people, particularly politicians and policymakers, understand the inner workings of AI. In addition, there is a concern that regulating such an explosive technology might throttle its development, putting the United States at a competitive disadvantage. However, rudimentary attempts at regulation are just emerging. Here are some specific efforts that are underway to understand and harness AI and data for the benefit of the public.

- *Government regulation:* The European Union has passed a stringent data protection law, *General Data Protection Regulation (GDPR)*, that protects an individual's right to privacy and control over how their personal data is used.

The United States government and some states have started to create laws to protect the public from the misuse of AI. The Take It Down Law mentioned in [Section 10.2](#) is a good example, where harmful personal information posted without consent is required to be taken down. The State of California has set up a crypto scam tracker  to provide guidance and help those scammed by fraudsters. California has a data privacy law *California Consumer Privacy Act (CCPA)* that protects consumer data from being sold without consent. Other data privacy laws include those that protect medical *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*, children's (*Children's Online Privacy Protection Act (COPPA)*), and financial *Gramm-Leach-Bliley Act (GLBA)* information.

Data Debate 10.5.1 *Use Gemini to explore each of the laws mentioned above, learn what the acronyms stand for, and what the laws state.*

- *Policies and compliance:* International organizations like the Organisation for Economic Co-operation and Development (OECD) have provided guidance in terms of principles that AI should adhere to, such as sustainability, equity, transparency, safety and accountability. The OECD also provides policy recommendations and identifies countries that adhere to these principles. Corporations too have implemented AI governance policies and regularly monitor compliance.
- *Education and outreach:* Governments, private institutions and international organizations have created policies and programs to educate children, teachers and the workforce about AI, its applications and ethical use. The general public, especially older adults, is vulnerable to scams and misinformation.

Community organizations such as local libraries and community centers provide much needed information sessions to educate this segment of our society. Read the blog in [Section 10.8](#) about the program offered by NYC public library.

The now defunct US Department of Education provided guidance for improving educational outcomes  by incorporating AI into K-12 education, training teachers, and offering personalized instruction for differential needs.

Private technology corporations like Google and Nvidia have provided funding and access to software tools to promote AI education. UNESCO (United Nations Educational, Scientific and Cultural Organization) has provided comprehensive guidelines  for AI education of students and teachers centered on an ethical and human-centric approach.

These are just a few efforts that are underway. As the use of AI expands, there will certainly be more initiatives on all three fronts discussed above.

10.6 Our Responsibility as AI Consumers

AI's power can be used responsibly and ethically by exercising caution and building guard-rails to prevent its abuse. Specifically, as individuals, we can take the following steps.

1. *Transparency*: AI applications, e.g., deep learning algorithms, can be quite opaque. The ProPublica article in [Section 10.1.1](#) mentioned that the recidivism risk scores assigned to convicts by the algorithm were never explained, nor was the data on which the algorithm was built shared. The parole applicants have a right to know why they were turned down, at the very least. Transparency in the decision making process would not only address issues of fairness and bias, but make it less prone to abuse.

Open data initiatives that make data available to the public are an important part of transparency. Federal government organizations such as NOAA do an outstanding job of making data accessible.

Another method of increasing transparency and providing visibility into AI systems is to use logs that are generated by AI systems. Such logs, also called AI audit logs, are often used to audit them, to identify what actions were taken, when and by whom.

GeminiBox 10.6.1: AI system logs

“What are AI system logs?” 

2. *Accountability*: Corporations and institutions should be held accountable for the consequences of their AI applications to incentivize them to monitor and manage their applications better. The Cambridge Analytica

example mentioned in [Section 10.4](#) is an example of the worst possible abuse of AI and data that required a congressional hearing to bring it out into the open. Holding individuals and corporations accountable for the abuse of AI should not need such escalation.

3. *Digital Trust*: Trust is an abstract and intangible concept. However, we are accustomed to measuring trustworthiness of companies through organizations like the *Better Business Bureau*, and of products through publications such as *Consumer Reports*. Professional organizations, e.g., the *American Statistical Association*, provide guidelines for the ethical practice of statistics, see [Section 10.8](#). Endorsements from such organizations foster trust in the public.

With the advent of AI, there is a similar push, all across the world, to provide ways to measure *digital trust* in organizations. The World Economic Forum has provided guidelines captured in [Figure 10.5](#), with the goals of security, accountability and inclusiveness, and dimensions to measure the progress towards the goals, shown in [Figure 10.5](#), left.

4. *Auditable Data and AI Protocol*: Many professions that deal with complex scenarios, such as medicine and space exploration, have a protocol checklist that ensures that a process is completed safely and fully. Such checklists are extremely useful for practitioners. The Royal Statistical Society of Britain has created a checklist for the safe practice of data science, described in the article *A Guide for Ethical Data Science* . The European Union commission has created a checklist *ALTAI*, or the *Assessment List for Trustworthy Artificial Intelligence*, for AI applications to engender digital trust, shown in [Figure 10.5](#), right. The checklist consists of principles AI systems should require such as:

- Human agency and oversight

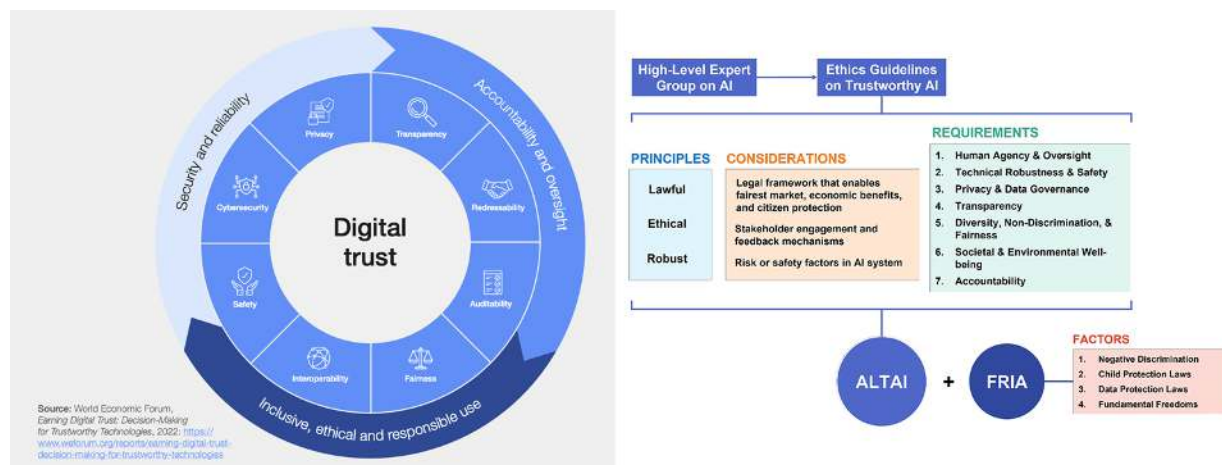
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

We have already discussed many of these concepts. The article *The assessment list for trustworthy artificial intelligence: A review and recommendations* in [Section 10.8](#) discusses the strengths and weaknesses of this list.

5. *Be aware of limitations of AI:* AI has no concept of “truth”, “ethics” or “shame.” It has no emotions or feelings, its goal is to perform whatever task it is prompted to do. If we request a list of ten books, it will provide a list of ten books, some of which might not exist, as we learned in the section on AI hallucinations. As consumers of AI, we need to examine and question the output of AI. In order to perform these sanity checks, we need to educate ourselves about how AI works.
6. *Human in the loop:* We have seen that AI systems, as of now, are prone to errors and can be quite brittle. For example, AI's recommendations and solutions are not consistent. An MIT and Penn State University study  analyzed Ring doorbell videos. AI models were asked: “Is there a crime?” and “Should the police be called?”. The results were inconsistent and were prone to false negatives, indicating that the models were calibrated to being conservative about declaring a crime. Similarly, there was inconsistency in recommending police intervention, which seemed to stem from the demographics of the neighborhood. The

models recommended calling the police more often in minority neighborhoods for the same type of crime which went unflagged in other neighborhoods. The models were also biased in the choice of words to describe people in the videos based on the neighborhood. This is clearly problematic, since such biases get reinforced with the increasing use of such data, algorithms and models.

One way to make AI models more robust is through human oversight called *human in the loop (HIL)*. Humans learn from the time they are born and have vast reserves of knowledge about the real-world that can be used to fine-tune AI and help it learn. HIL can be in the form of simple validation of AI results or more complex interaction through input at different stages. Medical diagnostics is a classic example where AI can screen imaging for abnormalities and a human expert can diagnose it further.



[Extended Description](#)

FIGURE 10.5: Left: Goals and dimensions of digital trust as proposed by the World Economic Forum. Right: A checklist proposed by the European Union for measuring trustworthiness of AI applications. [↗](#)

10.7 Conclusion

Like any revolutionary technology, AI engenders fear as well as unrealistic expectations. Will it take over the world; will it annihilate the human race? Or, will it create an army of faithful robots that will improve the quality of human lives through tasks such as elder care, bomb disposal, scouting for poisonous gases and so on. Will it take away jobs, or will it create a utopia where humans can pursue their dreams assisted by robots? AI technology is powerful but not perfect, and at this point in time, flawed in many ways.

AI is prone to hallucinations. In order to provide *an* answer, AI tends to invent facts in an authoritative and plausible manner. AI amplifies and reflects the flaws in its training data. If there is bias in the data it is trained on, it reinforces such biases. If the data has quality issues, they will affect the results and conclusions of AI models, in ways that might not be easily detectable. In the hands of unscrupulous parties, AI can, and has been, used to cause harm and divisions in society, to pit one group against another, and to create increasingly addictive harmful content. The privacy rights of individuals have been violated by harmful deepfakes. AI has also been used to perpetrate fraud in the form of identity theft, catfishing, and crypto scams.

However, AI and data have the potential to uplift society by expediting the progress of new medical technology like prosthetics and health monitors, by helping doctors refine diagnoses, by speeding up discovery of new life-saving medicines. It can improve infrastructure, find new ways of growing food, and help millions rise out of poverty.

The first step towards embracing this technology is education. This textbook provides a rigorous introduction to the statistics and data behind AI, and it serves as a stepping stone to more advanced understanding of AI and data.

10.8 Read & Reflect

- *Munk Debate on Artificial Intelligence — Bengio & Tegmark vs. Mitchell & LeCun.* 
- *Artificial intelligence could lead to extinction, experts warn.* 
- *FTC Staff Report Finds Large Social Media and Video Streaming Companies Have Engaged in Vast Surveillance of Users with Lax Privacy Controls and Inadequate Safeguards for Kids and Teens.* 
- *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software.* 
- *The geography of Pokemon Go.* 
- *Grok chatbot mirrored X users' "extremist views", xAI says.* 
- *A Conversation With Bing's Chatbot Left Me Deeply Unsettled.* 
- *Flawed Algorithms Are Grading Millions of Students' Essays.* 
- *Under Pressure, Teachers Tamper With Tests.* 
- *Reliance on metrics is a fundamental challenge for AI.* 
- *The New York Public Library and Data & Society present 'Understanding AI.'* 
- *5 telltale signs that an image is an AI fake.* 
- *The Houston Astros and the Ethical use of Data and Analytics.* 
- *New Scam Center Strike Force Battles Southeast Asian Crypto Investment Fraud Targeting Americans.* 

- *Weaponized AI: The New Frontline in Global Disinformation Warfare.* 
- *Year of elections: Lessons from India's fight against AI-generated misinformation.* 
- *Take It Down Act signed into law by President Donald J. Trump on May 19, 2025.* 
- *How to detect Deep Fakes, from MIT Labs.* 
- *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work.* 
- *Facebook-Cambridge Analytica data scandal.* 
- *Explainer: What is digital trust in the intelligent age?* 
- *Deploying Agentic AI - What Worked, What Broke, and What We Learned.* 
- *The assessment list for trustworthy artificial intelligence: A review and recommendations.* 

10.9 Exercises

1. Write an example where data collected about you could help, and where it could harm.
2. Give three examples of bias while collecting data. Are they deliberate or unintentional? What are the consequences? How can one determine if a given data set is biased?

3. Find one example of a country, and a state in the United States that have consumer data protection laws. What do these laws state?
4. Why is data privacy important? What are different ways in which your personal data can be stolen? What are possible consequences?
5. What steps can technology companies that are developing AI take to safeguard against biases in their AI models? Choose a tech company and research how they are addressing this issue.
6. Read the articles: *How to detect Deep Fakes, from MIT labs* and *5 telltale signs that an image is an AI fake* in [Section 10.8](#). Do you think these are helpful to detect a fake image or video generated by AI? What were the three ways to detect fake images that you felt were the best?

10.10 Summative Project

For this research project, find the latest information or news from articles, videos or reports about the following and write a summary. Verify that these articles are from trustworthy sources.

1. Find a news article or report on the ethical use of AI. This can be either a report on misuse of AI or of biases of AI. Who was impacted by this? Write a summary.
2. Find a news article or report that talks about new regulations or safeguards that are being put in place to help prevent scams or fraud using AI. Do you think this is an effective law or regulation? Write a summary.

3. Find a chart or graph that displays information about AI in the context of *Responsible use of Data* and write a brief summary.

Bibliography

- [1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. [↗](#)
- [2] P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number Volumes I-II Package (1st ed.). Chapman & Hall/CRC, 2015. [↗](#)
- [3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007. [↗](#)
- [4] Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Springer, Cham, 2024. [↗](#)
- [5] Johnny Blair, Ronald F. Czaja, and Edward A. Blair. *Designing Surveys: A Guide to Decisions and Procedures*. SAGE Publications, Inc., Los Angeles, third edition, 2013. [↗](#)
- [6] George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, New York, 2nd edition, 2005. [↗](#)
- [7] George Casella and Roger L Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2002. [↗](#)
- [8] C. Chatfield. *The Analysis of Time Series: An Introduction*. CRC Press, Florida, 6th edition, 2004. [↗](#)

- [9] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977. [↵](#)
- [10] D. R. Cox and N. Reid. *The Theory of the Design of Experiments*. Chapman & Hall/CRC, Boca Raton, Florida, 2000. [↵](#)
- [11] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, NY, 1991. [↵](#)
- [12] Tamraparni Dasu and Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc., USA, 1st edition, 2003. [↵](#)
- [13] Tamraparni Dasu and Ji Meng Loh. Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment (PVLDB)*, 5(11):1674–1683, 2012. [↵](#)
- [14] P. Diggle, S.L. Zeger, and K.Y. Liang. *Analysis of Longitudinal Data*. Oxford University Press, New York, N.Y., 1994. [↵](#)
- [15] Bradley Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982. [↵](#)
- [16] Herbert B. Enderton. *Elements of Set Theory*. Academic Press, New York, San Francisco, and London, 1977. [↵](#)
- [17] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton & Company, New York, 2007. [↵](#)
- [18] K. B. Gorman, T. D. Williams, and W. R. Fraser. Ecological sexual dimorphism and environmental variability within a community of

antarctic penguins (genus *pygoscelis*). *PLoS ONE*, 9(3):e90081, 2014. [↵](#)

- [19] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edition, 2009. [↵](#)
- [20] Thomas P. Hettmansperger and Joseph W. McKean. *Robust nonparametric statistical methods*. CRC Press, 2nd edition, 2010. [↵](#)
- [21] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. R package version 0.1.0. [↵](#)
- [22] Peter J. Huber. *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics, 1996. [↵](#)
- [23] Stefan Ivanovic, Ana-Maria Olteanu-Raimond, Sébastien Mustiere, and Thomas Devogele. Detection of outliers in crowdsourced GPS traces. In *Spatial Accuracy 2016 Symposium*, Montpellier, France, July 2016. [↵](#)
- [24] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2025. Online manuscript released August 24, 2025. [↵](#)
- [25] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, 2nd edition, 1975. [↵](#)
- [26] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Inc.,

USA, 2019.[↵](#)

- [27] Michael Lewis. *Moneyball*. WW Norton, 2004.[↵](#)
- [28] Kevin Matzen, Kavita Bala, and Noah Snavely. Streetstyle: Exploring world-wide clothing styles from millions of photos. *CoRR*, abs/1706.01869, 2017.[↵](#)
- [29] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Springer, 2014.[↵](#)
- [30] Douglas C. Montgomery. *Introduction to statistical quality control*. Wiley, New York, NY [u.a.], 3rd edition, 1997.[↵](#)
- [31] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vining. *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, 2006.[↵](#)
- [32] Ronald K. Pearson. *Mining Imperfect Data*. Society for Industrial and Applied Mathematics, 2005.[↵](#)
- [33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.[↵](#)
- [34] Raghu Ramakrishnan and Johannes Gehrke. *Database Management Systems*. McGraw-Hill, Inc., USA, 2nd edition, 2000.[↵](#)
- [35] C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. Wiley, 1973.[↵](#)
- [36] Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 9th edition, 2007.[↵](#)

- [37] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2010. [↵](#)
- [38] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, 1992. [↵](#)
- [39] Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer, Cham, fourth edition, 2017. [↵](#)
- [40] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining*. Pearson, 2018. [↵](#)
- [41] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, second edition, 2001. [↵](#)
- [42] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977. [↵](#)
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. [↵](#)
- [44] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017. [↵](#)

Extended Descriptions for Chapter 1

Extended Description for Figure 1.3

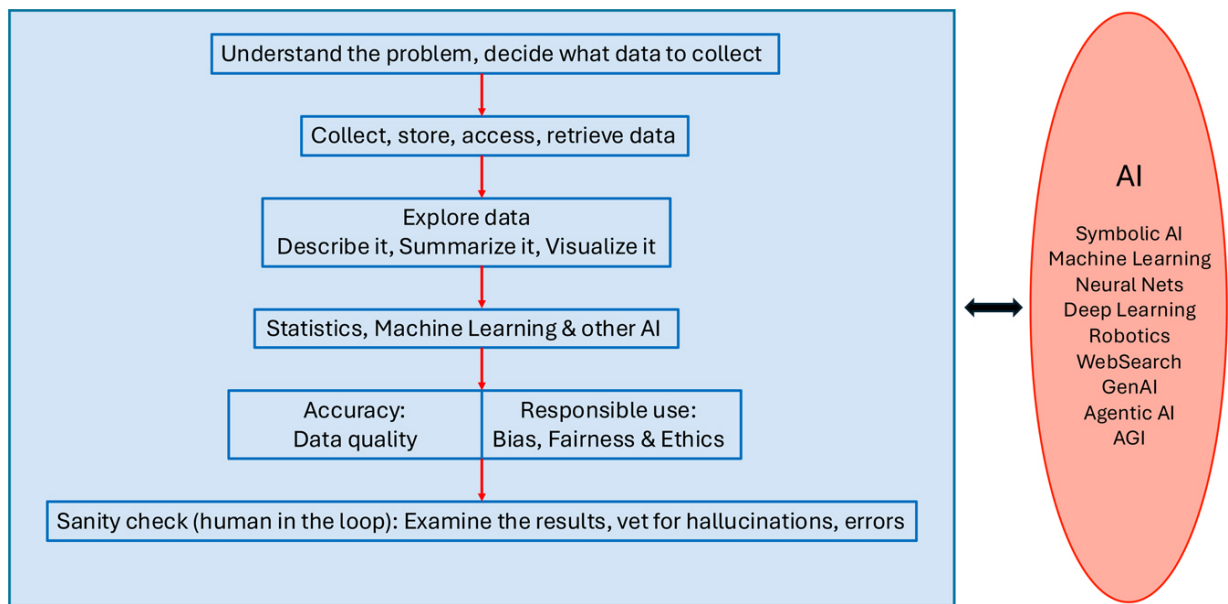


FIGURE 1.3: Statistics, data and AI are interconnected. AI builds on fundamental concepts from statistics and uses data for training, while statistical models gather data in the form of feedback loops from AI to improve performance and advance foundational concepts.

The sequence begins with understanding the problem and deciding what data to collect, then collecting, storing, accessing, and retrieving data. It continues with exploring, describing, summarising, and visualising data, then statistics, machine learning, and other A I. Accuracy links to data quality, while responsible use links to bias, fairness, and ethics. A sanity

check examines results for hallucinations and errors. An adjacent A I list includes symbolic A I, machine learning, neural nets, deep learning, robotics, web search, Gen A I, agentic A I, and A G I.



Extended Descriptions for Chapter 3

Extended Description for Figure 3.1



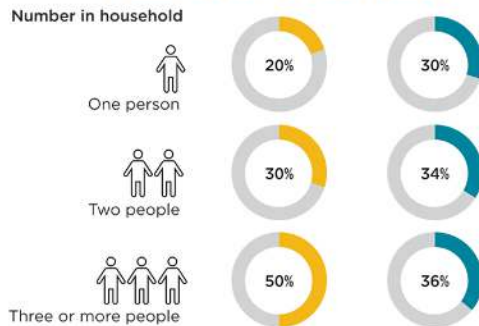
50 Years of the American Housing Survey: Households Then (1973) and Now (2023)



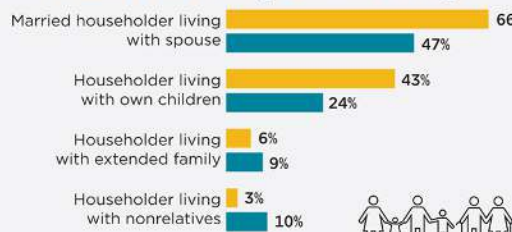
Changes in Householder Race and Ethnicity



Household Size Decreased

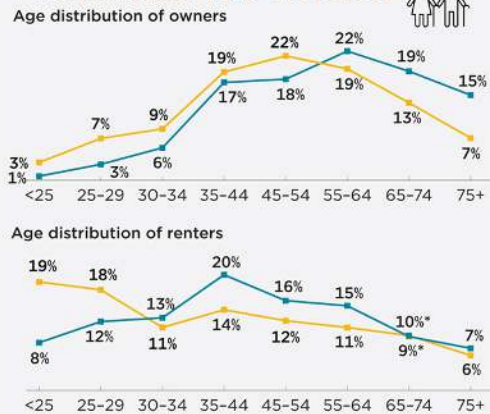


Household Composition Changed



Note: Data on race/ethnicity, marital status, household social characteristics, age, veteran or active military status, and move-in pertain to the householder (the primary owner, renter, or occupant). Household composition categories are not mutually exclusive. More than one category may apply to a unit. "Extended family" means a relative of the householder who is not their spouse or child. When necessary, variables from 2023 and 1973 were recoded for compatibility. Note that in 1973, the military variable reflects householder's service status as of 1970 (three years prior to the survey), whereas in 2023, the military variable reflects householder's current service status. All percentages are rounded to the nearest whole number. All comparisons between 1973 and 2023 are significantly different at the 90 percent level, except for one age comparison. Definitions and more information on confidentiality protection, methodology, and sampling and nonsampling error are available at www.census.gov/ahs. The U.S. Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product (Data Management System [DMS] number: P-7533599; Disclosure Review Board [DRB] approval number: CBDRB-FY25-SEHSD003-066).

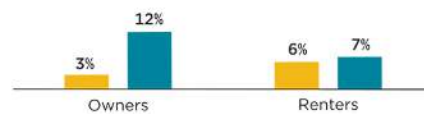
Householder Age Increased



*Not statistically different between 1973 and 2023.

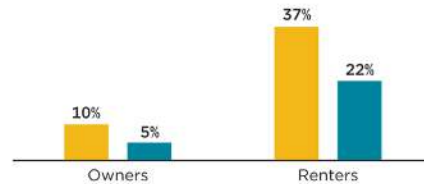
More Veterans and Service Members Became Homeowners

Percent of householders with veteran or active military status



Fewer Owners and Renters Recently Moved

Percent of householders who moved in within last 12 months



United States[®]
Census
Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov



Source: U.S. Census Bureau and U.S. Department of Housing and Urban Development, 1973 and 2023 American Housing Surveys.

FIGURE 3.1: Changes in American households between 1973 and 2023, EDA based on visualization and descriptive statistics. Source: United States Census Bureau.

The content combines bar charts, doughnut charts, line charts, and column charts. Household race and ethnicity categories compare White non-Hispanic, Black non-Hispanic, Hispanic or Latino of any race, and Other or Mixed Race non-Hispanic. Household size compares one person, two people, and three or more people. Age distribution line charts compare owners and renters across age groups from under twenty-five years to seventy-five years and older. Household composition compares married householders living with spouse, householders living with children, householders living with extended family, and householders living with nonrelatives. Column charts compare owners and renters by veteran or active military status and by households that moved within the previous twelve months. Icons represent household members and age groups. A note defines household categories and survey terms. Source and organisational information appear at the bottom.



Extended Description for Figure 3.2

Draw a multi-modal distribution with thick plotting lines

To see the multi-modal distribution plot, you need to be logged in to ChatGPT, as using Python requires that access.

However, you can run the code below on your local machine using any Python environment (like Jupyter Notebook, Google Colab, or even a script):

```
python
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm

# Create x values
x = np.linspace(-10, 20, 1000)

# Create a multi-modal distribution using a mixture of Gaussians
y = (0.4 * norm.pdf(x, loc=-2, scale=1) +
     0.3 * norm.pdf(x, loc=5, scale=0.8) +
     0.3 * norm.pdf(x, loc=12, scale=1.5))

# Plot the distribution with thick lines
plt.figure(figsize=(10, 6))
plt.plot(x, y, linewidth=3, color='blue')
plt.title("Multi-Modal Distribution", fontsize=16)
plt.xlabel("x", fontsize=14)
plt.ylabel("Density", fontsize=14)
plt.grid(True)
plt.tight_layout()
plt.show()
```

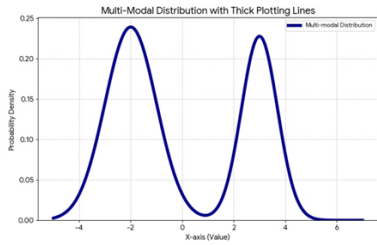
Let me know if you want to add more peaks or change the style of the plot!

...

Draw a multi-modal distribution with thick plotting lines

Show code

Analysis



A **multi-modal distribution** is a probability distribution with two or more distinct peaks (modes). I've drawn a **bimodal distribution**—a common type of multi-modal distribution with two modes—using a thick line as requested.

The graph shows the combined probability density function (PDF) of two normal distributions, resulting in two prominent peaks.

Key Features of the Plot

- **Bimodal Shape:** The plot clearly shows two peaks: one centered around $X = -2$ and another around $X = 3$.
- **Thick Line:** The plotting line is set with a large `linewidth` to make it prominent and thick.
- **Distribution:** The distribution is a mixture of two Gaussian (normal) distributions:
 - **Mode 1:** Centered at a mean (μ) of -2.0 with a standard deviation (σ) of 1.0 .
 - **Mode 2:** Centered at a mean (μ) of 3.0 with a standard deviation (σ) of 0.7 .

...

FIGURE 3.2: Output generated by two AI tools in response to the prompt “Draw a multi-modal distribution with thick plotting lines” in [GeminiBox 3.0.1](#). Left, OpenAI's ChatGPT. Right, Google's Gemini.

The first output gives Python code for a multi-modal distribution with thick plotting lines, including imports, x values, Gaussian mixtures, plotting settings, title, axis labels, grid, layout, and display command. The second output gives a rendered probability density function plot with two peaks, followed by explanatory text and key features describing bimodal shape, thick line setting, and distribution components.



Extended Description for Figure 3.3

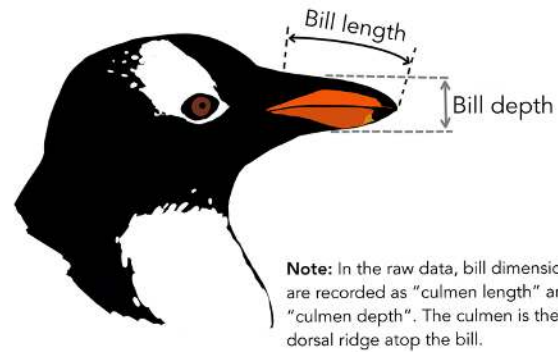


FIGURE 3.3: Left: Gentoo Penguins in Antarctica, walking along a “penguin highway”, a path that joins the sea and their nesting area on a rocky outcrop. (Photo credit: Arturo de Frias Marques. January 2014, [wikiimedia.org](https://commons.wikimedia.org/wiki/File:Gentoo_Penguins_in_Antarctica.jpg).) Right: Measuring bill length and bill depth. (Artwork by Allison Horst.)

The first panel presents five penguins walking along a narrow track across snow. The second panel illustrates a penguin head in profile, with bill length marked from the bill base to the tip and bill depth marked vertically across the bill. A note states that raw bill measurements use culmen length and culmen depth, where the culmen is the dorsal ridge atop the bill.



Extended Description for Figure 3.4

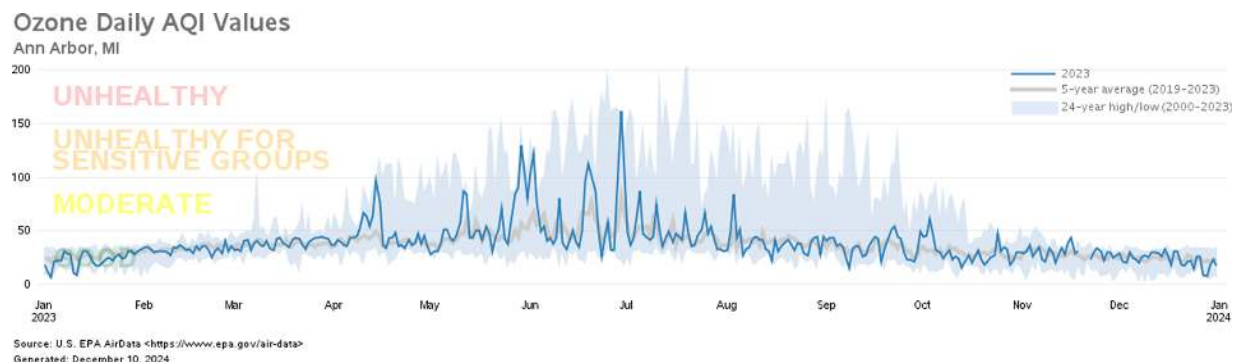


FIGURE 3.4: The United States Environmental Protection Agency collects air quality data and makes it available through its interactive air quality index dashboard. The plot shows ozone variations with respect to time during the year 2023 in Ann Arbor, MI.

Values stay mostly below moderate levels in winter, rise through spring, and peak repeatedly from late spring to mid-summer. The highest spikes occur around June and July. Values decline after summer and remain lower through autumn and winter. Reference bands mark moderate, unhealthy for sensitive groups, and unhealthy categories. The legend compares twenty twenty-three, the five-year average from twenty nineteen to twenty twenty-three, and the twenty-four-year high and low range from two thousand to twenty twenty-three.



Extended Description for Figure 3.5

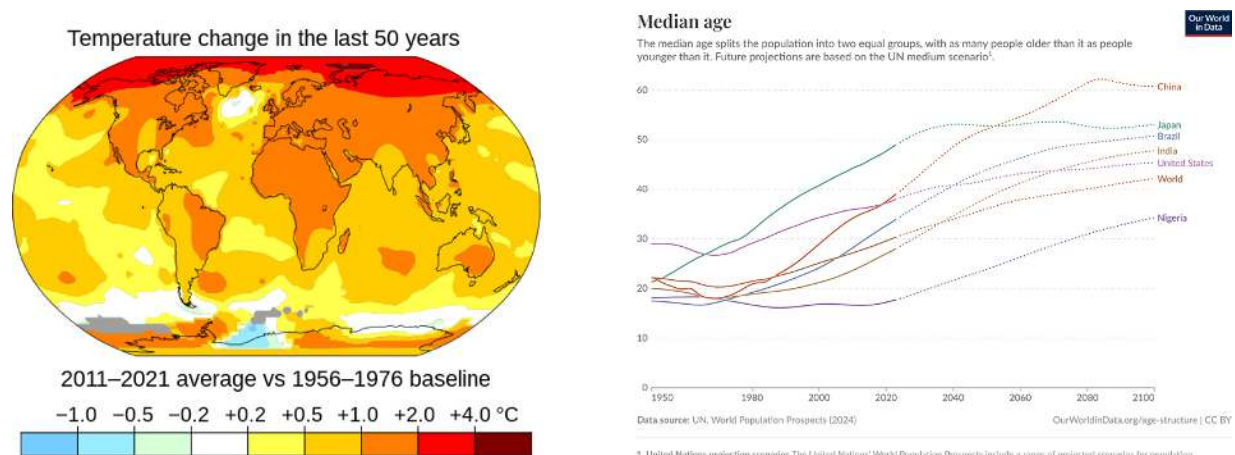


FIGURE 3.5: Left: Daily averages are used in climate studies to track changes (NASA). Right: Median ages are used in demographic studies (Our World in Data).

The map gives temperature change over the last fifty years, comparing the twenty eleven to twenty twenty-one average with the nineteen fifty-six to nineteen seventy-six baseline. The scale ranges from negative one point zero to plus four point zero degrees Celsius, with equivalent values from negative one point eight to plus seven point two degrees Fahrenheit. Most land areas indicate warming, with stronger increases across northern high latitudes. The line chart gives median age from nineteen fifty to twenty one hundred for China, Japan, Brazil, India, the United States, World, and Nigeria. China and Japan rise the highest by late century, while Nigeria stays lowest but increases steadily.



Extended Description for Figure 3.7

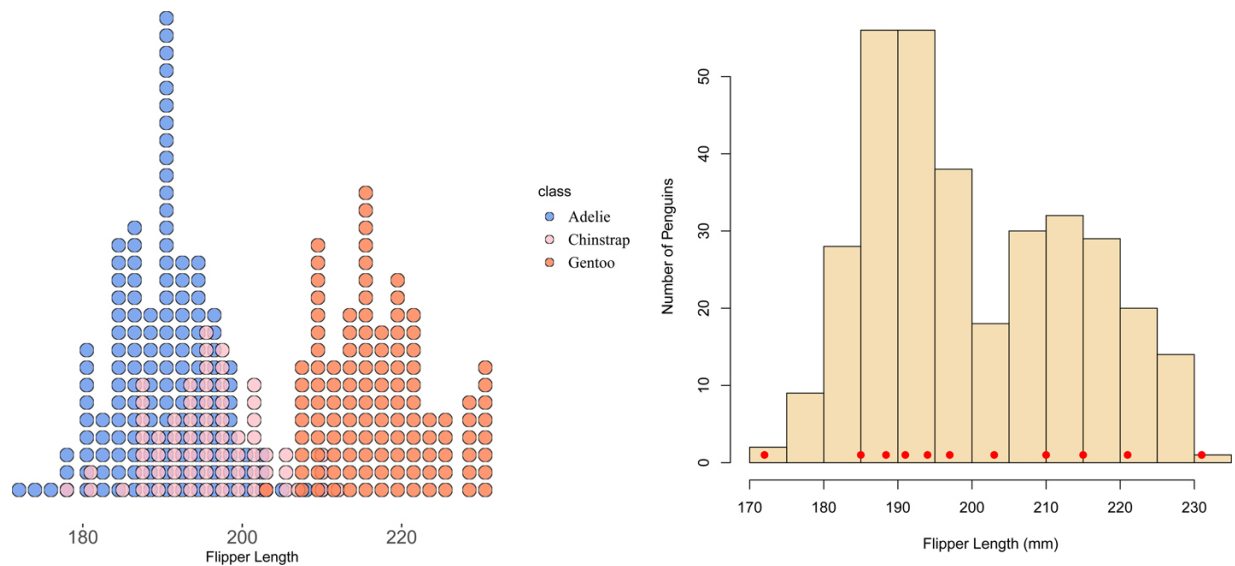


FIGURE 3.7: Data spread for the quantitative variable flipper length, Palmer Penguin data. Left: A dotplot where each dot represents a penguin's flipper length measured on the X -axis. Different colors represent different species. Stacked dots represent penguins with the same or similar flipper

length. Dotplots are discussed in [Section 3.8.5](#). Right: A histogram, discussed in [Section 3.8.3](#), of flipper length, where the red dots are deciles. The figures were drawn using the programming language *R*.

The dot plot groups penguins by species, with Adelie, Chinstrap, and Gentoo listed in the legend. Dots stack at repeated or similar flipper lengths from about one hundred seventy to two hundred thirty millimetres. Adelie and Chinstrap cluster mainly at shorter lengths, while Gentoo clusters at longer lengths. The histogram plots flipper length in millimetres against the number of penguins, with bars forming two main clusters and decile markers along the baseline.



Extended Description for Figure 3.8

Boxplot of Student Heights with Dotplot Overlay

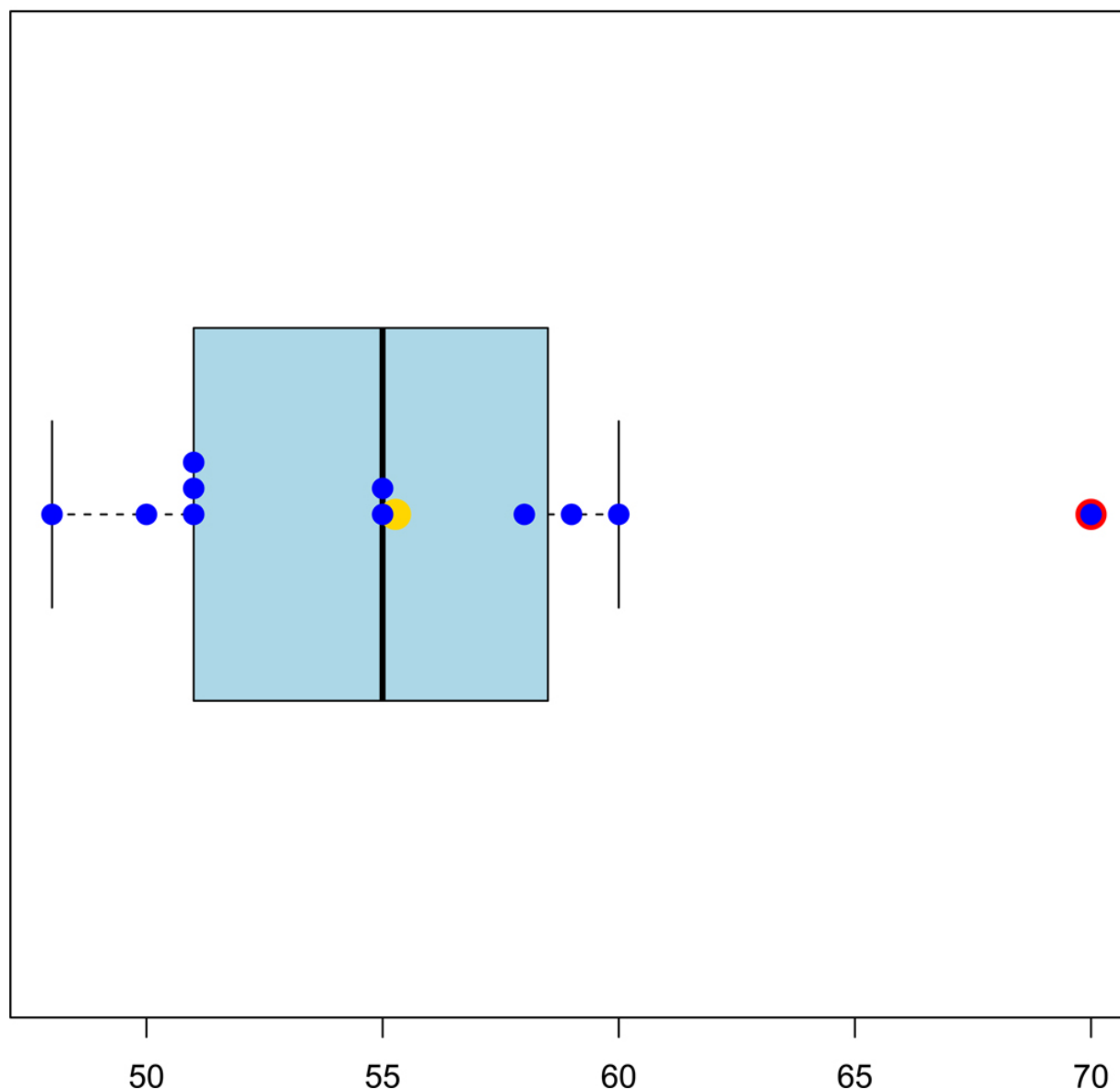


FIGURE 3.8: A horizontal bar plot of student heights where the dark line inside the box is the median, and the left and right edges of the blue box are Q_1 and Q_3 , respectively. The two small vertical lines connected to the blue box with dotted lines represent the whiskers. Any values outside the whiskers, such as the dot circled in red at the extreme right, represent outliers, or extreme values. bar plots are discussed in [Section 3.8.4](#). An

overlaid dotplot shows the 11 heights of individual students. The yellow dot is the mean. The distances of blue dots from the yellow dot (mean) represent deviations from the mean. The figure was drawn using the *R* programming language.

The x-axis ranges from fifty to seventy. The box extends from about fifty-one to fifty-eight, with the median near fifty-five. Whiskers extend to about forty-eight and sixty. Eleven dots mark individual student heights. Several dots cluster around fifty-one and fifty-five. One dot lies beyond the upper whisker near seventy, indicating an extreme value. A separate marker near fifty-five represents the mean.



Extended Descriptions for Chapter 4

Extended Description for Figure 4.2

Do young adults get what they paid for?

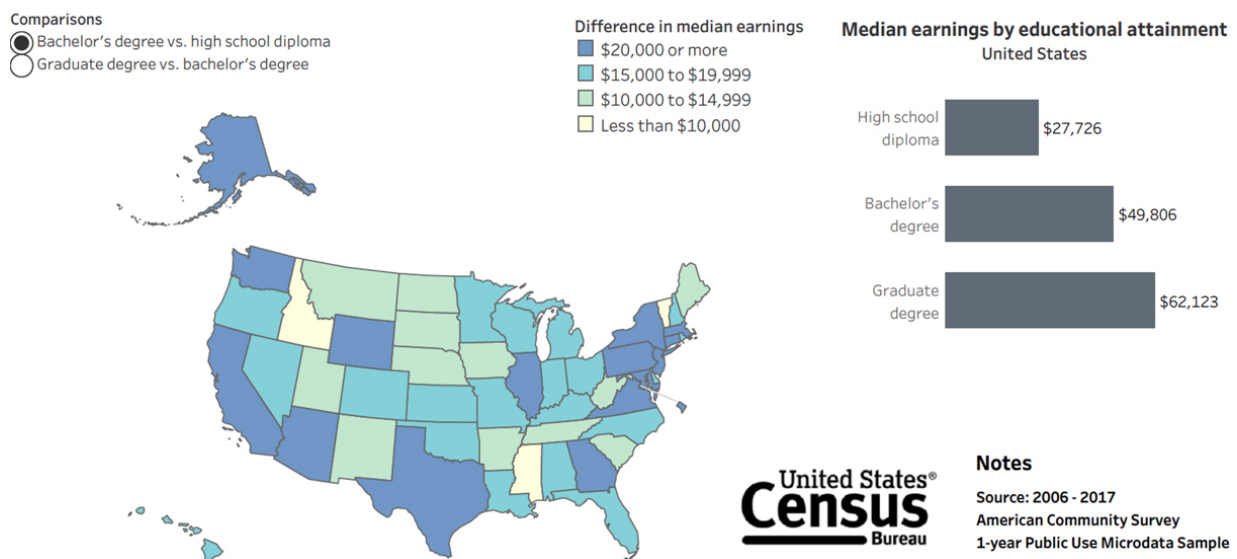


FIGURE 4.2: The United States Census Bureau provides small samples of census data to the public. The above visualization of income disparity based on education attainment is based on a 1% sample of the entire US population (335 million). Click for interactive graphics [🔗](#).

The map compares the difference in median earnings between a bachelor's degree and a high school diploma, or between a graduate degree and a bachelor's degree, across individual states using four earnings difference ranges. Alaska and Hawaii are included as inset maps. The accompanying bar chart compares median earnings for high school diploma, bachelor's

degree, and graduate degree in the United States. Notes identify the data source as the A C S Public Use Microdata Sample for the years two thousand six to two thousand seventeen.



Extended Description for Figure 4.3

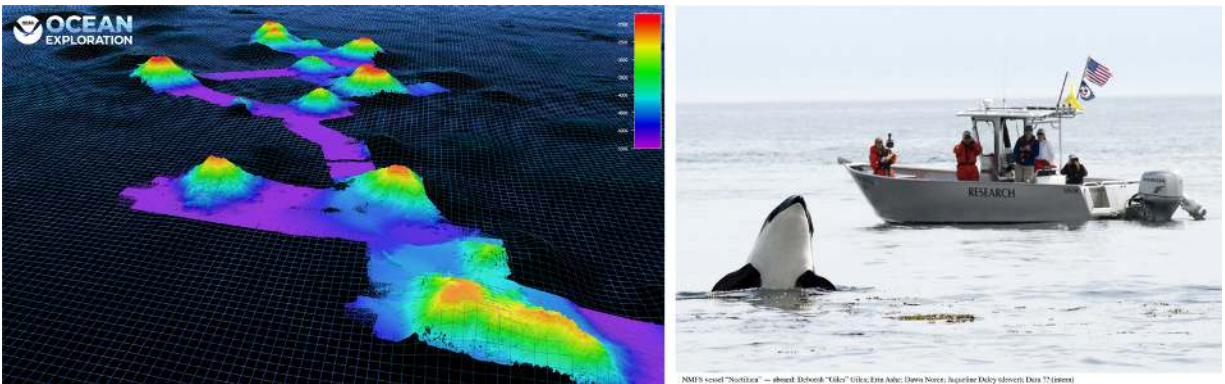


FIGURE 4.3: Left: NOAA's image of bathymetry data from ocean floor mapping through sampling. Right: NOAA Research: Collecting samples from a boat. A pantropical spotted dolphin surfacing alongside the NOAA Ship Gordon Gunter. Credit: NOAA Fisheries (Permit No.14450).

The first panel contains a three-dimensional bathymetry surface with raised underwater features, depth variation, and a scale bar. The second panel records a research vessel on calm water, with several crew members aboard and a dolphin surfacing nearby. Text on the vessel reads research, and a small credit line below the photograph names NOAA fisheries and includes a permit number.



Extended Descriptions for Chapter 5

Extended Description for Figure 5.3

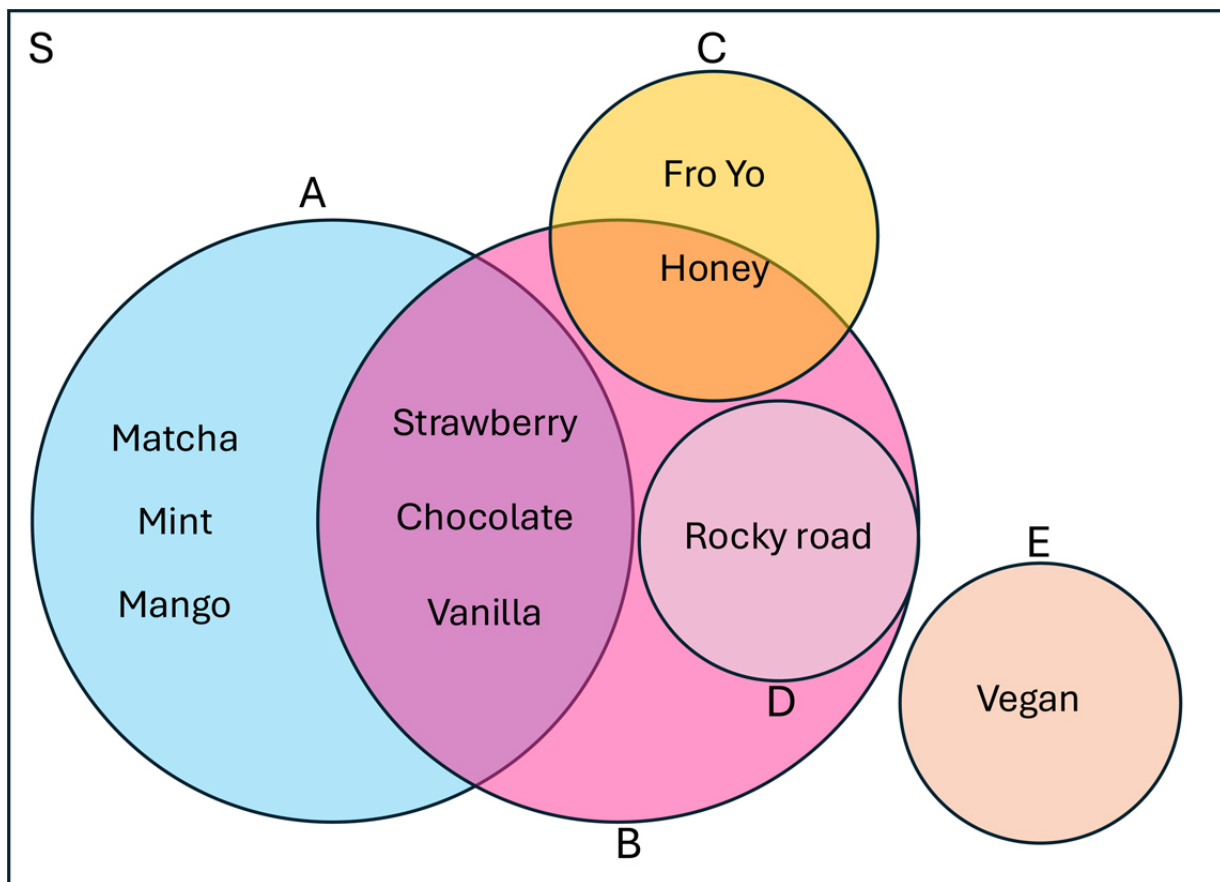


FIGURE 5.3: Events, outcomes and the probabilities of their occurrence.

A sample space contains overlapping sets labelled A, B, C, D, and E. Set A includes Matcha, Mint, Mango, Strawberry, Chocolate, and Vanilla. Set B overlaps with A and contains Strawberry, Chocolate, Vanilla, Honey, and

Rocky road. Set C overlaps with B and contains Fro Yo and Honey. Set D lies within B and contains Rocky road. Set E is separate and contains Vegan.



Extended Description for Figure 5.5

		Your Friend (Y)					
		Chocolate Chip	Strawberry	Mint	Mocha	OF	
You (X)	Chocolate Chip	0.1 (0.125)	0 (0)	0.2 (0.25)	0 (0)	0.5 (0.625) (0.91)	0.80 (1.0)
	Strawberry	0	0.1	0	0	0.05 (0.09)	0.15
	Mint	0	0	0	0	0 (0)	0.00
	Mocha	0	0	0.05	0	0 (0)	0.05
	OF	0	0	0	0	0 (0)	0.00
		0.1	0.1	0.25	0	0.55 (1.0)	1.00

FIGURE 5.5: Probability distributions of choices of ice cream flavors for X and Y : Joint (black), conditional $P(Y|X = CC)$ (green), marginal of X and Y (red), and conditional updated using Bayes' rule $P(X|Y = OF)$ (highlighted in yellow).

Rows list friend X choices as chocolate chip, strawberry, mint, mocha, and O F. Columns list friend Y choices as chocolate chip, strawberry, mint, mocha, and O F. Each cell gives joint probability, with conditional values added for selected chocolate chip and O F cases. Marginal probabilities

appear beside the rows and below the columns, and the totals equal one point zero zero.



Extended Description for Figure 5.8

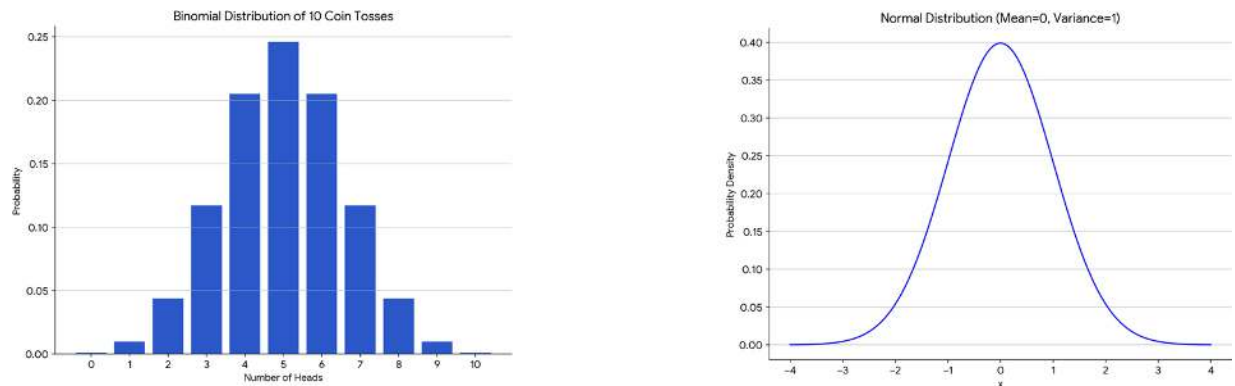


FIGURE 5.8: Plots generated by Gemini AI using the prompts shown in [GeminiBoxes 5.4.2](#) and [5.4.3](#). Left: A Binomial distribution, a classic example of a discrete probability distribution, depicts the probability of number of heads in 10 tosses of a fair coin with equally likely outcomes. Right: The Normal (Gaussian) distribution with mean 0 and variance 1. The Normal distribution is one of the most widely used distributions.

The first plot presents a binomial distribution for ten coin tosses, with probability on the vertical axis and number of heads on the horizontal axis. The bars rise towards five heads and then decrease symmetrically towards ten heads. The second plot presents a normal distribution with mean zero and variance one, with probability density on the vertical axis and x on the horizontal axis. The curve peaks at zero and decreases evenly on both sides.



Extended Descriptions for Chapter 6

Extended Description for Figure 6.2

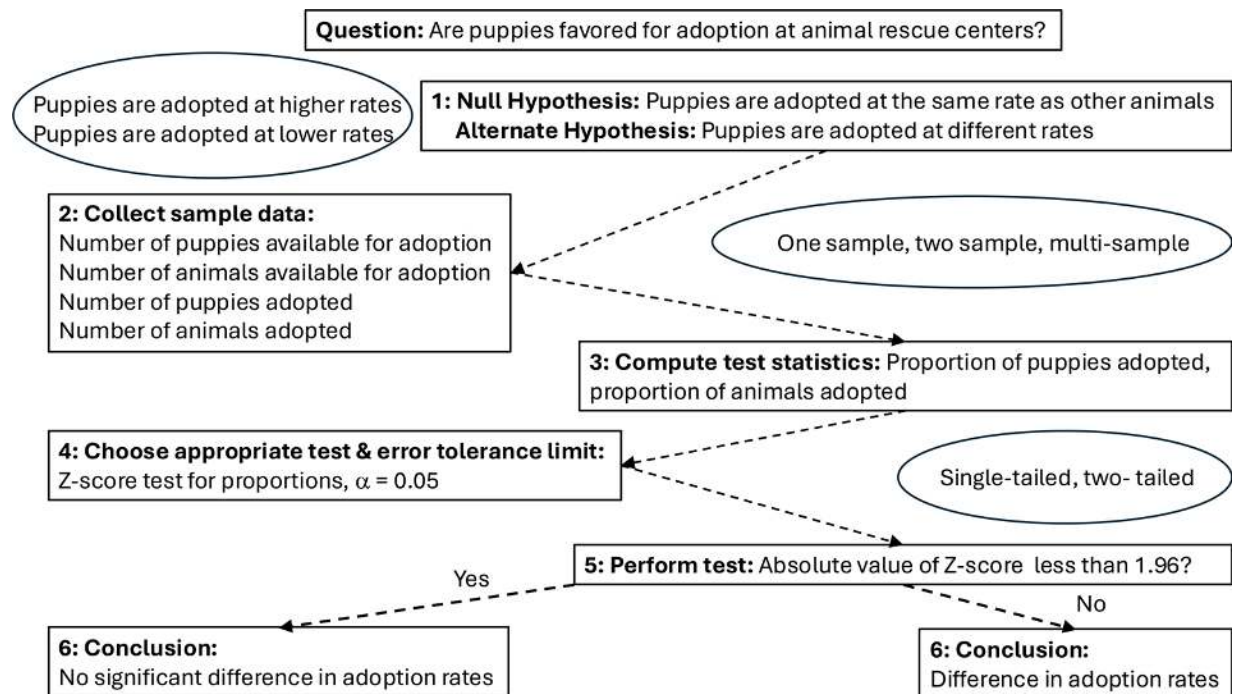


FIGURE 6.2: Are puppies favored for adoption over other animals at animal rescue centers? The process of statistical hypothesis testing is designed to answer such questions in an objective, data-driven manner.

A question about puppy adoption leads to null and alternate hypotheses, then to sample data collection, test statistic computation, test and error tolerance selection, and a Z-score test for proportions with alpha at zero point zero five. The decision step checks whether the absolute Z-score is

less than one point nine six. Yes leads to no significant difference in adoption rates. No leads to a difference in adoption rates.



Extended Description for Figure 6.4

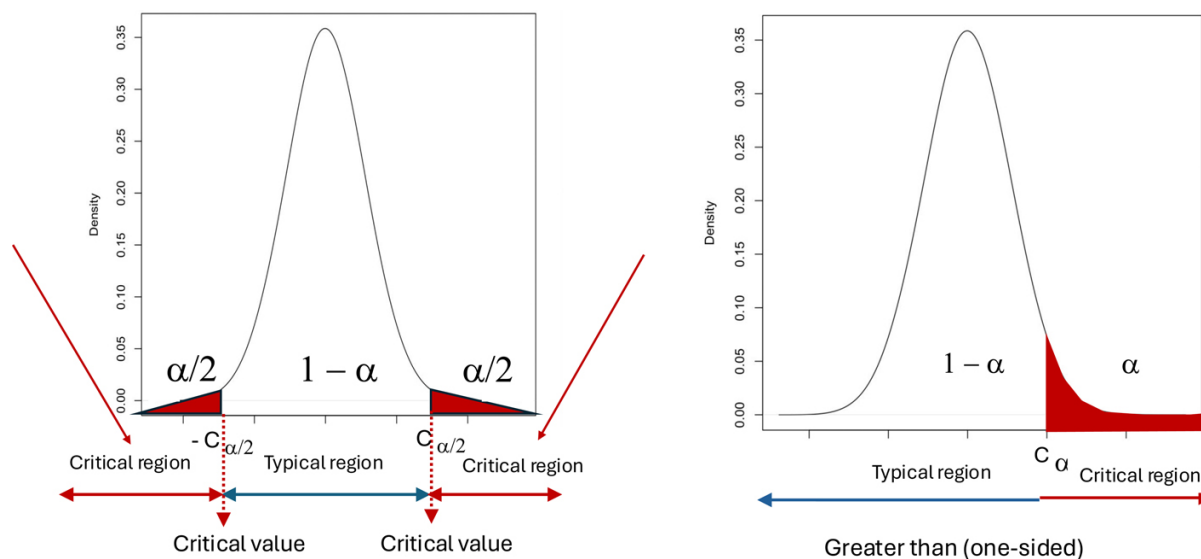


FIGURE 6.4: Left: A two-tailed test, showing critical values, critical and typical regions where the level of significance α is split between the two extremes of the distribution. Right: Single-tailed test.

The first plot gives a two-tailed test, with alpha over two marked at both tails, one minus alpha in the central region, critical values at both boundaries, and critical regions at both extremes. The second plot gives a one-sided greater-than test, with one minus alpha in the typical region, alpha in the right tail, one critical value, and the critical region beyond it.



Extended Description for Figure 6.5

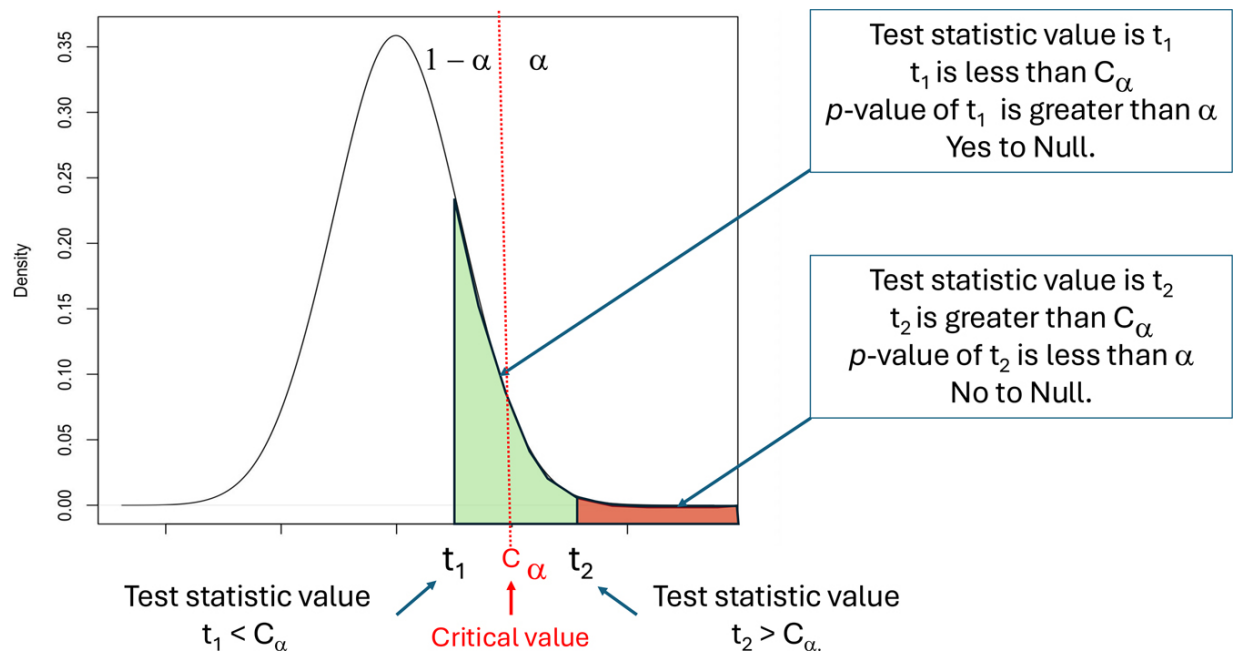


FIGURE 6.5: Making decisions regarding validity of the null hypothesis using probabilities rather than the equivalent critical region. Values of sample statistics t_1 and t_2 , corresponding p -values and critical value C_α .

The curve marks one minus alpha before the critical value and alpha beyond C_α . Test statistic t_1 lies before C_α , with its p value greater than alpha and decision yes to null. Test statistic t_2 lies beyond C_α , with its p value less than alpha and decision no to null.



Extended Description for Figure 6.8

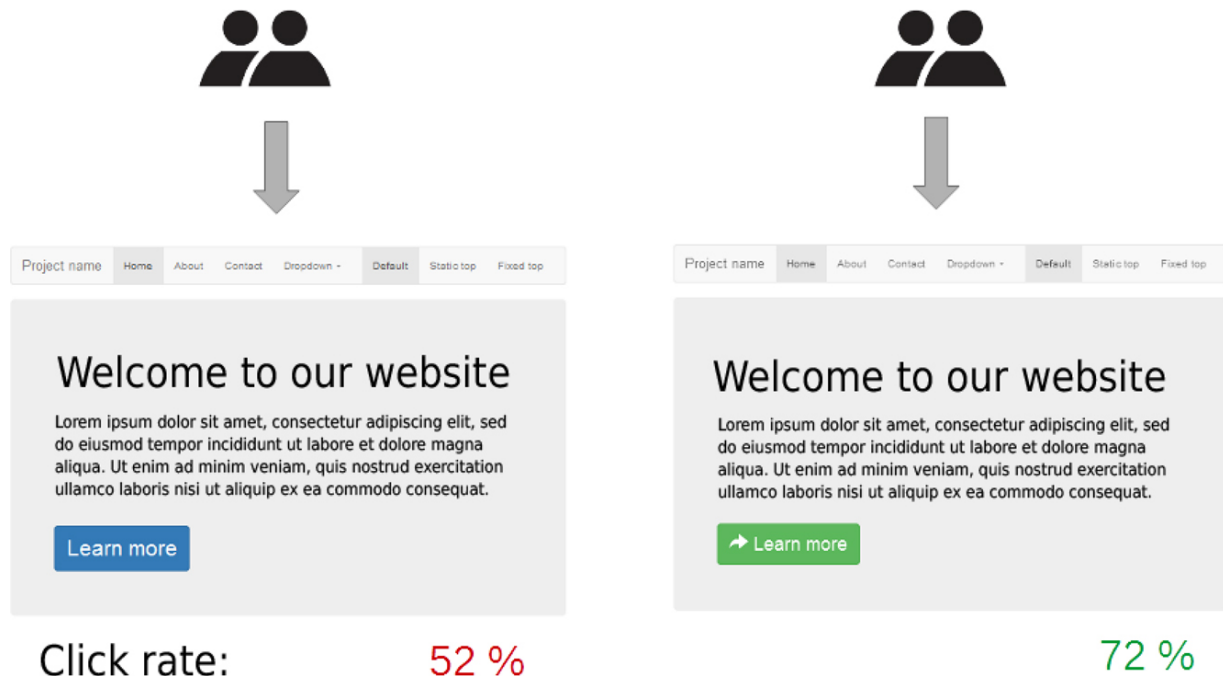


FIGURE 6.8: Social media companies and e-commerce websites assess behavioral psychology and user preferences through A/B testing. Image: Maxime Lorant, CC BY-SA 4.0. [🔗](#)

Each mock-up shows a group icon above a webpage with a navigation bar, a welcome heading, placeholder text, and a learn more button. The first version reports a click rate of fifty-two per cent. The second version reports a click rate of seventy-two per cent. An arrow connects the group icon to each webpage, indicating users interacting with the alternative versions.



Extended Description for Figure 6.10

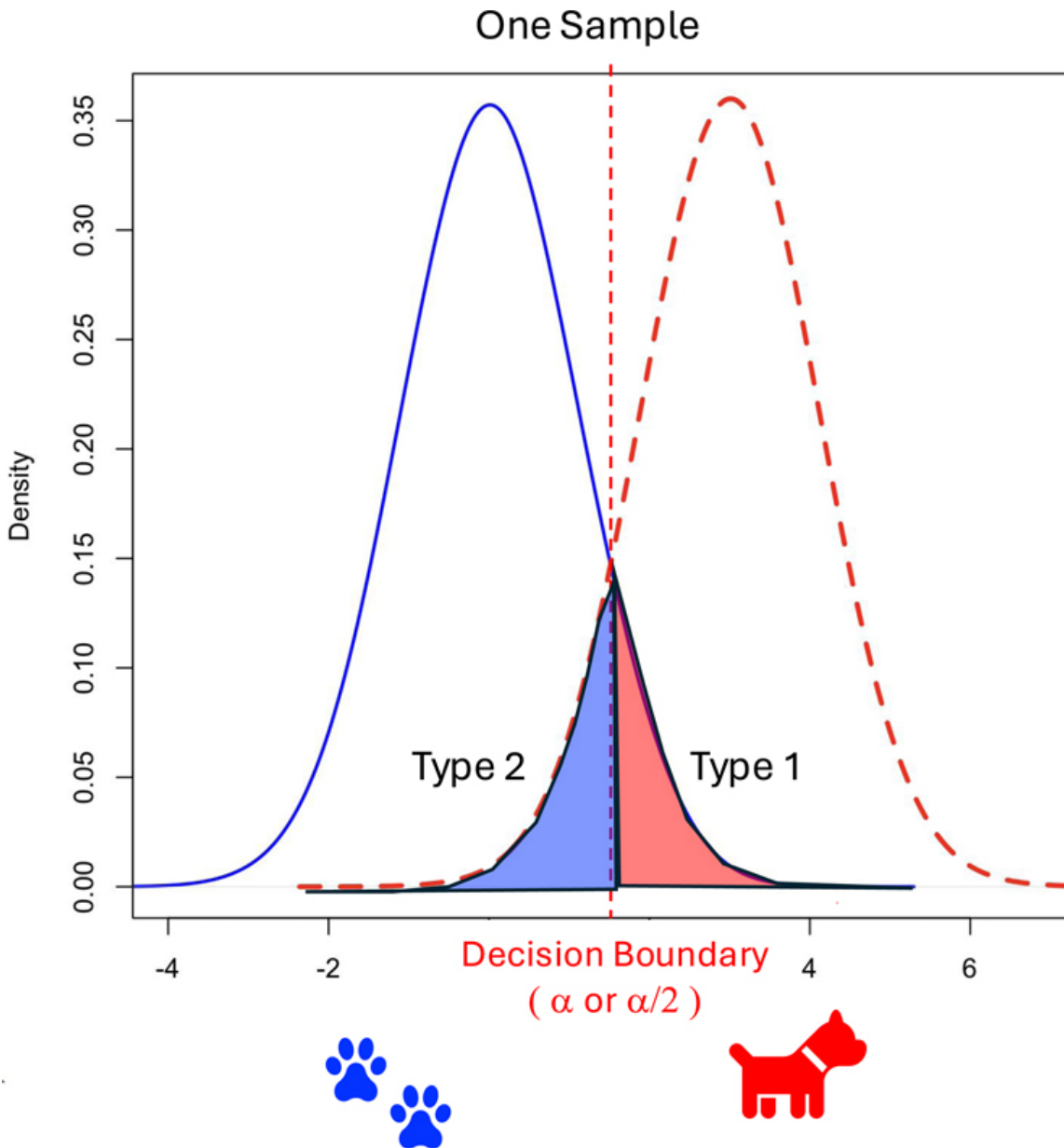


FIGURE 6.10: Type 1 error or false positive: Reject the null hypothesis due to an abnormal sample value from the null distribution. Type 2 error or false negative: Mistakenly accept the null hypothesis (i.e., reject the alternate hypothesis) due an abnormal sample from the alternate distribution.

The vertical axis is density, and the horizontal axis ranges from negative four to six. One curve peaks near one, and the other peaks near three. The decision boundary is marked at about two. Type two is marked on the lower

central area before the boundary, and type one is marked on the lower area after the boundary. Paw and puppy icons appear below the plot.



Extended Description for Figure 6.11

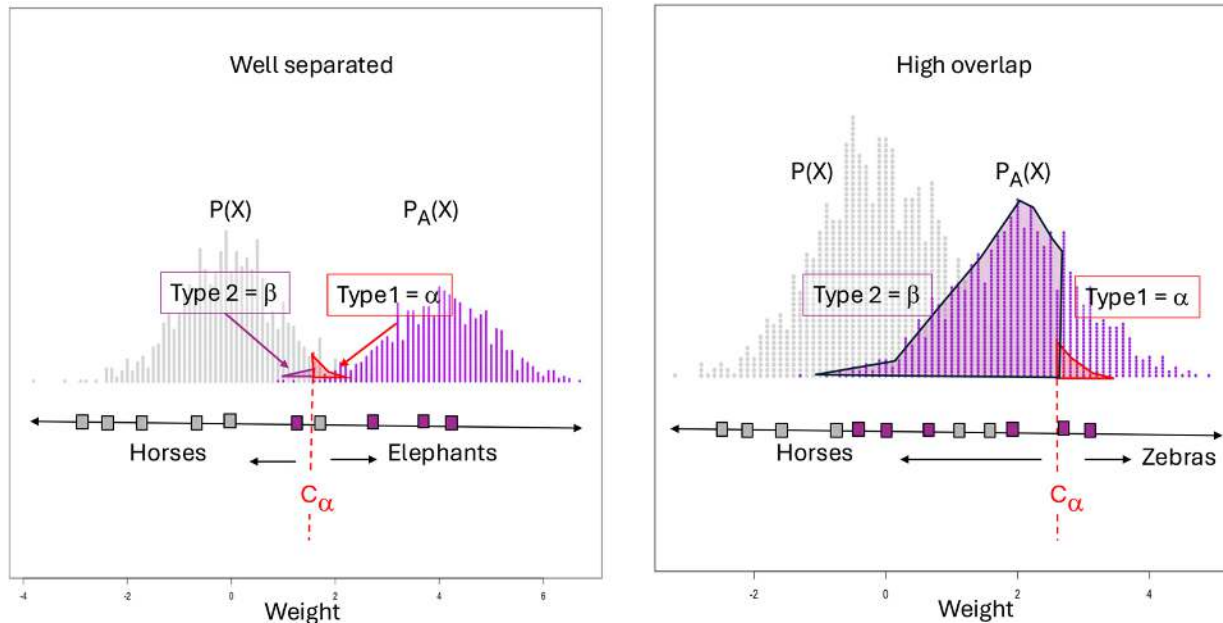


FIGURE 6.11: The sample mean has a high power of distinguishing between well-separated null ($P_{\mathcal{N}}(X)$) and alternate hypotheses ($P_{\mathcal{A}}(X)$) (e.g., distribution of weights of horses and elephants) but low power when the two distributions have a high overlap (e.g., distribution of weights of horses and zebras).

The first plot compares well-separated P of X and P_A of X weight distributions for horses and elephants, with a critical value C_α between them. Type two equals β is small, and type one equals α is marked near the boundary. The second plot compares highly overlapping P of X and P_A of X weight distributions for horses and zebras, with a larger type two equals β region and a type one equals α region beyond C_α .



Extended Description for Figure 6.12

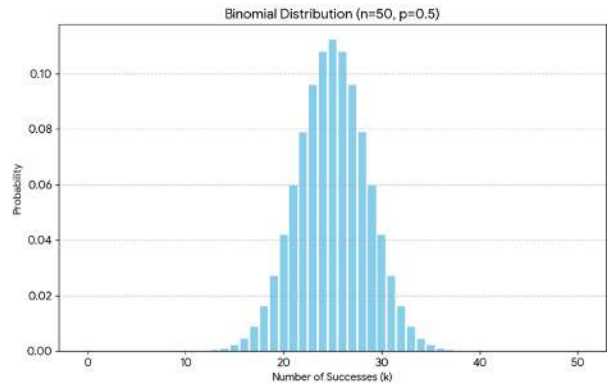


FIGURE 6.12: AI StreetStyle can identify global fashion trends in headpieces. Left: A headpiece from the Kathakali art form in India. (Photo: Maneesha Shahani ). Right: Testing whether the median weight of a Kathakali headpiece is 1.75 pounds using a sign test. Gemini generated Binomial distribution with $n = 50$ and $p = 0.5$ .

The left visual includes a performer wearing a tall layered headpiece, facial makeup, jewellery, shoulder ornaments, and costume, while holding a second decorated headpiece on a vertical support. The right histogram is titled Binomial Distribution with n equals fifty and p equals zero point five. The horizontal axis is Number of successes k , and the vertical axis is Probability. The bars rise from low probabilities at smaller success counts, peak near the middle around twenty-five successes, and fall in a similar pattern towards higher success counts.



Extended Description for Figure 6.13



About 1 in 5 Americans say their diets are extremely or very healthy

% of U.S. adults who say that overall, their diet is ...



Note: Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Feb. 24-March 2, 2025.
"Americans on Healthy Food and Eating"

PEW RESEARCH CENTER

FIGURE 6.13: Shopper checks ingredients on boxed cereal. Wikimedia Commons. US Department of Agriculture. Office of Governmental and Public Affairs.

The left visual includes shelves filled with boxed breakfast cereals and shelf price tags. The right chart states that about one in five Americans rate their diet as extremely or very healthy. The bar chart gives twenty-one per cent for Extremely or Very healthy, fifty-nine per cent for Somewhat healthy, and twenty per cent for Not too or Not at all healthy.



Extended Descriptions for Chapter 7

Extended Description for Figure 7.1

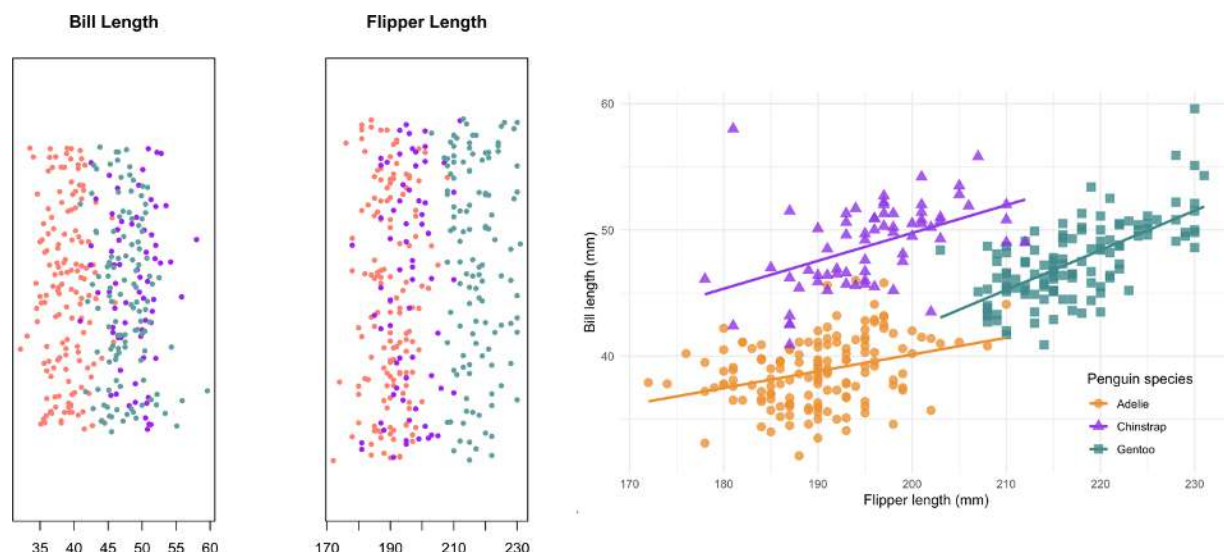


FIGURE 7.1: Plots of three variables from the Palmer Penguin data taken from Horst *et al.*'s [21] manual. Bill length and flipper length values for the three species are stacked on the X -axis. Some species are very similar and have considerable overlap in range of values, e.g., Chinstrap and Gentoo with respect to bill length, and Adélie and Chinstrap with respect to flipper length. Right, a scatterplot of flipper length vs. bill length from Horst *et al.*'s [21] manual. What happens to the overlap between species when you consider the two variables flipper length and bill length together?

The first two narrow dot plots list Bill Length and Flipper Length, with values clustered by species along horizontal scales. Bill length ranges from

about thirty-five to sixty millimetres, and flipper length ranges from about one hundred and seventy to two hundred and thirty millimetres. The scatter plot compares Flipper length in millimetres on the horizontal axis with Bill length in millimetres on the vertical axis. Adelie, Chinstrap, and Gentoo species form separate point groups with fitted trend lines. Bill length generally increases as flipper length increases.



Extended Description for Figure 7.3



Region	Brand			Row Total
	Celine	Prada	Birkin	
NY	100	10	10	120
NJ	20	20	500	540
CT	30	10	100	140
CA	30	40	30	100
IL	0	100	0	100
Column Total	180	180	640	1000

FIGURE 7.3: Left: A U.S. customs officer examines fake handbags (Photo: U.S. Customs and Border Protection). Are there more fakes in certain regions of the US? Right: Hypothetical data about fake handbags and regions.

The view includes an officer examining handbags on stacked boxes in a warehouse. The table lists Region, Brand, Celine, Prada, Birkin, and Row Total. N Y has one hundred Celine, ten Prada, ten Birkin, and one hundred

and twenty total. N J has twenty Celine, twenty Prada, five hundred Birkin, and five hundred and forty total. C T has thirty Celine, ten Prada, one hundred Birkin, and one hundred and forty total. C A has thirty Celine, forty Prada, thirty Birkin, and one hundred total. I L has zero Celine, one hundred Prada, zero Birkin, and one hundred total. Column totals are one hundred and eighty Celine, one hundred and eighty Prada, six hundred and forty Birkin, and one thousand overall.



Extended Description for Figure 7.5

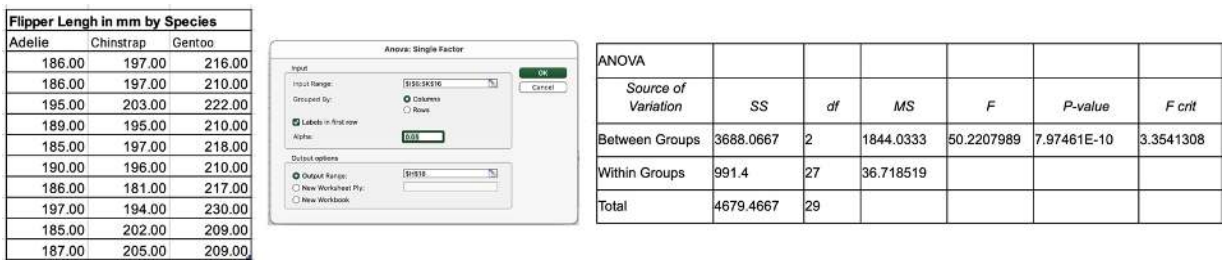


FIGURE 7.5: One-way ANOVA with Excel: Input data. Dialog box selections. Output of one-way ANOVA.

The first table lists flipper length in millimetres for Adelie, Chinstrap, and Gentoo species. The middle dialog box shows Excel A N O V A settings, with grouped by columns, labels in first row, and alpha set to zero point zero five. The final A N O V A table reports between groups, within groups, and total results, including S S, d f, M S, F, P-value, and F crit.



Extended Description for Figure 7.8

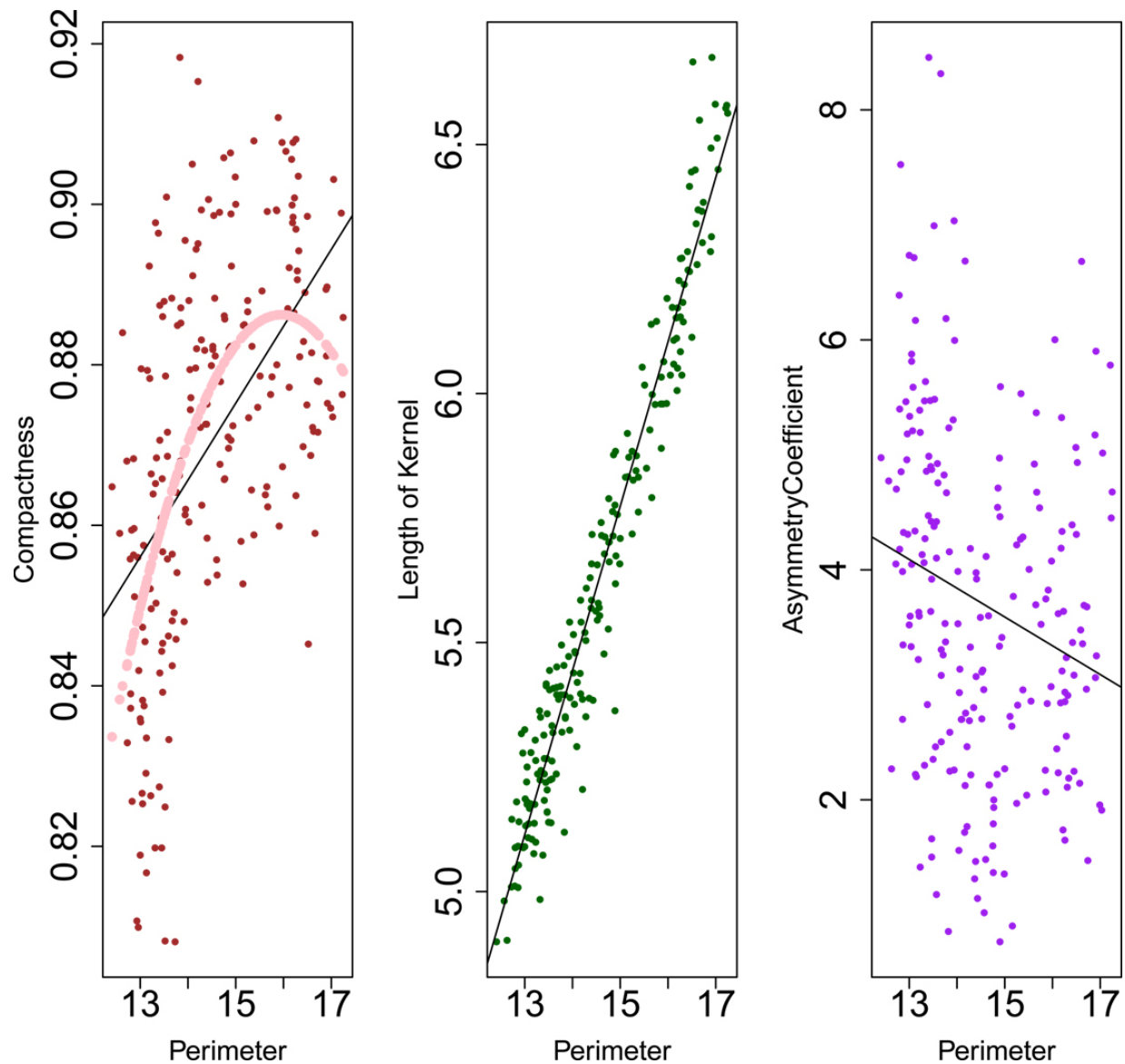


FIGURE 7.8: Scatterplots and lines of best fit for wheat kernels based on the open source UCI Machine Learning Repository data. The figures were created using *R*. $\rho=0.5292436$, 0.9724223 , -0.2173404 , respectively, for the data in the three panels. Is that consistent with your understanding of what ρ means?

The first plot compares perimeter with compactness and includes a curved fitted line that rises and then levels slightly. The second plot compares perimeter with the length of kernel and includes a fitted line with a strong

increasing trend. The third plot compares perimeter with the asymmetry coefficient and includes a fitted line with a decreasing trend. Each plot contains scattered data points with a fitted trend line illustrating the relationship between the two variables.



Extended Description for Figure 7.9

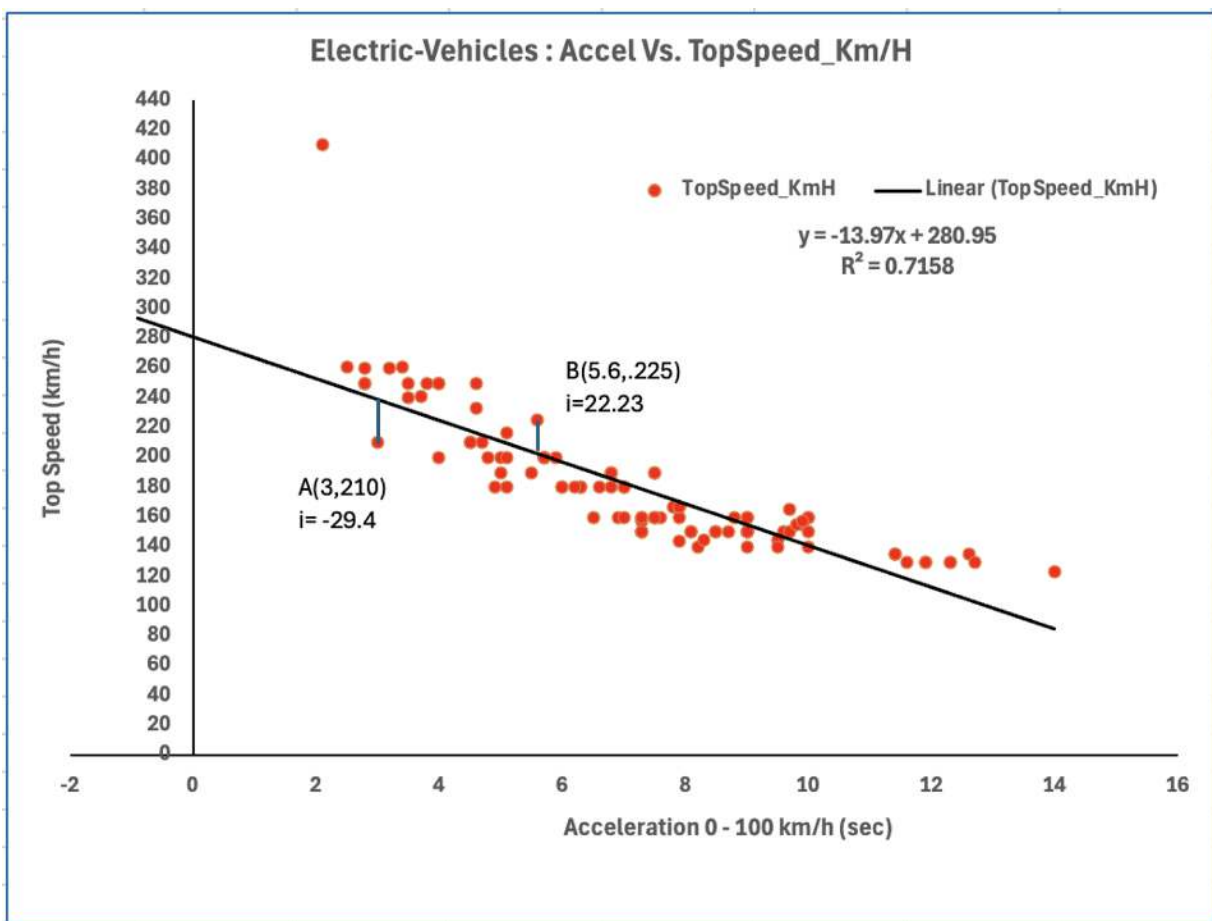


FIGURE 7.9: This chart shows the line of regression for Specs of Electric Vehicles: Acceleration time from 0-100 km/h vs. Top Speed in km/h. EV data set from [kaggle.com](https://www.kaggle.com). Notice the residuals marked on the figure for two data points A and B.

The horizontal axis gives Acceleration zero to one hundred kilometres per hour in seconds, and the vertical axis gives Top Speed in kilometres per hour. The data points generally decrease as acceleration time increases. The fitted line slopes downward and is labelled y equals negative thirteen point nine seven x plus two hundred and eighty point nine five, with R-squared equal to zero point seven one five eight. Two marked residuals identify point A at three and two hundred and ten, and point B at five point six and two hundred and twenty-five.



Extended Description for Figure 7.10

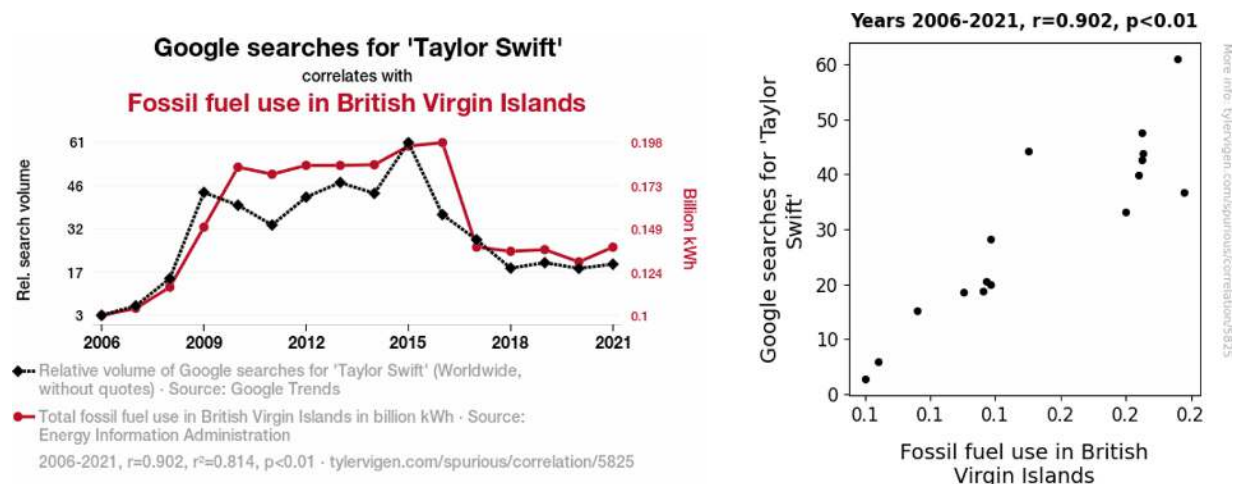


FIGURE 7.10: Tyler Vigen's website [Spurious Correlations](http://tylervigen.com/spurious/correlation/5825) contains hundreds of examples of spurious correlations, and explanations of why they are suspect. Left: Time series of fossil fuel use in the British Virgin Islands and the number of Google searches for “Taylor Swift” during 2006 - 2021. Right: Scatterplot of the same data.

The line chart compares relative searches for Taylor Swift with fossil fuel use in the British Virgin Islands from two thousand and six to two thousand

and twenty-one. Both lines rise to high points around the middle years and then decline. The scatter plot compares the same variables and shows an upward association across the plotted points.



Extended Description for Figure 7.11

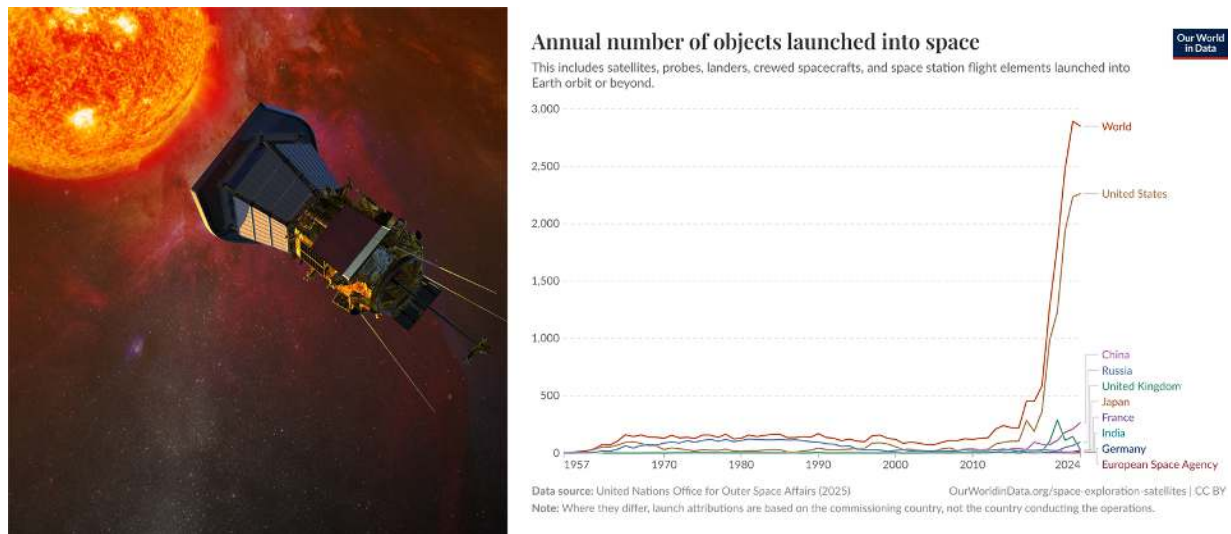


FIGURE 7.11: Left: Artist's rendering of the Parker Solar Probe (formerly Solar Probe Plus) launched in 2018 that plunged into the solar atmosphere in 2021 and sent back data that had never been collected before. The progress of such probes is tracked through telemetry data that continuously monitors health and functioning of the probes to detect anomalies early and ensure mission success. AI methods are increasingly used to handle the staggering amounts of time series data generated and collected by spacecraft. (Source: NASA Goddard Space Flight Center from Greenbelt, MD, USA, CC BY 2.0.) Right: The number of objects launched into space has increased exponentially. (Source: Our World in Data.)

The illustration places a spacecraft near the sun in space. The line chart gives the annual number of objects launched into space from nineteen fifty-seven to two thousand and twenty-four. The world and United States lines stay relatively low for most years, then rise sharply after about two thousand and sixteen. Smaller lines list China, Russia, the United Kingdom, Japan, France, India, Germany, and the European Space Agency.



Extended Description for Figure 7.13

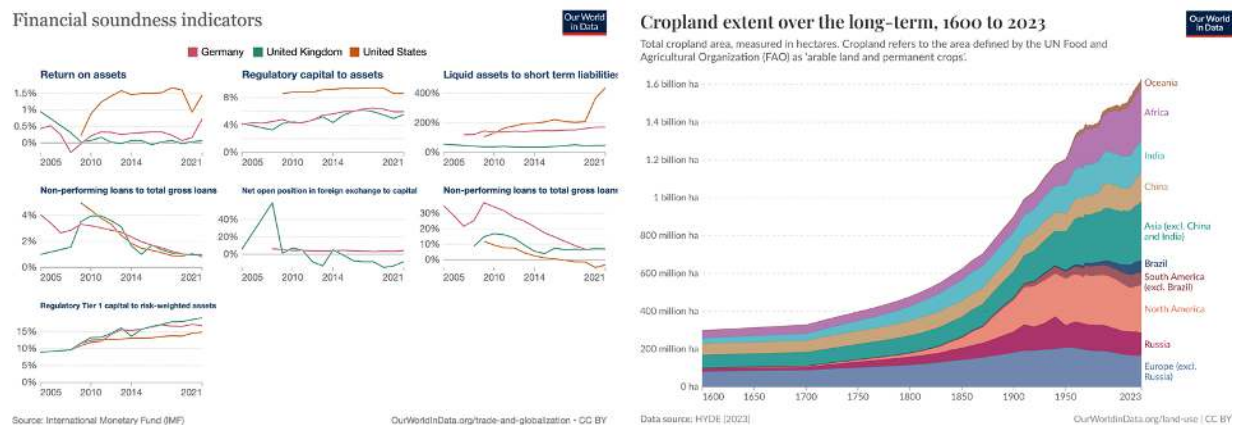


FIGURE 7.13: Time series with trend. Left: Financial soundness indicators. Right: Expansion in cropland by region.

The financial charts compare Germany, the United Kingdom, and the United States across indicators such as return on assets, capital to assets, liquid assets to liabilities, and non-performing loans. The cropland chart traces regional expansion from sixteen hundred to two thousand and twenty-three, with the total area increasing slowly at first, then rising much faster after the nineteenth century.



Extended Description for Figure 7.14

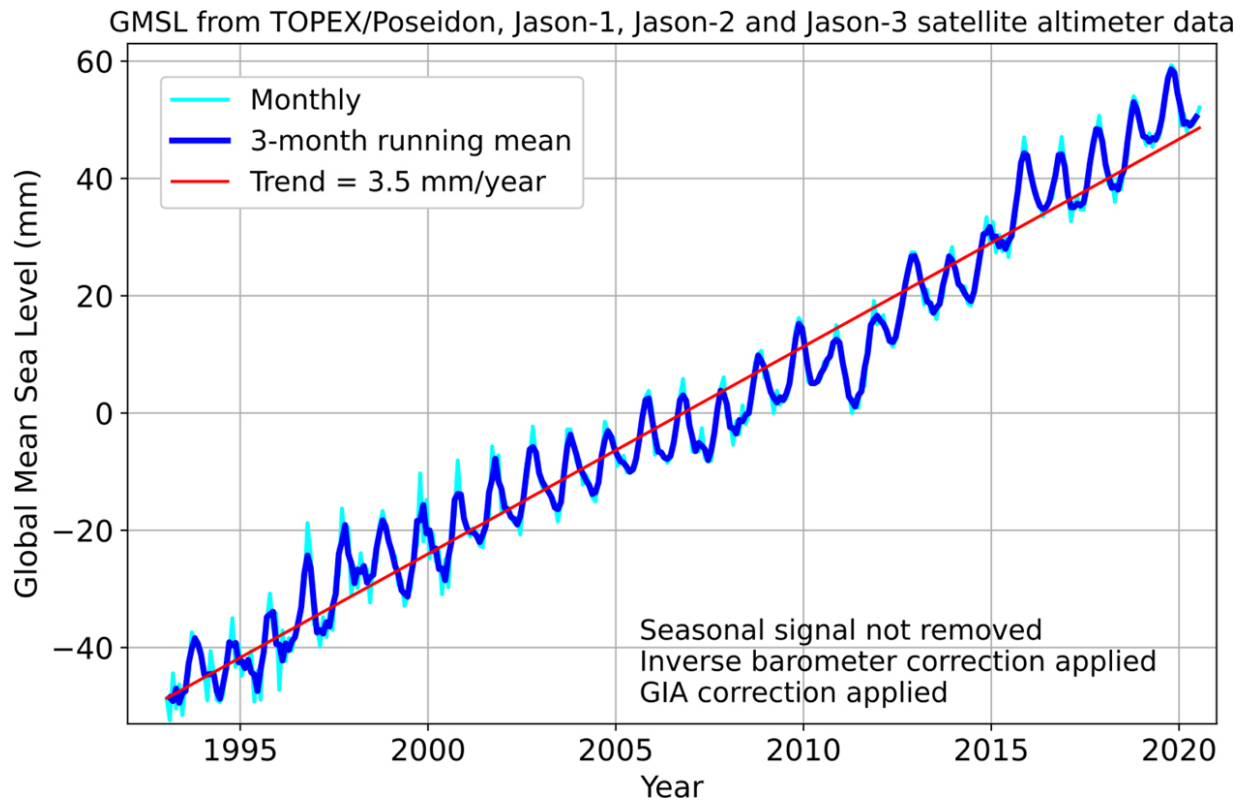


FIGURE 7.14: The development of global sea level from 1992 to 2020 with seasonal variations. Measurements have been conducted via satellites TOPEX/Poseidon, Jason-1 and Jason-2. The trend line clearly shows the constant increase in global sea level of 3.5 ± 0.4 mm/year since the 1990s. Image credit: Geek3.

The horizontal axis gives year, and the vertical axis gives global mean sea level in millimetres. Monthly values fluctuate seasonally, while the three-month running mean follows a steady upward trend from the early nineteen nineties to two thousand and twenty. The trend line indicates an increase of three point five millimetres per year. Notes state that the seasonal signal is not removed, inverse barometer correction is applied, and G I A correction is applied.



Extended Description for Figure 7.17



FIGURE 7.17: Left: Innovative Warn-on-Forecast System from NOAA. Right: Plotting the path of an autonomous vehicle.

The left visual presents forecast tracks with rotating storm signatures near Carter County, intensity categories from E F zero to E F four, a strength scale from weak to very strong, a valid time of eight thirty to ten thirty post meridiem C D T, and agency logos. The right visual presents a vehicle travelling on an urban street, with detected pedestrians, cyclists, vehicles, lane boundaries, traffic information, speed, and the planned driving path overlaid on the road.



Extended Description for Figure 7.19

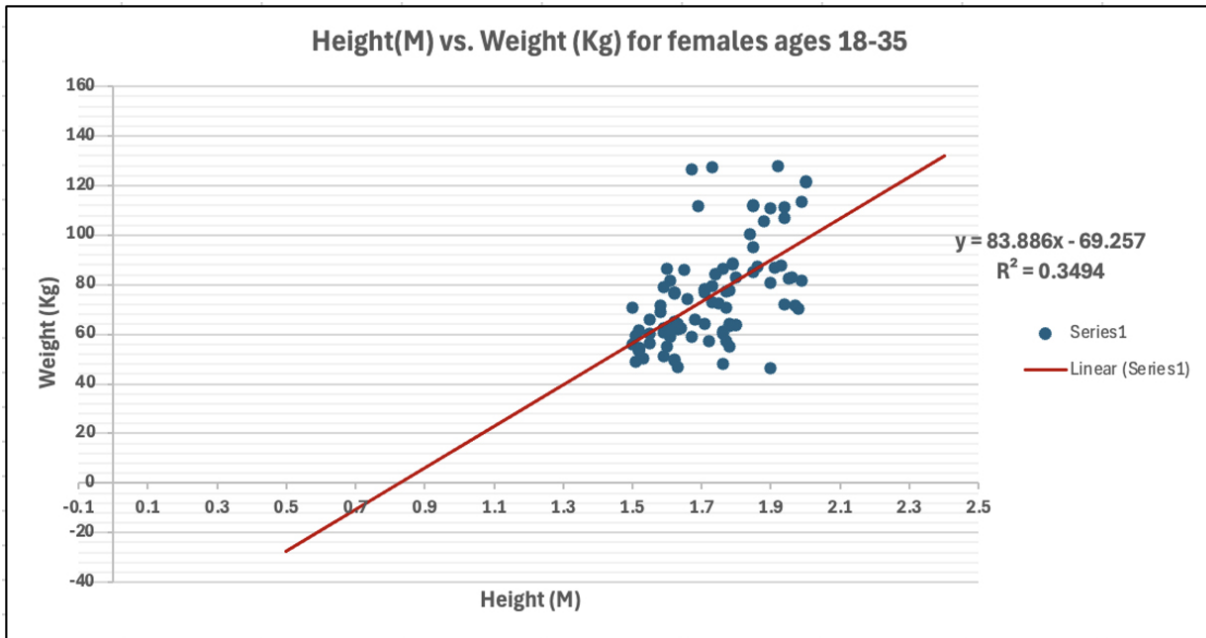


FIGURE 7.19: The relationship between height and weight of women members of a gym between the ages 18-35. The chart was created using Excel. The data was derived from [kaggle.com](https://www.kaggle.com).

The horizontal axis gives height in metres, and the vertical axis gives weight in kilograms. Most points cluster between about one point five and two metres in height and around fifty to one hundred and twenty kilograms in weight. The trend line slopes upward, with the equation y equals eighty-three point eight eight six x minus sixty-nine point two five seven and R -squared equal to zero point three four nine four.



Extended Description for Figure 7.20

Maine Average Temperature

September-August

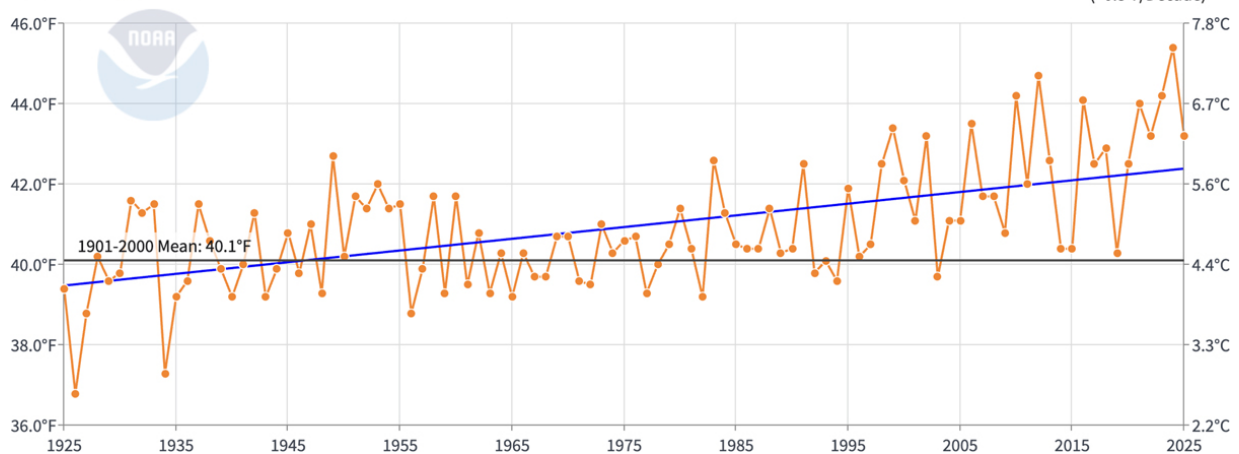


FIGURE 7.20: Average temperature for the month of August in Maine from 1925 to 2025 from NOAA, with blue trend line.

The horizontal axis covers nineteen twenty-five to two thousand and twenty-five, and the vertical axis gives temperature in degrees Fahrenheit and degrees Celsius. The annual values fluctuate, while the trend line rises gradually. The chart marks the nineteen hundred and one to two thousand mean at forty point one degrees Fahrenheit and states an eighteen ninety-six to two thousand and twenty-five trend of plus zero point three degrees Fahrenheit per decade.



Extended Descriptions for Chapter 8

Extended Description for Figure 8.6

Name	Gender	Weight	Name	Gender	Weight	Name	Gender	Weight
George	M	144	George	M	144	George	M	144
Kumar	M	135	Kumar	M	135	Kumar	M	135
Nina		125	Nina	F	125	Nina	F	125
	F	132	Jane Doe	F	132	Jane Doe	F	132
Tom	M		Tom	M	130.4	Tom	M	139.5
Dawn	F	116	Dawn	F	116	Dawn	F	1116

FIGURE 8.6: Missing values in a data table; age imputed based on everyone; age imputed based only on men.

Each table contains the headers Name, Gender, and Weight. The first table includes missing entries for one name, one gender, and one weight. The second table fills the missing values with Jane Doe, F, and one hundred and thirty-two. The third table retains the imputed name and gender but changes Tom's weight to one hundred and thirty-nine point five and Dawn's weight to one thousand one hundred and sixteen.



Extended Descriptions for Chapter 9

Extended Description for Figure 9.4

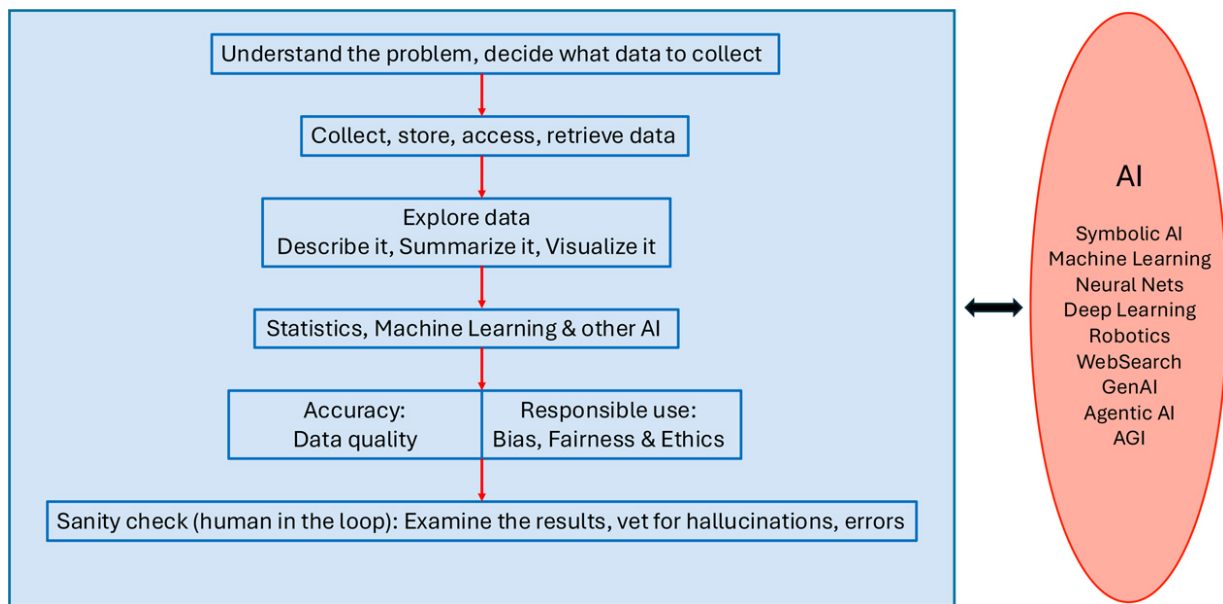


FIGURE 9.4: Data literacy and AI revisited: Concepts from the preceding chapters form the foundation of today's AI which in turn improves the concepts and automates the processes and tasks discussed in those chapters.

The sequence moves from defining the problem and choosing data, to collecting and retrieving data, exploring and summarising data, applying statistics, machine learning, and other A I, checking accuracy and responsible use, and ending with a human sanity check. A linked A I panel lists symbolic A I, machine learning, neural nets, deep learning, robotics, web search, generative A I, agentic A I, and A G I.



Extended Description for Figure 9.10

Workers are more worried than hopeful about future AI use in the workplace

% of employed adults saying they generally feel ____ about how AI may be used in the workplace in the future



Note: Based on all employed adults, including 17% who have not heard about AI use in the workplace and were not asked this question.

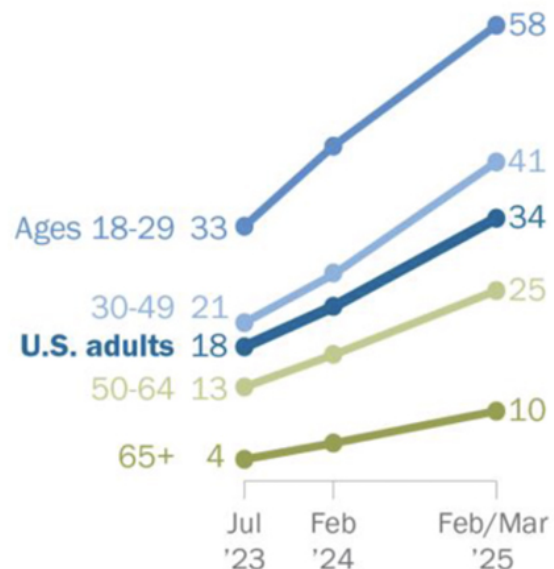
Source: Survey of U.S. workers conducted Oct. 7-13, 2024.

"U.S. Workers Are More Worried Than Hopeful About Future AI Use in the Workplace"

PEW RESEARCH CENTER

ChatGPT use continues to rise; a majority of adults under 30 have used it

% of U.S. adults who say they have ever used ChatGPT



Note: Those who did not give an answer are not shown.

Source: Survey of U.S. adults conducted Feb. 24-March 2, 2025.

PEW RESEARCH CENTER

FIGURE 9.10: Pew Research Center graphics. Left: How Americans feel about AI use in the workplace. Right: Adoption of AI, specifically ChatGPT, in different age groups.

The bar chart reports employed adults' feelings about future workplace A I use, with worried highest at fifty-two per cent, followed by hopeful, overwhelmed, and excited. The line chart shows Chat G P T use rising from July two thousand and twenty-three to February and March two thousand and twenty-five, with the highest use among ages eighteen to twenty-nine and lower use in older age groups.

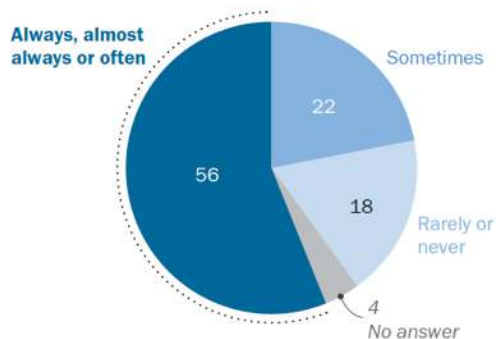


Extended Descriptions for Chapter 10

Extended Description for Figure 10.4

Nearly 6 in 10 Americans frequently skip reading privacy policies

% of U.S. adults who say they ___ agree to online privacy policies right away, without reading what the policies say



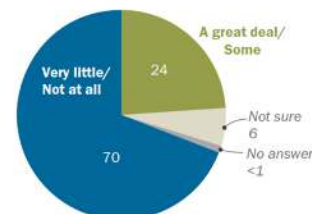
Note: Figures may not add up to 100% due to rounding. "No answer" includes those who did not give an answer or who do not use the internet.

Source: Survey of U.S. adults conducted May 15-21, 2023.
"How Americans View Data Privacy"

PEW RESEARCH CENTER

Most Americans who have heard of AI don't trust companies to use it responsibly and say it will lead to unease and unintended uses

Among those who have heard of artificial intelligence, % who say they trust companies to use AI responsibly ...



Among those who have heard of AI, % who say that as companies use AI to collect and analyze personal information, this information will be used in ways that ...



Note: "Definitely/probably will happen" are combined. Figures may not add up to 100% due to rounding. Those who did not give an answer or who gave other responses are not shown.

Source: Survey of U.S. adults conducted May 15-21, 2023.

"How Americans View Data Privacy"

PEW RESEARCH CENTER

FIGURE 10.4: How Americans feel about companies using their personal information for AI systems. Retrieved from Pew Research Center.

The pie chart states that fifty-six per cent of United States adults often skip reading privacy policies, while smaller groups choose sometimes, rarely or never, and no answer. The second chart reports that seventy per cent of adults who have heard of A I have very little or no trust in companies using

A I responsibly, and bar results list concerns about discomfort, unintended use, and easier deception.



Extended Description for Figure 10.5

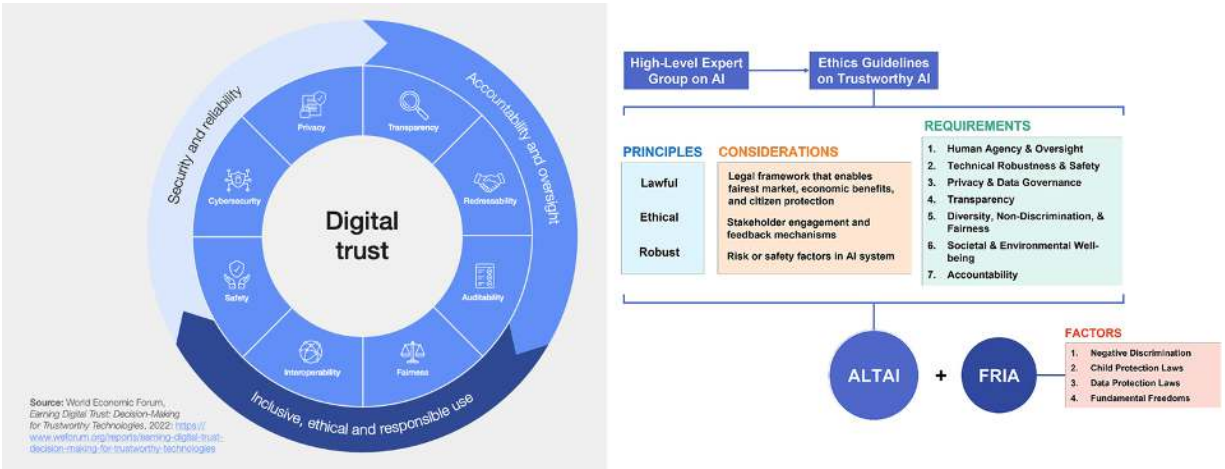


FIGURE 10.5: Left: Goals and dimensions of digital trust as proposed by the World Economic Forum. Right: A checklist proposed by the European Union for measuring trustworthiness of AI applications.

The circular framework centres on digital trust and divides it into privacy, transparency, redressability, auditability, fairness, interoperability, safety, and cybersecurity. The checklist links trustworthy A I guidance to principles, considerations, requirements, and factors. Requirements include human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal well-being, and accountability.

