Coming 2018!

DISCOVERING STATISTICS AND DATA

Third Edition

Instructor Preview

James S. Hawkes

HAWKES LEARNING

FOR REVIEW ONLY

Discovering Statistics and Data

Third Edition

Instructor Preview

FOR REVIEW ONLY

James S. Hawkes

HAWKES LEARNING

FOR REVIEW ONLY

Editor: Robin Hendrix

Assistant Editors: Wesley Duckett, Amber Widmer

Designers: Trudy Gove, D. Kanthi, E. Jeevan Kumar, U. Nagesh, James Smalls, Patrick Thompson, Rebekah Wagner, Tee Jay Zajac

Cover Design: James Smalls and Patrick Thompson

VP Research & Development: Marcel Prevuznak

Director of Content: Kara Roché

HAWKES LEARNING

A division of Quant Systems, Inc.

546 Long Point Road, Mount Pleasant, SC 29464

Copyright © 2017 by Hawkes Learning / Quant Systems, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written consent of the publisher.

Printed in the United States of America

ISBN: 978-1-946158-72-7

FOR REVIEW ONLY

Instructor Sample Contents

FOR REVIEW ONLY

FOR REVIEW ONLY

Letter from the Author

We strived to maintain the friendly conversational writing style of previous editions. The new edition pays homage to the technology-driven data explosion by the incorporation of larger real data sets. Additionally, we have a strong focus on data visualization. Some of the more significant improvements we have made in this edition include the following.

- Technology instructions, data sets, and interactive activities will now reside on our web resource for easier access and to allow us to update technology as new versions become available, as well as add other technologies requested by users in the future.
- The recommendations included in the Guidelines for Assessment and Instruction in Statistics Education (GAISE), published by the American Statistical Association, were carefully considered and incorporated whenever possible in this edition. As a result, you will find more real data sets, and hopefully more relevant data sets for students in this edition. In addition, we have incorporated more technology screen shots into the text so students can see the expected output from an analysis.
- Basic concept questions have been added to help students understand and reflect on what they have learned in each section (a GAISE guideline).
- Expansion of the solution content in our examples to better model the statistical thinking process for students and aid them in making valid conclusions from data (also a GAISE guideline).
- In this edition we have placed more emphasis on "Big Data" and the problems arising from having large data sets. You will find some unique visualization techniques that can be used on large data sets to reveal significant findings that might have remained hidden otherwise.
- One of the key feature requests from our customers was a modernization of our hypothesis testing procedures. To accomplish this, we streamlined our hypothesis testing steps from ten steps to six, revised the null hypothesis to reflect strict equality, incorporated both critical values and *P*-values for all tests, and established stricter checking of assumptions prior to conducting a hypothesis test. We also published the new ASA guidelines on hypothesis testing in the text.
- New web-based resources and apps have been developed to illustrate and provide interaction with significant statistical concepts often misunderstood by students. Along with our courseware, these apps will help students with the comprehension of difficult concepts such as:

The Central Limit Theorem and sampling distributions

Recognition of different distribution types

Multi-variable visualization techniques

Constructing confidence intervals using bootstrapping

Understanding the relationship between Type I and Type II errors in hypothesis testing

The Direct Mail Game -- using confidence intervals and hypothesis tests

FOR REVIEW ONLY
FOR REVIEW ONLY

- We have added over 60 new examples and over 200 new exercises in this edition.
- The Table of Contents has also been streamlined to contain fewer individual sections in a chapter, while at the same time expanding the content to make each section more robust and complete.
- The regression chapter was split into two chapters to discuss Linear Regression and Multiple Regression separately.
- Important content has been summarized in procedure, formula, or definition boxes to enhance student learning and to aid students in reviewing the content for assessments.
- Updated the technology images and output to reflect the latest versions of Excel, Minitab, and the TI-83/84 Plus calculator. We have also incorporated geospatial data visualization examples using the R Statistical Programming Language.

FOR REVIEW ONLY
FOR REVIEW ONLY

Discovering Statistics and Data Third Edition Table of Contents

Note: Content subject to change

FOR REVIEW ONLY
FOR REVIEW ONLY

FOR REVIEW ONLY
FOR REVIEW ONLY

FOR REVIEW ONLY
FOR REVIEW ONLY

An Ocean of Data

Although you usually can't see it, data is being created by everything electronic—phones, cars, computers, appliances, cameras, airplanes, medical equipment, telescopes, atom smashers, DNA sequencers, environmental monitors, manufacturing sensors, social media sites, email, text, and a multitude of other places. We live in a world where the amount of data being generated is incomprehensible, and it keeps growing. The phenomenal growth of stored data is one of the significant achievements of modern civilization and can be considered a measure of the technical advancement of any civilization.

Between 1986 and 2020 the data storage capacity worldwide will have increased by a factor of more than 15,000. Large amounts of data are impacting all the natural sciences and leading to new discoveries in physics, biology, astronomy, and cosmology. In addition, new online businesses are changing the way people shop, find jobs, find relationships, get directions, get recommendations, and find the answers to many questions. Our society is in the midst of a data revolution whose eventual impact may be greater than the industrial revolution. New wealth and convenience have already been created on a massive scale and there is much more to come. One of the companies that has participated in this technological revolution is Amazon.

Amazon has changed the way we shop. The story of Amazon's success is a story in which statistics and data play a very large role. Originally, Amazon just sold books. Because they were born as an internet book store, keeping data about their customers was relatively easy and straightforward. Amazon tracked,

- what customers purchased,
- what they looked at and didn't buy,
- how they navigated through their site,
- whether they were affected by promotions, reviews, or web design layouts, and
- relationships between individuals and groups.

Amazon saw their data as an opportunity to understand the kind of books customers wanted to read. As Chinese general Sun Tzu said,

"Opportunities multiply as they are seized."

-Sun Tzu

Wow, did the opportunities multiply for Amazon!

Amazon's success is based on their connection to their customer through the data they have collected and analyzed. Basically, they have set the standard in online retail for the data a company needs to collect to compete. So far, Amazon is at the top of the online retail mountain. The story you will hear about as you read this book is that data represents opportunity to learn something. Because the amount of data being stored in the world is doubling every two years, it seems like there is going to be a lot of opportunity for individuals willing and able to tangle with it.

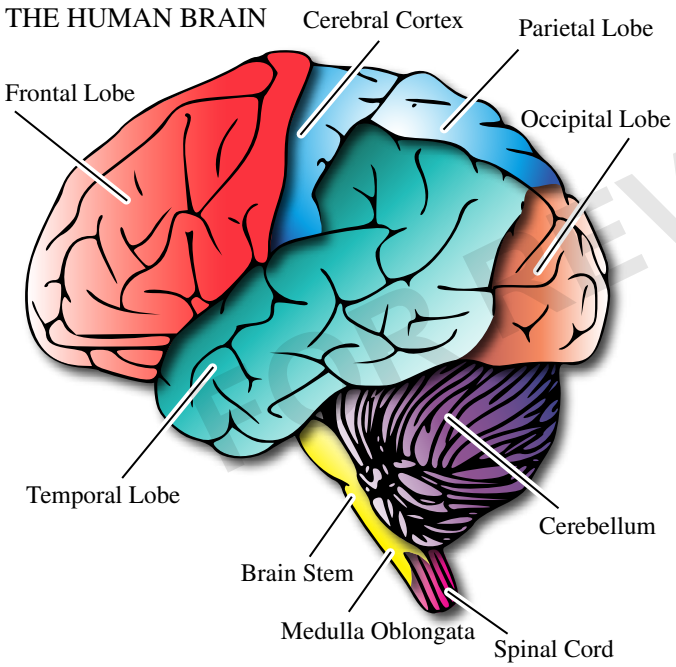
FOR REVIEW ONLY

1.1 The Meaning of Data

Historically we have associated data with measurements and numbers that were purposefully generated to help solve a problem. For example, in 3800 BC the Babylonian empire was interested in things that could be taxed or have some potential military value (especially the availability of adult males for their armies). Trying to solve their taxation and military problem caused Babylonians to perform the first census by counting people, livestock, butter, honey, milk, and other consumables in their territories. Obtaining data in those days must have been a time consuming and rather expensive task, but the data came from measurements or counts, had a purpose, and there was some expectation the data would be examined later.

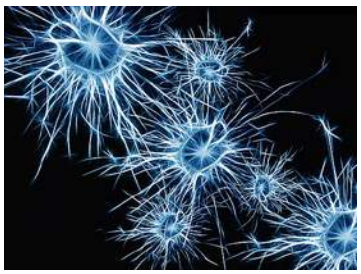
What constitutes data is changing. Presently, the average smartphone owner uses about 3,000,000,000 (3 billion) bytes of data per month, and this number is growing rapidly. The word "data" in this context is different from the historical notion of purposeful measurement to solve a problem. When you stream a video on your phone the data will never be analyzed by anyone. However, anytime you use your web browser your movements around the web are probably being recorded in a database at your browser provider, who in turn will sell the data on your browser activity to a digital marketing group. The marketing group will employ statistical methods to determine what and how they can market to you for their business clients. At least this data is still a measurement of sorts.

Another category of data comes from the desire to create artificial intelligence. As researchers confront the problem of reproducing human "intelligence" they must solve the same data problems we humans do—comprehending large volumes of visual and audio data. An enormous amount of what is considered data in the quest for artificial intelligence are not even measurements in the traditional sense. For example, recently an artificial intelligence company taught a computer how to play an old Atari video game in the same way humans learn, by looking at the screen using a video camera as "eyes". In other words, the pixels on the screen were the data for the machine learning model.



Fortunately, humans are born with the ability to perform powerful sensory and data analytic feats. The brain receives the equivalent of 100 million bytes of sensory information (data) for each second of sensory experience. The eyes alone generate the equivalent of about 90 million bytes of information per second. Assuming we are awake 16 hours then the eyes produce roughly the equivalent to 5.18 trillion (5,180,000,000,000) bytes of data per day. This data plus audio, olfactory, and tactile data are analyzed by the brain's 100 billion neurons. Some neurons in the cerebellum are thought to have 200,000 inputs per neuron. Neurons in the cerebral cortex are thought to have around 10,000 connections per neuron.

FOR REVIEW ONLY



At the other end of the connection spectrum are the neurons in the retina which only have a few connections. Which means there are hundreds of trillions of neural connections in your brain. Essentially, your brain is a supercomputer that produces your model of physical reality that you call the "now".

In addition to creating the "now" our brains produce data driven predictive reality models that are extremely useful in making decisions. Do I have enough time to safely cross the street now or should I wait until the oncoming car goes by? Should I make a left turn now or should I wait until the oncoming bus passes? The brain also designs experiments to gather relevant data for decision making. For example, every morning when you take a shower most people stick their hand or foot into the shower to sense temperature (gather relevant data) before deciding to jump in.

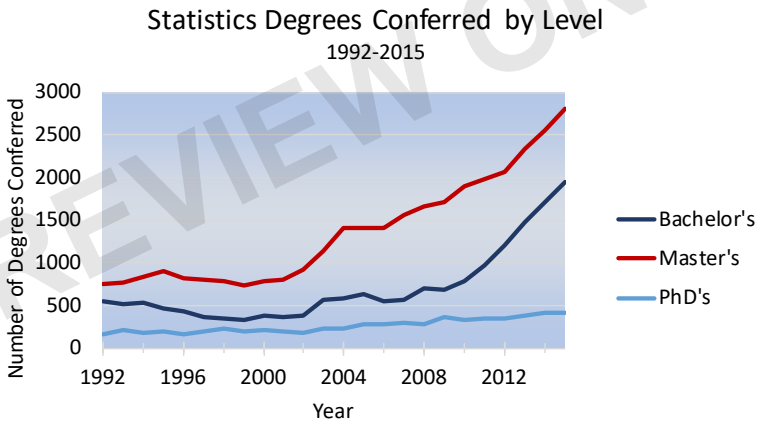
Statisticians do the same sort of things you do unconsciously as you go about your daily life. They analyze data using pictures, summary measurements and build data driven predictive models. They develop methods of designing experiments and gathering data that are cost effective and diminish bias. Essentially, statistics is a "formal" way of thinking with data.

FOR REVIEW ONLY

1.2 Statistics as a Career

We live in the information age, an economy and culture based on computers and information. While, there is an overwhelming amount of new data being produced every year, there have not been large increases in the number of statisticians being produced annually until very recently. That is why Hal Varian, Google's chief economist said,

"I keep saying the sexy job in the next 10 years will be statisticians. People think I'm joking but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades."



The chart depicts the number of statistics degrees conferred in the U.S. by year and by degree. It also illustrates the growing interest in the field of statistics. However, notice that the number of PhD statisticians isn't growing nearly as fast as the number with Master's and Bachelor's degrees. Many industry experts are worried that the U.S. isn't producing enough highly-trained data scientists. Therefore, companies are providing considerable incentives in the form of lucrative salaries to convince people that careers in statistics and data science are well worth their while.

We will begin our journey into statistics by first discussing some of the reasons there is so much data being produced.

1.3 The Data Explosion

To discuss the data explosion, we need to understand numbers of very large magnitudes. It is not uncommon to have a cell phone or computer that has 64 gigabytes (64 billion) bytes of memory. However, data is accumulating in databases so fast we need much larger numbers to describe the sizes of modern databases.

Table 1.3.1 - Quantifying Bytes

	Data Storage Quantities	
1,000,000,000	10^9	1 gigabyte (1 billion bytes)
1,000,000,000,000	10^{12}	1 terabyte (1 trillion bytes)
1,000,000,000,000,000	10^{15}	1 petabyte (1 quadrillion bytes)
1,000,000,000,000,000,000	10^{18}	1 exabyte (1 quintillion bytes)
1,000,000,000,000,000,000,000	10^{21}	1 zettabyte (1 sextillion bytes)

As was mentioned earlier the eyes produce the equivalent of 5,180,000,000,000 bytes of data per day. Using data storage language this would be 5.18 terabytes. In April 2011, the Library of Congress claims it had 235 terabytes of data storage. Data files containing multiple terabytes of data are fairly common, but are still considered large files from an information processing perspective.

A petabyte is 1,000 terabytes. One petabyte of data is equivalent to the data required to generate high-definition video 24 hours a day for 13.3 years. Example of data on the petabytes scale include the following.

- AT&T is thought to transfer more than 30 petabytes of data through its networks every day.
- YouTube currently generates in the neighborhood of 80-100 petabytes of new data annually.
- In 2014 Facebook's data warehouse had upwards of 300 petabytes of data storage capacity.
- The Internet Archive collects digitized materials, including websites, software applications/games, music, movies/videos, moving images, and nearly three million public-domain books. Its collection is currently approaching 20 petabytes.
- New optical telescopes are coming online that will produce 15 terabytes of new data for astronomers every day. The total amount of data collected over the life of the project will be roughly 60 petabytes.

From an information technology perspective, a database that contained several petabytes of data would be considered an extraordinarily large database and very challenging to process. It is unknown what the world's largest database is, but in 2017 the Large Hadron Collider in Cern, Switzerland had about 600 petabytes of data available for analysis. Analyzing data on the petabyte scale usually requires large clusters of computing power with tens of thousands of processors.

An exabyte is 1,000,000 terabytes or one quintillion bytes. One exabyte would be the data required to watch high definition video 24 hours a day for 13,300 years. One exabyte could hold one hundred thousand times the printed material stored at the Library of Congress. Exabytes is the unit of data storage that historically has been used to express the world's technical capacity to store data.

FOR REVIEW ONLY
FOR REVIEW ONLY

.

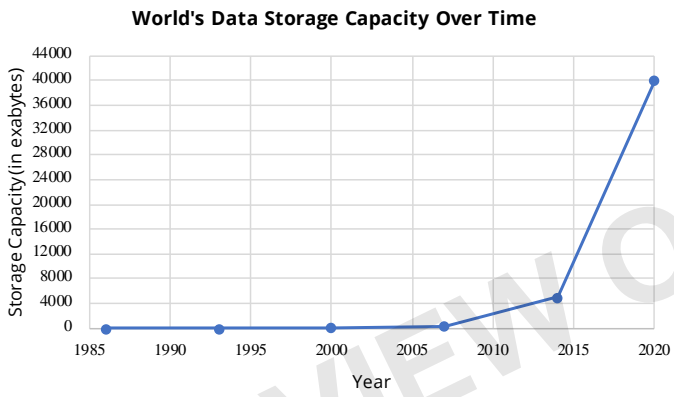


Figure 1.3.1

Table 1.3.2

Year	World's Data Storage Capacity (exabytes)
1986	2.6
1993	15.8
2000	54.5
2007	295
2014	5000
2020	40000

According to IBM, 2.5 exabytes of data are created every day worldwide. More recent estimates have been as large as 5 exabytes per day. Thankfully most of this data is not committed to long term storage. In 2016, it is estimated that Google's data storage capacity was about 10-15 exabytes of data. It has been estimated that the storage requirement to store every word ever spoken by human beings in all history recorded in 16-bit audio would be about 42 exabytes.

A zettabyte is 1,000,000,000 terabytes (one billion terabytes). In 2016 global internet traffic exceeded one zettabyte annually. In 2014 the world's capacity to store data had reached 5 zettabytes. If we added every single grain of sand on the earth and multiplied it by 75 that would roughly equal 40 zettabytes, which is the expected worldwide volume of stored data in 2020.

Why is So Much Data Being Created Now?

New Sources of Data

The internet and World Wide Web have created enormous opportunity to generate and collect data. Every time we go online, carry our GPS-equipped smartphones, use social media and chat applications, post photos or videos or audio files, and send email, we are leaving our digital "footprints" in databases all over the internet.

Essentially, there are two ways in which data is stored, structured and unstructured. Relational databases store data in structured form. That is, data is stored in row/column format plus some other sophisticated data structures. Unstructured data are data that don't reside in a relational database, in other words everything else. Things like email, videos, web pages, texts, and audio files are all examples of unstructured data.

Unstructured data is growing exponentially, particularly in social media and communication applications. For example,

- Facebook has 1.15 billion mobile daily active users.
- In 2015, 205 billion email messages were sent per day.
- In 2017, 8.6 trillion text messages are sent annually worldwide and 4 million text messages are sent each minute in the US.

FOR REVIEW ONLY
FOR REVIEW ONLY

Example of Remote Sensing Technologies

Satellite: There are hundreds of remote sensors used in different types of satellites and other space born systems. These sensors collect massive amounts of data about the earth, the solar system, and even collision detection to avoid other space born objects.

Conventional radar: A system for detecting the presence, direction, distance, and speed of aircraft, ships, and other objects, by sending out pulses of high-frequency electromagnetic waves that are reflected off the object back to the source.

Doppler radar: A radar tracking system using the Doppler effect to determine the location and velocity of a storm, clouds, precipitation, etc.

Seismograph: An instrument used to detect and measure earthquakes.

Sonar: A system for the detection of objects under water and for measuring the water's depth by emitting sound pulses and detecting or measuring their return after being reflected.

A cross-industry study found that less than half of an organization's structured data is actively used in decision-making. Far less than 1% of unstructured data is analyzed or used.

One of the more interesting studies using unstructured data used tweets to forecast the spread of an influenza epidemic through real time tweets relating to symptoms and treatments of the disease. A similar study was done on the spread of chikungunya in Puerto Rico. In infectious disease control, the faster a model can predict where an epidemic has spread, the more influence organizations like the Center for Disease Control (CDC) can have on the propagation of the disease. Although the influenza "Twitter" model has been criticized, it was two weeks ahead of the CDC models in forecasting the spread of the disease.

Surge in Sensors

The poet William Butler Yeats wrote a poem called Under Ben Bulben that contained the line "measurement began our might." Yates died in 1939 and probably could never have imagined how prophetic he was. Even an older iPhone like the 6S Plus has the following sensors: proximity sensor, ambient light sensor, 12MP camera (photon sensor), accelerometer, gyroscope, compass, barometer, near field communication (NFC) technology (a relative of RFID), touch ID fingerprint scanner, and a pressure sensitive display.

An estimated 10 billion sensors were sold in 2011. Since 2011 there has been a rapid increase in sensor use. Each sensor can generate large volumes of data. For example, one sensor at the Large Hadron Collider can generate 14,000,000 measurements per second.

Sensors are low cost ways of collecting measurements and provide data critical to automating any machine or process. Modern factories will have thousands of sensors that feed data into their process control system. Sensors are also used in environmental monitoring systems, machine optimization and control, games, phones, security systems, computers, and appliances—they are even used in trash cans.

Sensors have enabled manufacturers to capture and harness extraordinary levels of data. In industrial settings, they are used to detect position, speed; proximity, pressure, temperature, humidity, level, viscosity, flow, current and voltage, vibration, weight, and almost anything you can think of. A chemical plant making ink or paint or lubricants or hydraulic fluids or fuels might use sensors to measure flow, pressure, level, temperature, and viscosity to feed their process control system. The process control system monitors, store, and analyze sensor data for irregularities and either take automated action or inform maintenance personnel to act.

Sensors are also used to gather data in environmental setting. For example, diamond mines in Canada have heavy metal sensors to detect any chemical leakage from the mine. These sensors are monitored 24 hours per day. Superfund sites (land identified by the EPA containing extremely hazardous waste) are similarly monitored to detect movement in the contaminates. Earthen dams are monitored for changes in soil viscosity which may endanger the structural integrity of the dam. If you are interested in designing control systems using computers and sensor data, there is an engineering major you might wish to consider called control engineering.

Remote sensing means acquiring information about an object or phenomenon without making physical contact with the object. It is also used in numerous fields

FOR REVIEW ONLY

including geography, hydrology, ecology, oceanography, forestry, economics, military, intelligence, geology and even archeology.

There are over 4300 satellites orbiting the earth but only about one third of them are operational. However, one satellite can produce enormous quantities of data For example, the Sentinel-1 (launched in 2014) is a European satellite that collects imagery using radar (day and night) and in all weather conditions. This satellite generated 5 petabytes of data in its first two years of operation. Sentinel 1B was launched in 2016 and two other Sentinel satellites are planned. In addition, there are lots of tiny satellites being launched. In 2017 a company called Planet Labs launched over 100 shoe box size satellites for imaging the earth. Their flock of satellites can completely image the earth every 24 hours and can produce very large image datasets.

Virtually every machine we use has integrated sensors. Automobiles have 60-100 sensors on board. Since reliable self-driving cars and trucks seem to be the future of the automobile industry, it is not surprising that by 2020 the number of sensors on a vehicle is expected to reach 200. In 2020, the automobile industry is forecasted to produce 107 million vehicles worldwide. These vehicles will use over 20 billion sensors. Imagine the data output of 20 billion sensors. However, 20 billion is nothing compared to the estimated one trillion sensors that may be deployed and connected to the internet by 2020—as part of the creation of the "Internet of Things (IoT)." The IoT is creating huge volumes of structured and unstructured data. Assume each sensor on the IoT can produce 10 measurements per second and the 2020 sensor estimate is cut in half and only 500 billion sensors are connected to the internet, then sensor data alone could generate 5 trillion pieces of data per second. If this happens, in one year these sensors could generate an incomprehensible 157 exabytes of data (157,000,000,000,000,000,000 bytes).

The Host with the Most

South Korea is the world's most connected nation, with broadband connections reaching 98% of homes. However, only about 85% are thought to actually use the Internet regularly.

1.4 The Fusion of Data, Computing, and Statistics

The Apollo moon landing was an epic event in American History. Without digital computers, it probably wouldn't have happened. The Apollo Guidance Computer (AGC) provided computation guidance, navigation, and control of the Command module and the Lunar Module. It was the first computer to use integrated circuits. Additionally, NASA used IBM System/360 Model 75 mainframes on the ground to perform independent computations and communication to the spacecraft as well as monitor astronaut's health and environmental data.

The iPhone 6 is a relatively old version of the iPhone. It uses an Apple-designed processor that can execute 3.36 billion instructions per second. Its processor can perform instructions 120,000,000 times faster than the AGC. If you had given an iPhone 6 to NASA engineers in 1969 it would have appeared to be alien technology. The iPhone 6 is at least 1000 times more powerful--some have estimated over 1,000,000 times—than all the NASA combined computer resources in 1969.

Why is this important to statisticians? When you apply statistical methods to increasingly larger data sets, there are very few data sets that you are likely to encounter that will challenge modern computers. Using current computing

Modern Computing: Alan Turing

There have been many contributors to modern computing, but a few contributors deserve special mention. Early electronic computers had to be physically rewired to change its "program". Alan Turing was the first person to come up with the notion of a "stored program" in which the program was stored in memory rather than literally "hard wired." He presented this idea in a 1936 paper. While this was a fundamental idea, before modern computers were created a couple of the building blocks needed to be invented.

FOR REVIEW ONLY

The First Analog Computer

Mechanical aids to computation have been around for a long time. The Sumerian abacus appeared somewhere between 2700-2300 BC. The Greeks invented what is considered the first analog computer called the Antikythera mechanism; it was used for astronomical calculations in 150-100 BC. It took one thousand years for another computing device to attain the complexity of the Antikythera.

The Prototype

The prototype of what we call a computer today (digital computers) were created in the late 1940s. However, until about 1978 digital computers were only affordable by business and government entities. The first tidal wave of new data was generated by personal computers. But the data generated by personal computers really wasn't an important contributor to global data until data networks evolved.

technologies data sets that are feasible to process simple statistical models in a reasonable amount of time are on the order of gigabytes. We will discuss some exceptions in the section on Big Data. This leads us to data networks.

Data Networks

All the personal computers and sensors that are available today would not be very useful if they could not easily share data over a private or public network. Private computer networks exist in almost every company, educational institution, and business that use computers. You even have a private network in your home if you have wi-fi there. In the 1980s private networks ran on proprietary software and one network vendor's software did not easily communicate with other network vendor systems.

For large organization like the US Department of Defense, which had many different networks, this was a problem. A group within the defense department, the Advanced Research Projects Agency Network (ARPANET), took the first steps to "connect" private networks. In 1983 a standard communication protocol TCP/IP (Transmission Control Protocol/Internet Protocol) was adopted by ARPANET. This was the beginning of the internet. Now, TCP/IP is used by virtually every private network in the world and almost all private networks are connected (the internet) to one another and can share data. In the early days of the internet private networks were connected but there was not a great "system" for sharing data. In 1990 an English computer scientist, Tim Berners-Lee, working at the Large Hadron Collider (LHC) in Cern, Switzerland proposed a global hyperlinked information system for sharing information between computer systems. His memo to Cern management proposed a "Hypertext project" named Worldwide Web as a "web" of "hypertext documents" to be viewed by "browsers" using client-server architecture. Berners-Lee built the first web server in 1991. Now, there are over 80 million web servers and 1.2 billion web sites worldwide.

1.5 Big Data

Big Data is a loosely defined concept used to describe data sets produced by our globally networked, internet driven, sensor-laden world. While there isn't broad agreement on exactly what big data is, there is broad agreement that Big Data will accelerate the pace of discovery in science as well as innovation in commerce. In fact, that has already happened.

There have been numerous ideas about what makes data "Big Data." Most experts would agree that it is a large volume of data, structure or unstructured. Beyond that, opinions differ. One criteria that is appealing is any data set that is too large to process on commonly available computer systems. More recently the term has been used to refer to a set of analytical models that are used on data sets, regardless of their size. At the moment, the most common meaning of "Big Data" is a set of data sufficiently large to be challenging to analyze at a "typical" data center. As a frame of reference, the minimum for a "large" data center would be on the order of tens of thousands of servers and thousands of data storage arrays.

FOR REVIEW ONLY

Another characteristic of "Big Data" is that it requires teams of programmers, database programmers, statisticians and machine learning experts to analyze the data. The "Big Data" team will usually be using highly scalable cloud computing resources.

Data has many attributes. However Big Data seem to have four attributes that make it different: volume, variety, velocity and veracity. These characteristics constitute the four Vs of Big Data.

- **Volume** is the scale of the data and Big Data implies large volumes of data. According to IBM, most companies in the U. S. have at least 100 terabytes of stored data. But some companies have exabytes of data and are receiving 100's of terabytes of new data every day.
- **Variety** is the different forms data can take—from traditional data elements in a structured database to highly unstructured images, twitter feeds, movies, and audio.
- **Velocity** is how fast data is coming at your data processing and data management infrastructure. There is a technical term called "streaming data" which is data generated continuously and usually in small amounts by thousands of data sources. Streaming data would be common in ecommerce, gaming, social networks, stock trading, and telemetry data from monitoring systems. One aspect of Big Data is that the data streams have substantial velocity. YouTube, for example, has an amazingly large data stream.
  - → 300 hours of video are uploaded every minute
  - → 5 billion videos are watched every day
  - → 30 million visitors per day

  It is quite a technical challenge to neatly place high velocity data into the appropriate data repository every minute of every hour of every day without fail.
- **Veracity** is the trustworthiness of the data. Data is a major asset of any company, institution, or government agency. Uncertainty, bias, or inaccuracies in the data make it less valuable for meaningful analysis and decision–making.

Flash Boys: Data Velocity

Flash Boys is a book written by Michael Lewis about high frequency stock trading. Part of the book describes the great lengths a firm went to reduce the time it takes to send a buy or sell order between New York and Chicago. The best available time in 2008 was 14.65 milliseconds (14.65 thousands of a second). But theoretically it should be possible to communicate between the two cities over a fiber line in 12 milliseconds. The book tells the story of what it took to build and market a "direct" fiber run between New York and Chicago at a cost of $300 million dollars. Note, that is 300 million dollars spent to improve data velocity by slightly more than 2 milliseconds.

Sources of Big Data in Science

Despite the availability of enormous computing power, some areas of science and industry have data sets so large that they overwhelm modern computing systems. In the sciences, particle physics, astronomy, genomics, meteorology, and internet searches have amassed enormous quantities of data.

Science is being profoundly affected by an abundance of measurements. Almost all large natural science data sets grow because the data are being measured and gathered by specially designed automated measurement systems (machines).

- In the case of genomics, the development of very fast and relatively inexpensive DNA sequencers.
- In the case of astronomy and cosmology it is new telescopes with very large digital camera arrays.
- In the case of meteorology, it is satellite imagery and automated weather sensors.

FOR REVIEW ONLY
FOR REVIEW ONLY

"The Most Important Master's Thesis of the 20th Century"

In 1948 Claude Shannon wrote a paper entitled "A Mathematical Theory of Communication" which was the foundational work for a field now called information theory. The paper introduced the term "bit" and demonstrated that a series of bits "1s and 0s" (eight of them make a byte) could be used to represent all information. The bit/byte would become the standard unit for data storage and network communication of the future. Shannon's foundational work in information theory was not his only contribution. His masters thesis has been called the most important masters thesis of the 20th century. It showed that electrical switches could be configured to perform Boolean logic functions (i.e. digital logic). Shannon's work became the foundation of digital circuit design. Digital circuits are the fundamental component of all digital computers and without them we would not have modern computers, nor modern statistics.

- In the case of particle physics, it is particle colliders (like the eight-billion-dollar Large Hadron Collider LHC).

These sensing machines are generating enormous quantities of data from their sensor arrays. The data they are providing offer a huge opportunity to advance our understanding of science and medicine.

Medicine

It used to be that doctors recorded everything they did on your chart. Now, it all goes into a database. This happens every day on every patient. For example, your doctor may order a test such as an MRI of your brain, or an echo cardiogram of your heart. It would not be unusual for an MRI of your brain to be 220+ megabytes of data. An echocardiogram could be as little as 40 megabytes and an interventional study (a surgical procedure) could be as much as one gigabyte. Even a chest x-ray would be about 20 megabytes. Medical technology generates an enormous amount of data. Figure 1.5.1 shows the number of echocardiograms paid for just by Medicare from 2007 to 2011. Just to store the annual echocardiography data that Medicare pays for would require roughly 280 petabytes.

Table 1.5.1 - Echocardiograms Paid for by Medicare 2007-2011

Year	Number of Echocardiograms
2007	6,816,517
2008	6,918,949
2009	6,951,249
2010	7,054,577
2011	7,077,554

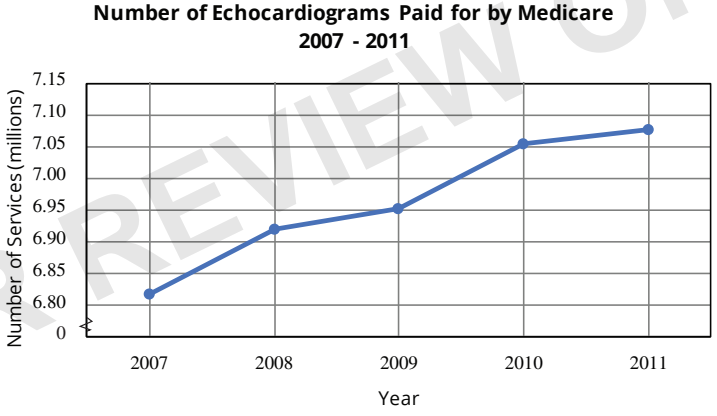


Figure 1.5.1

Once a patient's data is anonymized it can be combined and aggregated. Looking at disease from a broad perspective of aggregated patient health data can provide new insight in a disease process. It can reveal biomarkers that were unknown and could more easily predict the trajectory of a disease and perhaps offer an intervention.

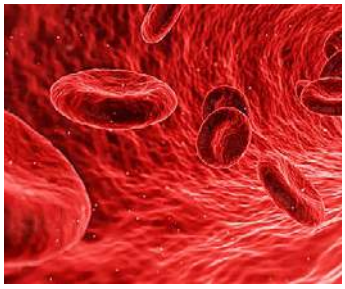
For example, there are two large cancer databases that have followed hundreds of thousands of cancer victims for 15 years or more. Once an oncologist diagnoses the

FOR REVIEW ONLY

specific cancer they need to develop a treatment plan. A good oncologist usually will recall 6-8 similar cases to help formulate the plan. Now, the oncologist can call upon a cancer database that will have thousands of similar cases and the information system attached to the database can make recommendations for the treatment plan.

Also, the oncologist might use the cancer genome atlas—which classifies cancers by their genome—looking for treatments against a specific cancer genome. The oncologist might utilize the new field of proteomics, which is a study of the proteins in your blood. One drop of blood passed through a superconducting magnet can generate 40+ gigabytes of data on all the proteins in the blood, which is the environment that the cancer cells are growing in.

Genomics



Genomics is a field that maps and studies the DNA (genomes) of biological entities. Every plant, animal, bacteria, virus has a design that is contained in its genetic material (DNA) stored in each cell. The DNA is a blueprint for the organism. It determines whether an organism will produce leaves or legs and of course many other things. In 1995 the genomes of two bacteria, homopholus influenza and microplasm genitalia, were sequenced, meaning the letters of their DNA were read and stored. The influenza genome is 2,000,000 base pairs long. Once there were large volumes of data, it wasn't long before computer scientist and statisticians began entering the field of biology.

The DNA strand is made up of four chemical building blocks, called nucleotides [adenine (A), thymine (T), guanine (G) and cytosine (C)]. Essentially, DNA encodes information. Human DNA has about 3 billion base pairs of nucleotides. To sequence a human genome means to determine the specific base pairs for nearly all 3 billion pairs associated with the individual's genome. So, one entry into a human genomic database contains various combinations of [ATGC] for the individuals 3 billion base pairs.

Genomics is around 20 years old. The Human Genome Project originally took 10 years to process one human genome; now this can be achieved in less than a week. As of 2015 all genomic data represented approximately 25 petabytes. The amount of data being produced in genomics daily is currently doubling every seven months. Within the next decade, genomics is looking at generating somewhere between 2 and 40 exabytes a year, depending upon whether the data doubles every seven months or every 18 months as shown in Figure 1.5.2.

It is estimated that 1 billion people will have their DNA sequenced by 2025. If this happens genomic databases will likely be the largest databases in existence.

There are databases that contain large numbers of completely sequenced human genomes. There are databases with completely sequenced genes for people with autism, cancer, muscular dystrophy, heart disease and virtually any other disease that might have a genetic component. There are DNA databases that specialize in specific mammals, insects, bacteria, viruses, and plants.

FOR REVIEW ONLY
FOR REVIEW ONLY

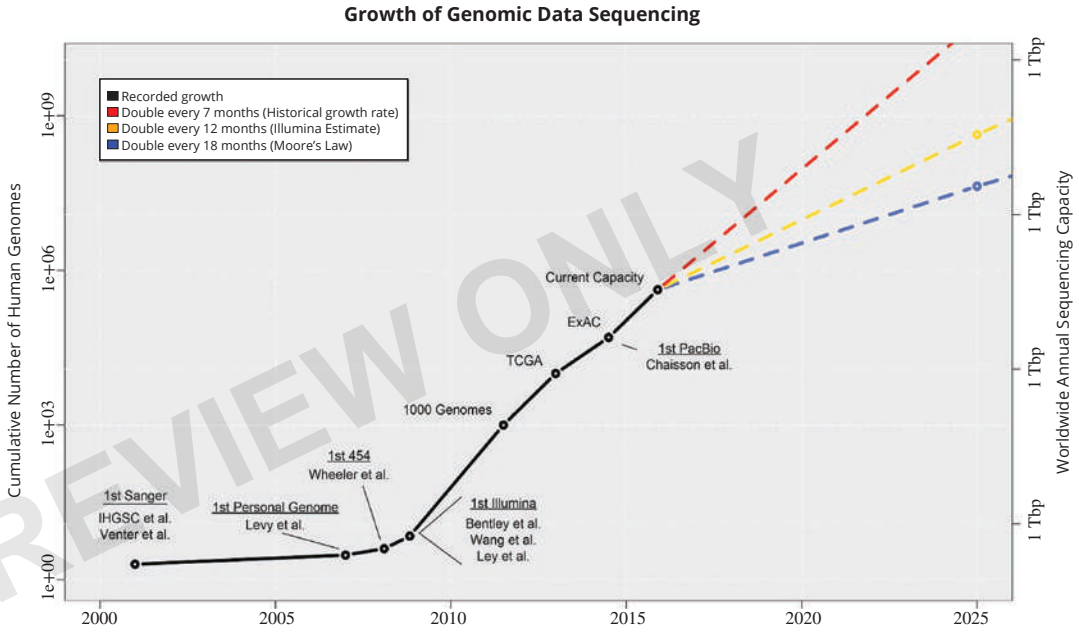


Figure 1.5.2

Astronomy and Cosmology

The Sloan Digital Sky Survey (SDSS) began collecting astronomical data in 2000. The telescope is collecting data at the rate of about 200 gigabytes per night. In its first few weeks of operation it collected more data than all the data collected in the history of astronomy.

The Large Synoptic Survey Telescope (LSST) being built in Chile is the successor to SDSS. When it becomes fully online in 2018 it will acquire 15 terabytes of data per night or 1.28 petabytes annually. That is, each night it acquires 75 times more data than the SDSS. The data will be images and will be analyzed by computer programs that require massive amounts of computing power.

The James Webb telescope is scheduled to be launched in 2018. It will operate 1,000,000 miles from earth and will be 5 times more powerful than the Hubble telescope. It will be able to directly image exoplanets of nearby stars. If the downlinks on the satellite world perfectly for 10 years it will generate 209 terabytes of data over its life.

NASA has 100 active missions. In the time it took to read the previous sentence NASA downloaded 1.73 gigabytes of data from its missions. And, the rate of NASA's data gathering is growing exponentially. NASA has plans for missions that will stream 24 terabytes a day.

But the mother of all telescopes is the forthcoming $2.1 billion Square Kilometre Array (SKA) radio telescope. When it is completed in 2020, it's designers believe it could generate more data in one day than the entire internet in one year. Further, it will be 10,000 times more powerful than any other telescope. Given the amount of data it produces, SKA will require three times the computing power of the world's largest supercomputer in 2017.

FOR REVIEW ONLY
FOR REVIEW ONLY

Physics

The Large Hadron Collider (LHC) is a 17-mile ring filled with superconducting magnets that send protons around in opposite directions at nearly the speed of light only to have them smash into one another. It has an annual budget of over one billion dollars and cost many billions to build. The LHC is regarded as the pinnacle of modern day science.

The biggest finding from the LHC thus far has been the discovery of the Higgs Boson, a particle predicted by the standard model of physics but never shown to exist. The LHC has an amazing 150 million sensors that deliver data at an incredible 14 million times per second. Inside the accelerator there are 600 million collisions per second. Since only a few of the collisions are of interest, the LHC *only* stores about 25 petabytes of data per years. As of 2017 the analysis of the LHC data is done on a computing grid with 500,000 processors and 500 petabytes of storage.

Sources of Big Data in Business and Industry

In business and industry most large machines have sensors that monitor system components many times per second. For example,

- General Electric (GE) manufactures a gas turbine with 200 embedded sensors which generate about 600 gigabytes of data per day. One gas turbine would generate 219 terabytes of data in a year. GE is the largest producer of gas turbines in the world with more than 10,000 gas and steam turbines operating throughout the world. Assuming all these turbines have the same number of embedded sensors, the data generated by all these machines would be on the order of 2 exabytes annually.
- GE also produces aircraft engines that are generating 10 data points per second on 1000 parameters (sensors). On a flight from New York to London one of these engines would generate about 8 gigabytes of data. One model of GE aircraft engine is used on approximately 2,000 Boeing 737 aircrafts. Most commercial aircraft fly about 3000 hours per year. At two engines per aircraft, the Boeing 737 aircraft fleet generates about 92 petabytes of data per year.
- A modern commuter train's sensors will collect and send 9,000,000 data points per hour.
- A smart energy meter could send 35 gigabytes of data per day.
- Modern buildings are full of sensors that monitor sound, temperature, humidity, and motion.
- Computer logs are used to monitor and diagnose computer system problems. Even a 50 server data center will generate 100 gigabytes of log data per day.
- Worldwide there are about 100 billion credit card transactions per year. Building fraud prevention models for credit cards has become a big data problem.
- By the year 2020 there is expected to be 50 billion machines connected to the internet.

An Aircraft Engine

In industrial applications—like gas turbines, aircraft engines, and train motors—sensor data is used to determine an optimal operations strategy and to detect root

FOR REVIEW ONLY
FOR REVIEW ONLY

cause of failures and defects in near real-time. Companies will also will be using sensor data to look for correlations among variables that may signal a design improvement in the system.

In 2016, a company developed a model to forecast corn yield per acre. It was an unusual model because the model used one petabyte of satellite imagery data that was run through a cluster of 30,000 computer processors to predict average corn yields per acre for 2016. Interestingly the satellite image model predicted an average corn yield of 169 bushels per acre yield for 2016. The US Department of Agriculture (USDA) predicted 175.1 and the actual corn yield for 2016 was 178 bushels per acre. This is a case where big data modeling is not always good. But, the accuracy of this kind of modeling will undoubtedly improve.

Satellite image data is also being used to produce revenue forecast for large box retailers (e.g., Walmart, Target). Using satellite imagery, computer programs are counting cars in these retailers parking lot every day and connecting this data with quarterly revenue estimates.

FOR REVIEW ONLY

FOR REVIEW ONLY

3.4 Histograms and Other Graphical Displays of Quantitative Data

A **histogram** is a common graphical method that reveals the distribution of a set of data. Histograms are often constructed based on frequency distributions of quantitative data. Histograms look similar to bar graphs, but are used to analyze quantitative data rather than qualitative data.

Histogram

A **histogram** is a graphical representation of a frequency or relative frequency distribution. The horizontal scale corresponds to classes of quantitative data values and the vertical scale corresponds to the frequency or relative frequency of each class.

DEFINITION

Each of the classes in the frequency distribution is represented by a vertical bar whose height is proportional to the frequency of the class interval. The horizontal boundaries of each vertical bar correspond to the class boundaries. Once the frequency distribution has been calculated, all the information necessary for plotting a histogram is available. In Figure 3.4.1, the histogram is created from the frequency distribution of the heart rate data in Table 3.3.1 of the previous section.

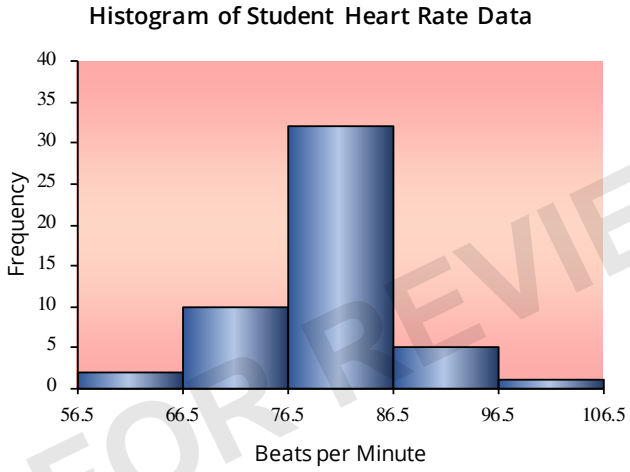


Figure 3.4.1

You can quickly see that most of the heart rates are in the third class interval from 76.5 to 86.5. The center of the data appears to be near 80 and the data appears to have a mound shape with most of the data in the middle.

Histograms are one of the more frequently used statistical tools. A histogram is not only easy to interpret; it also reveals a great deal about the structure of the data. By examining the histogram, one can determine the shape of the distribution of the data. That is, the data could be **symmetric** or **skewed**.

Technology

For instructions on how to create a histogram in Excel or Minitab, please refer to the web resource at **Tech > Graphs > Histograms**

FOR REVIEW ONLY
FOR REVIEW ONLY

Symmetric vs. Skewed

If you split the histogram of a distribution down the center, and the left and right sides of the histogram are approximately mirror images of one another, the distribution is said to be **symmetric**.

A **skewed distribution** is a nonsymmetric (or asymmetric) distribution that extends more to one side than the other. The distribution is said to be **skewed to the right** (or positively skewed) if the tail to the right of the peak of the distribution is longer than the tail to the left of the peak. The distribution is said to be **skewed to the left** (or negatively skewed) if the tail to the left of the peak of the distribution is longer than the tail to the right of the peak.

DEFINITION

Here are a few common shapes of distributions used to describe data.

Shapes of Graphs

1. **Uniform:** For data with a uniform distribution, the frequency of each class is approximately the same. Histograms of data that are distributed uniformly are rectangular in shape. An example of data that are distributed uniformly are the number of occurrences of the digits 0 through 9 used in the last four numbers of cell phone numbers. Since these numbers are generated randomly, each digit should occur approximately the same number of times.

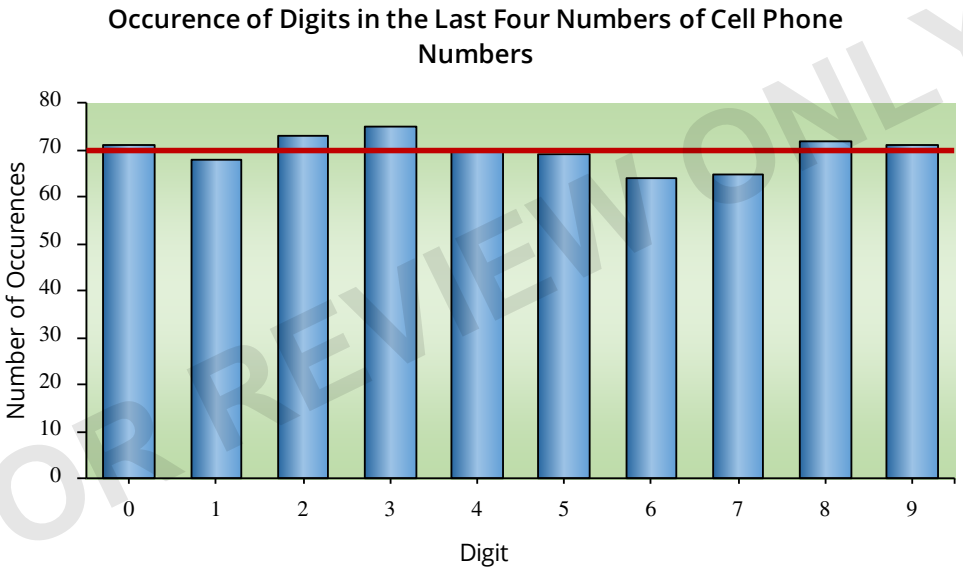


Figure 3.4.2

2. **Bell-shaped:** A symmetric graph that peaks in the center and then slopes off gradually to the left and right in relatively the same proportions is called a bell-shaped graph. A vertical line drawn through the peak of the graph would result in approximately mirror images of the data on either side. An example of data having a bell-shaped distribution are the heights of adult women.

FOR REVIEW ONLY

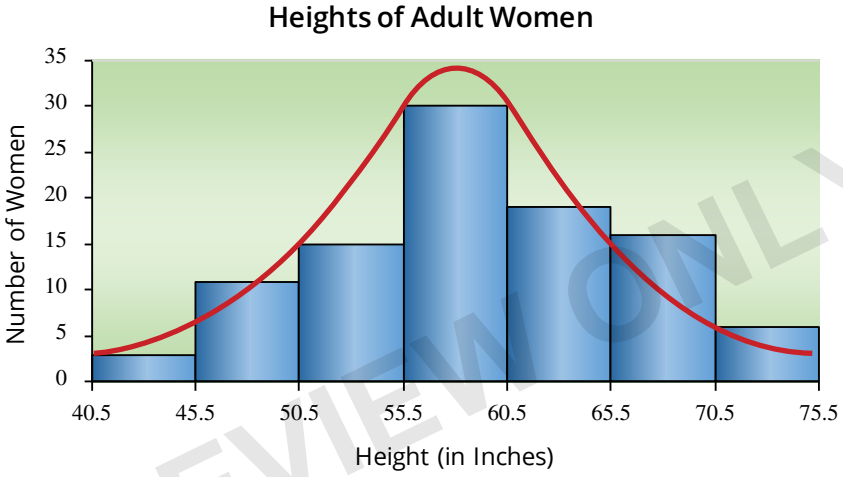


Figure 3.4.3

3. **Skewed to the Right:** A graph is said to be skewed to the right when the majority of the data fall on the left side of the distribution. The tail of the distribution extends to the right. A graph may be skewed to the right because there is an extreme value on the right side of the graph. An example of a graph that is skewed to the right is the distribution of salaries in San Francisco shown in Figure 3.4.4. The graph is skewed to the right because it is being pulled towards the extreme salaries above $200,000.

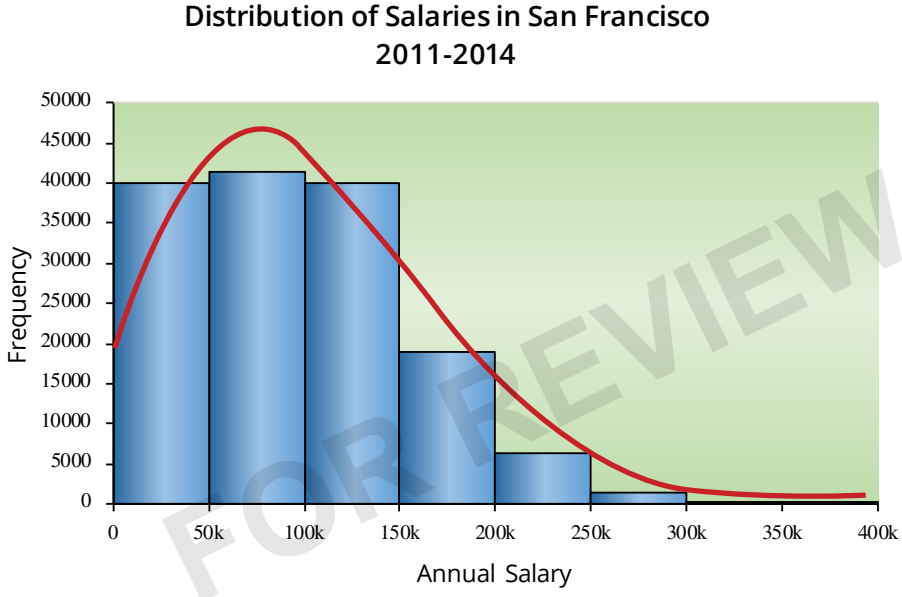


Figure 3.4.4

4. **Skewed to the Left:** We say that a graph is skewed to the left when the majority of the data fall on the right side of the distribution. This is the opposite of the previous example. A graph may be skewed to the left because there is an extreme value on the left side of the graph. An example of a graph that is skewed to the left is the graph of high school graduate GPAs shown in Figure 3.4.5.

FOR REVIEW ONLY
FOR REVIEW ONLY

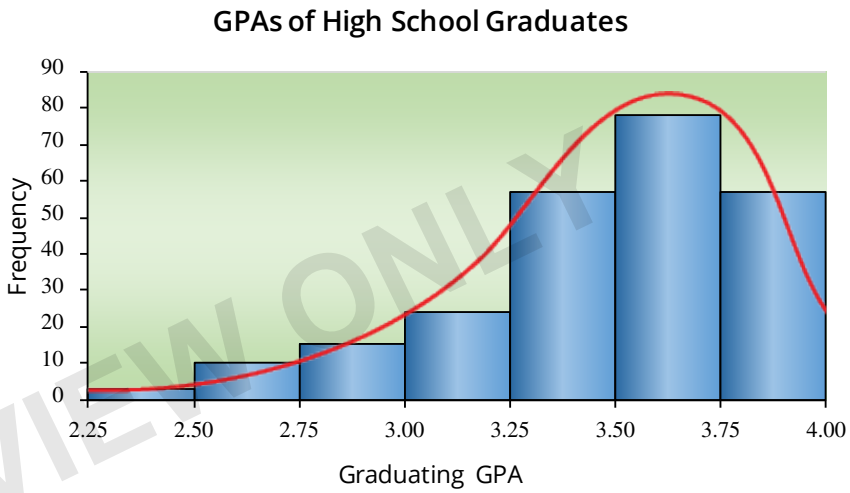
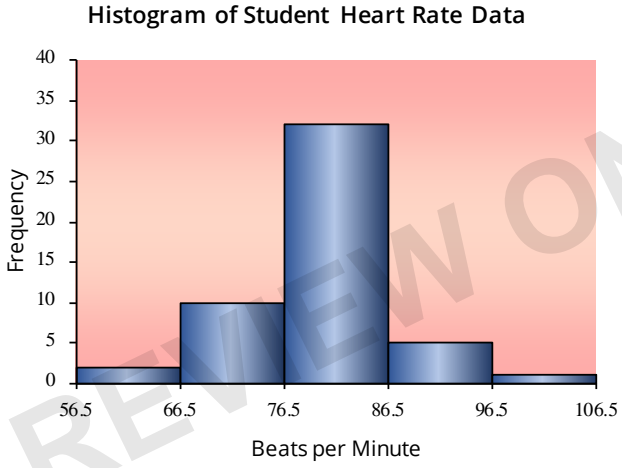


Figure 3.4.5

Example 3.4.1

Shapes of Graphs

Describe the overall shape of the student heart rate data from Figure 3.4.1.



Solution

If we draw an imaginary vertical line down the center peak of the graph, we can see that both sides of the graph are relatively mirror images of each other. The graph has a high peak in the middle and then gradually slopes off on both sides, similar to the shape of a bell. Therefore, the distribution of student heart rates is symmetrical and bell-shaped.

There are many statistical software programs that can be used to construct histograms. There may be times though, when you need to manually create a histogram for a set of data. Although there isn't a universal method for constructing a histogram, the procedure outlined in this section can help you create one.

FOR REVIEW ONLY
FOR REVIEW ONLY

Constructing a Histogram

1. First determine the number of classes to be displayed in the histogram. Typically the number of classes to use results from trial and error, but a good starting point is to use $\sqrt{n}$ or $\sqrt[3]{2n}$.

2. Determine the extreme observations in the data set, the minimum and maximum. Depending on the size of the data set, it may be a good idea to sort the observations from smallest to largest. This will make it much easier later on when trying to determine the interval in which a particular observation falls.

3. Calculate the difference between the minimum and maximum observations. (Note that this measure is called the range, which will be discussed further in Chapter 4.)

4. Calculate the class width, using the following formula. For simplicity, we usually round this number up to the next largest integer to make the calculation of the class intervals easier.

$$\text{Class Width} = \frac{\text{Maximum Value} - \text{Minimum Value}}{\text{Number of Classes}}$$

5. Calculate the upper and lower limits of each class. A good rule of thumb for determining the class intervals is to choose the lower class limit of the first class interval so that it is either 0 or a multiple of the class width, such that the smallest observation falls in the first class interval. The lower class limit of all subsequent class intervals can be found by taking the lower limit of the first class and adding successive multiples of the class width.

6. Calculate the class boundaries by subtracting 0.5 from the lower class limit and adding 0.5 to the upper class limit for each class. The class boundaries will be used to determine in which interval an observation falls in order to create a frequency distribution for the data. The rule for determining if an observation falls in an interval is as follows.

$$\text{Lower Class Boundary} < x_i \leq \text{Upper Class Boundary}$$

where x_i is observation i.

7. Tally each observation into its appropriate interval based on class boundaries. Now determine the class frequency, denoted by f_i, which is the number of observations in each interval.

8. It is a good idea to summarize the information for the histogram in a frequency distribution table or a relative frequency distribution table in order to accurately construct the histogram.

PROCEDURE

FOR REVIEW ONLY
FOR REVIEW ONLY

Example 3.4.2

A random sample of 28 applicants took a test designed to measure their aptitude for a job in sales. The resulting scores were obtained. The observations are already sorted from smallest to largest.

Aptitude Scores (%)						
38	49	53	56	58	58	60
62	66	67	69	69	71	74
75	76	77	77	77	78	78
81	82	83	84	87	88	88

Construct a relative frequency distribution for the test scores, using 6 class intervals, and construct a histogram.

Solution

To solve this problem, we will work through the procedure for constructing a histogram.

1. First, we need to determine the number of classes for the histogram. We are told in the problem to use 6 class intervals, so there will be 6 classes.
2. The second step is to identify the maximum (largest observation) and the minimum (smallest observation) from the data set. Since the data in the table are already ordered from smallest to largest, the minimum test score is 38 and the maximum score is 88.
3. Next, we calculate the difference between the minimum and maximum.

$$\text{Maximum Score} - \text{Minimum Score} = 88 - 38 = 50$$

4. Now calculate the class width.

$$\text{Class Width} = \frac{\text{Maximum Score} - \text{Minimum Score}}{\text{Number of Classes}} = \frac{50}{6} \approx 8.33$$

Since the class width is not an integer, we round up to the next largest integer, 9. Therefore, we will use a class width of 9 when constructing the histogram.

5. The next step is to calculate the upper and lower limits of each class. We want the lower limit of the first class interval to be either 0 or a multiple of the class width, and the first interval must include the smallest observation. It isn't reasonable to make the lower limit of the first class interval to be 0, since the minimum score, 38, is much larger than the class width. Therefore, it must be a multiple of the class width, which is 9. We will set the lower limit of the first class interval to be 36, which is the largest multiple of 9 that is less than the minimum value in the data. The first class interval then is 36–44, the second interval is 45–53, and so on. Therefore, the class limits are as follows.

Class Limits					
36–44	45–53	54–62	63–71	72–80	81–89

FOR REVIEW ONLY
FOR REVIEW ONLY

6. Once we have the class limits, we calculate the class boundaries by subtracting 0.5 from the lower class limit of each interval and by adding 0.5 to the upper class limit of each interval.

Class Boundaries					
35.5–44.5	44.5–53.5	53.5–62.5	62.5–71.5	71.5–80.5	80.5–89.5

7. Tally each observation to determine the class interval in which each observation falls. Remember that the observation falls in the interval if it is greater than the lower class boundary and less than or equal to the upper class boundary. This gives you the class frequencies.

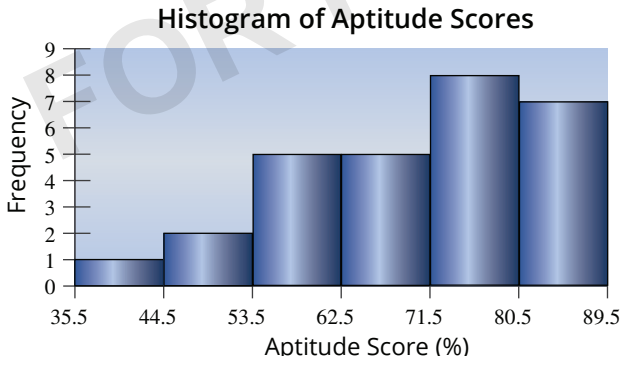
Class Frequencies					
36–44	45–53	54–62	63–71	72–80	81–89
1	2	5	5	8	7

8. The relative frequency distribution table provides a summary of the data and is used to construct the histogram.

Relative Frequency Distribution			
Class Limits	**Class Boundaries**	**Frequency** (f_i)	**Relative Frequency**
36–44	35.5–44.5	1	0.0357
45–53	44.5–53.5	2	0.0714
54–62	53.5–62.5	5	0.1786
63–71	62.5–71.5	5	0.1786
72–80	71.5–80.5	8	0.2857
81–89	80.5–89.5	7	0.2500
	Total	**28**	**1.0000**

The sum of the frequency column should sum to the total number of observations in the data set. Also, the sum of the relative frequency column should be approximately equal to 1. If the sum of the relative frequency column is slightly different from 1, it is probably due to rounding.

9. Finally, construct the histogram.



FOR REVIEW ONLY

Stem-and-Leaf Plot

The **stem-and-leaf plot** (or **stemplot**) is a hybrid graphical method. The plot is similar to a histogram, but the data remains visible to the user. Like all graphical displays, the stem-and-leaf plot is useful for both ordering and detecting patterns in the data. It is one of the few graphical methods in which the individual data is not lost in the construction of the graph. As the name implies, there is a "stem" to which "leaves" will be attached in some pattern.

Stem-and-Leaf Plot

The **stem-and-leaf plot** is a graph representing quantitative data that separates each data value into two parts: the stem and the leaf.

DEFINITION

Consider the following data: 97, 99, 108, 110, and 111. If we are interested in the variation of the last digit, the stem and leaves are as shown in Table 3.4.1 and displayed in Figure 3.4.6. The leaves in this case are the *ones* digit and the stems are the *tens and hundreds* digits. All of the data values that have common stems are grouped together, and their leaves branch out from the common stem.

Table 3.4.1 - Data, Stems, and Leaves

Data	Stem	Leaf
97	09	7
99	09	9
108	10	8
110	11	0
111	11	1

Stem-and-Leaf Plot

Stem	Leaf
09	7 9
10	8
11	0 1
Key: 10	8 = 108

Figure 3.4.6

Notice the key at the bottom of the stem-and-leaf plot to show how the data is to be read from the plot. It is also important to keep the spacing between the leaves consistent.

If we are interested in the variation of the last two digits, the stem and leaves are shown in Table 3.4.2 and displayed in Figure 3.4.7. The leaves in this case are the *tens and ones* digits and the stems are the *hundreds* digits. Again all of the data values that have common stems are grouped together, and their leaves branch out from the common stem.

FOR REVIEW ONLY
FOR REVIEW ONLY

Table 3.4.2 - Data, Stems, and Leaves		
Data	**Stem**	**Leaf**
97	0	97
99	0	99
108	1	08
110	1	10
111	1	11

Stem-and-Leaf Plot

Stem	**Leaf**
0	97 99
1	08 10 11
Key: 1	08 = 108

Figure 3.4.7

Deciding which part to make the stem and which part to make the leaf depends on the focus or purpose of the analysis. Sometimes the choice of stem and leaf is easy. With the heart rate data, using the *tens* digit as the stem will break the data into four classes.

Stem-and-Leaf Plot	
Stem	**Leaf**
6	7 9 8 5 2
7	7 9 4 8 9 7 3 7 9 0 2 7 0 5 4 9
8	4 4 2 1 6 3 3 4 2 0 1 0 3 2 0 2 1 5 4 0 3 0 5 7 8
9	0 4 3 8
Key: 9	0 = 90 bpm

Figure 3.4.8

It is often advantageous to place the leaves of a stem-and-leaf plot in numerical order. Ordering the leaves in the heart rate data pictured above gives us the following.

Ordered Stem-and-Leaf Plot	
6	2 5 7 8 9
7	0 0 2 3 4 4 5 7 7 7 7 8 9 9 9 9
8	0 0 0 0 0 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 6 7 8
9	0 3 4 8
Key: 9	0 = 90 bpm

Figure 3.4.9

FOR REVIEW ONLY
FOR REVIEW ONLY

The stem-and-leaf plot in Figure 3.4.9 with only four stems isn't very revealing as to the shape of the distribution. The data seem clumped into the middle two stems 7 and 8. In order to spread the data out a little and get a better picture of the distribution, we can use **split stems**. For example, for data in the 70-79 range, we will place data ranging from 70 to 74 into one stem and the data ranging from 75 to 79 in another. This splits the ten digits from 0 to 9 into two groups of 5. Using split stems we get a much more revealing picture of the distribution of the heart rate data.

Ordered and Split Stems	
6	2
6	5 7 8 9
7	0 0 2 3 4 4
7	5 7 7 7 7 8 9 9 9 9
8	0 0 0 0 0 1 1 1 2 2 2 2 2 3 3 3 3 4 4 4 4
8	5 5 6 7 8
9	0 3 4
9	8
Key: 9	0 = 90 bpm

Figure 3.4.10

If you look at this stem-and-leaf plot closely, you will see that it resembles a histogram turned on its side. Each stem of the plot plays a similar role to a class in a histogram. The leaves represent the frequency or height of the bars in a histogram, which is why it's important to keep the spacing of the leaves the same. The biggest advantage to using a stem-and-leaf plot over a histogram to display a set of data, is the fact that you don't lose the individual data in a stem-and-leaf plot. We will find this advantageous in the next chapter when we talk about descriptive statistics.

Stem-and-leaf plots lose their advantages, however, when the data set is large or the data contains a wide range of data values. Constructing a stem-and-leaf plot by hand can also become tedious when there are a large number of observations. Luckily, most statistical software programs, like Minitab, provide a way to construct a stem-and-leaf plot with minimal effort.

From the ordered stem-and-leaf plot using split stems in Figure 3.4.10, note that the distribution of the heart rate data appears symmetric with a large concentration of heart rates between 80 and 84. You can also readily see that the most frequently occurring value in the data is 80.

Stem-and-leaf plots can be used for many types of data, even data containing decimals. Let's suppose that you have 50 data values ranging from 20.4 to 29.7. The most likely choice of stems might be the integers 20 to 29 with the number appearing after the decimal point as the leaf. For basketball scores ranging from 78 to 122 you could use the stems 07, 08, 09, 10, 11, and 12 with one-digit leaves. For data on Per Capita GNP ranging from $160 to $980, it wouldn't make sense to have two-digit stems ranging from 16 to 98 unless you had a large amount of data, which would then make the stem-and-leaf plot impractical. For sparse data sets with a large range of values, like the GNP data, you could use one digit stems and two-digit leaves. The spacing of the leaves would be extremely important in this case to be able to interpret the shape of the distribution.

FOR REVIEW ONLY
FOR REVIEW ONLY

Because of the wide range of data that stem-and-leaf plots can accommodate, it is very important to use a key at the bottom of the plot to show how the data is read. If the data are measurements, then the units should be included in the key as well.

Example 3.4.3

Construct a stem and leaf plot that compares the annual homerun production Babe Ruth and Barry Bonds hit per season.[1]

Barry Bonds						
Year	1986	1987	1988	1989	1990	1991
HR	16	25	24	19	33	25
Year	1992	1993	1994	1995	1996	1997
HR	34	46	37	33	42	40
Year	1998	1999	2000	2001	2002	2003
HR	37	34	49	73	46	39
Year	2004	2005	2006	2007		
HR	45	5	26	28		

Babe Ruth						
Year	1914	1915	1916	1917	1918	1919
HR	0	4	3	2	11	29
Year	1920	1921	1922	1923	1924	1925
HR	54	59	35	41	46	25
Year	1926	1927	1928	1929	1930	1931
HR	47	60	54	46	49	46
Year	1932	1933	1934	1935		
HR	41	34	22	6		

Solution

Homeruns Hit per Season: Babe Ruth vs Barry Bonds		
Ruth		**Bonds**
0 4 3 2 6	0	5
1	1	6 9
9 5 2	2	5 4 5 6 8
5 4	3	3 4 7 3 7 4 9
1 6 7 6 9 6 1	4	6 2 0 9 6 5
4 9 4	5	
0	6	
	7	3
Key: 0	6	1 = 60 HR for Ruth, 61 HR for Bonds

FOR REVIEW ONLY
FOR REVIEW ONLY

Note that in order to compare the two sets of data more readily, we needed to use the same stems for each. This was accomplished in this example by sharing the common stems and creating one plot instead of two separate ones. This is commonly referred to as a **side-by-side (or back-to-back) stem-and-leaf plot**.

The Ordered Array

An **ordered array** is a listing of all the data in either increasing or decreasing magnitude. Data listed in increasing order are said to be listed in **rank order**. If listed in decreasing order, they are listed in **reverse rank order**. Listing the data in an ordered way can be very helpful. It allows you to scan the data quickly for the largest and smallest values, for large gaps in the data, and for concentrations or clusters of values.

Example 3.4.4

The personnel records for a clothing department store located in the local mall are examined, and the current ages for all employees are noted. There are 25 employees, and their ages are listed in the following table. It is desired that their ages be placed in rank order.

Ages of Employees												
32	21	24	19	61	18	18	16	16	35	39	17	22
21	60	18	53	18	57	63	28	20	29	35	45	

Solution

Ages (Ordered)												
16	16	17	18	18	18	18	19	20	21	21	22	24
28	29	32	35	35	39	45	53	57	60	61	63	

It is always a good idea to get a look at the ordered array of the data early in your analysis. Examining the ranked data produces a good intuitive sense for the data. Looking at the ordered array, it is evident that over half of the employees are younger than 25 and only three employees are within 5 years of retirement. We can also easily see that the youngest employee is 16 years old and the oldest employee is 63 years old. Ordering the data makes it easy to analyze the data quickly and easily.

Technology

In Microsoft Excel, the Sort tool allows the user to sort any number of data values in ascending or descending order. To learn how to do this with Excel , or with other tools, please refer to the web resource at **Tech > Data Manipulation > Sorting**.

Ordered arrays are easy to create. Virtually all statistics, spreadsheet, and database programs enable the user to quickly sort the data in ascending or descending order. If a spreadsheet or database program is not available, a stem-and-leaf plot can be helpful in sorting the data.

Dot Plots

A **dot plot** is a graph where each data value is plotted as a point (or a dot) above a horizontal axis. If there are multiple entries of the same data value, they are plotted one above the other. Dot plots are useful when you are interested in where the data are clustered and which values occur most often.

FOR REVIEW ONLY

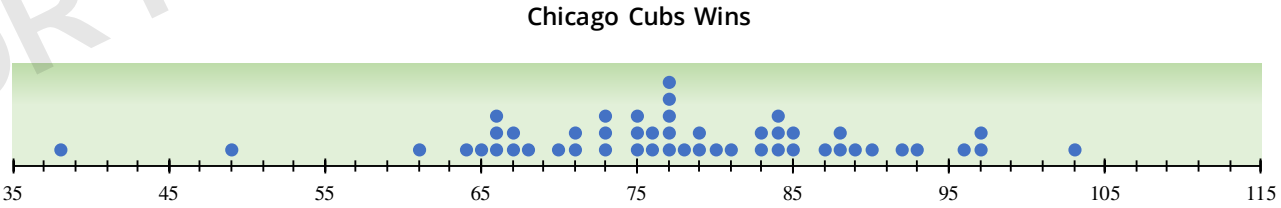
Example 3.4.5

The following table contains the number of wins by baseball's Chicago Cubs for a recent 50 year period. Use this data to construct a dot plot.

Wins by the Chicago Cubs (1967-2016)									
61	85	67	68	78	76	73	81	85	87
66	97	88	90	84	77	71	79	77	84
73	83	89	67	49	93	96	80	66	92
97	75	79	65	73	77	77	64	75	84
103	71	66	88	76	77	70	38	75	83

Solution

We plot each data value on the axis. For values where there are multiple entries, such as for 77, we stack the points on top of one another.

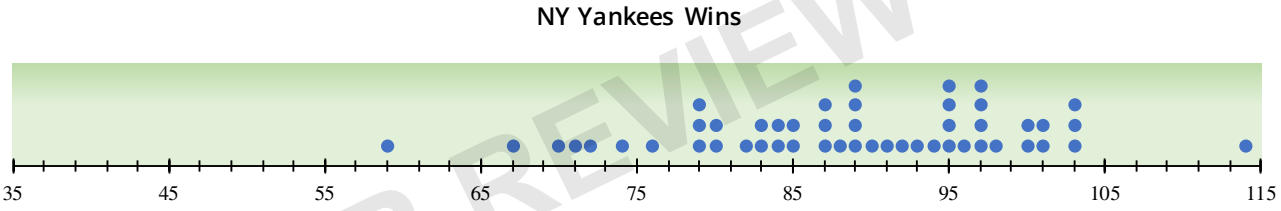


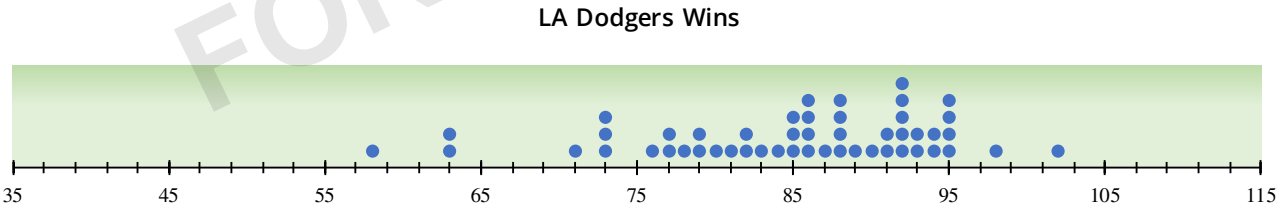
From the dot plot we see that most values occur between 64 and 97 and the value that occurs most frequently is 77.

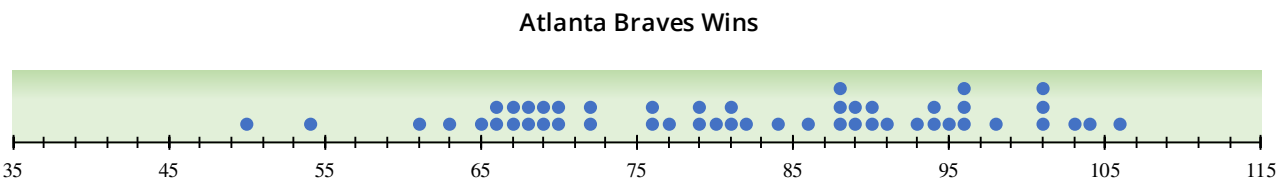
Technology

For instructions on how to create a dot plot in Excel or Minitab, please refer to the web resource at **Tech > Graphs > Dot Plot.**

By way of comparison, the dot plots for a few other teams are given below.







FOR REVIEW ONLY

Time Series Plots

Recall our discussion of time series data from Chapter 2. When looking at time series data, we are typically interested in whether the process is stationary or nonstationary. To determine whether a time series is stationary or nonstationary, we use a **time series plot**. A time series plot graphs quantitative data using time as the horizontal axis. Most often, a time series plot comes in the form of a **line graph**, which connects consecutive points over time with line segments. You have already seen some line graphs as examples of time series plots in Chapter 2.

A good example of time series data is the census measurements taken every decade since 1790. The population of the United States has been growing steadily since the first census (as shown in Table 3.4.3)[2]. Interestingly, the change in population between 2000 and 2010 was 27.3 million, which is greater than the entire population of the United States in 1850.

Table 3.4.3 - US Census Data

Year	Population	Population (in millions)
1790	3,929,214	3.9
1800	5,308,483	5.3
1810	7,239,881	7.2
1820	9,638,453	9.6
1830	12,866,020	12.9
1840	17,069,453	17.1
1850	23,191,876	23.2
1860	31,443,321	31.4
1870	38,558,371	38.6
1880	50,189,209	50.2
1890	62,979,766	63.0
1900	76,212,168	76.2
1910	92,228,496	92.2
1920	106,021,537	106.0
1930	123,202,624	123.2
1940	132,164,569	132.2
1950	151,325,798	151.3
1960	179,323,175	179.3
1970	203,211,926	203.2
1980	226,545,805	226.5
1990	248,709,873	248.7
2000	281,421,906	281.4
2010	308,745,538	308.7

In a line graph, time is always labeled on the horizontal axis, with the variable being measured labeled on the vertical axis. Points are then plotted for each time period, and a line is drawn that connects each consecutive point. In Figure 3.4.11, each of the points is plotted with the year on the horizontal axis and the corresponding population (in millions) on the vertical axis.

FOR REVIEW ONLY
FOR REVIEW ONLY

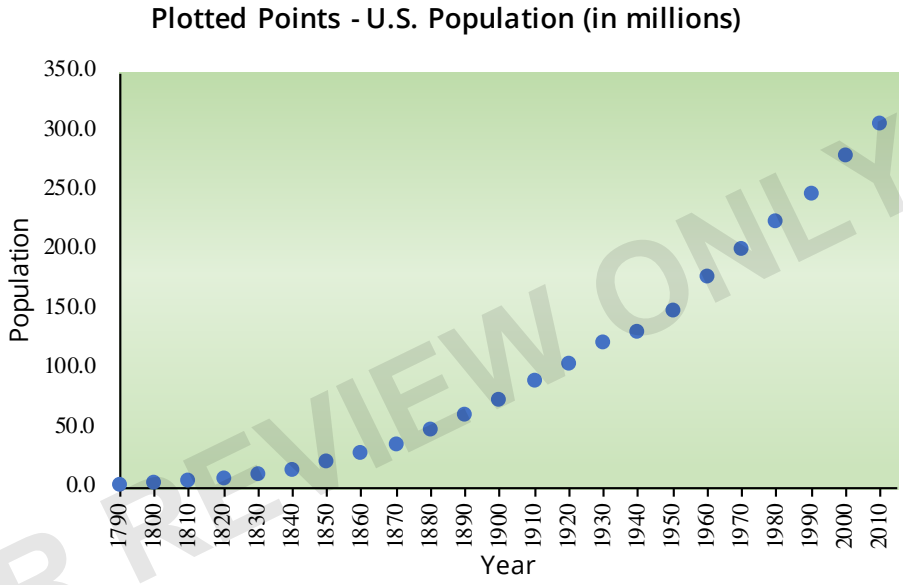


Figure 3.4.11

In Figure 3.4.12, a line segment connects consecutive points in the time series to give a line graph. From the line graph we can easily see that the series is nonstationary, and that there is an upward trend in the data.

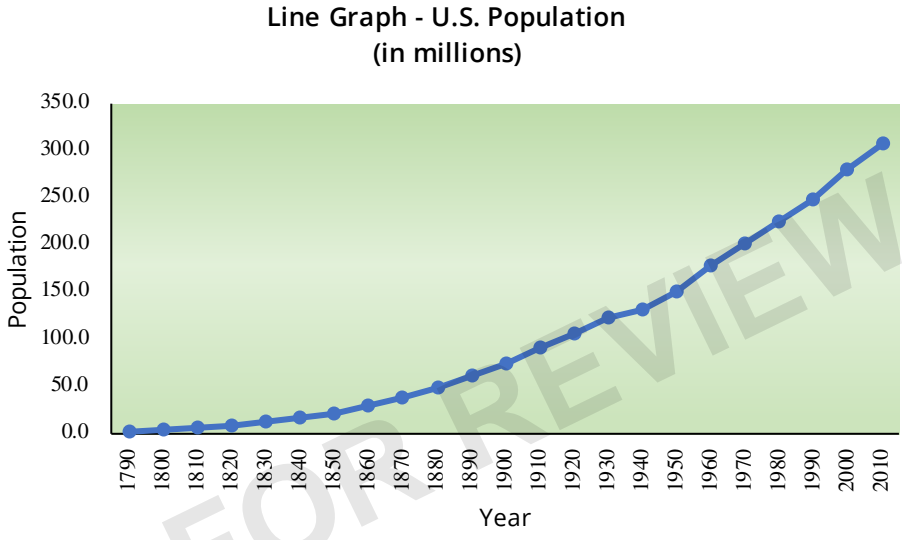


Figure 3.4.12

Does it make sense to draw a histogram for a nonstationary time series? We could create a histogram of the population data we have been examining, but it would not reveal anything very interesting. For a nonstationary time series (a time series with a trend), a histogram is usually not warranted.

It is not uncommon to see time series graphs with multiple sets of data plotted. It is also not uncommon to see two vertical axes used if the data sets have different scales. In Figure 3.4.13, both the sea level change and the global temperature change since 1901 are shown on the same graph. The vertical scale on the left refers to the change in sea level. The vertical scale on the right refers to the change in global

FOR REVIEW ONLY
FOR REVIEW ONLY

temperature. By plotting both sets of data on the same graph, we can see the obvious trend that the two sets of data share.

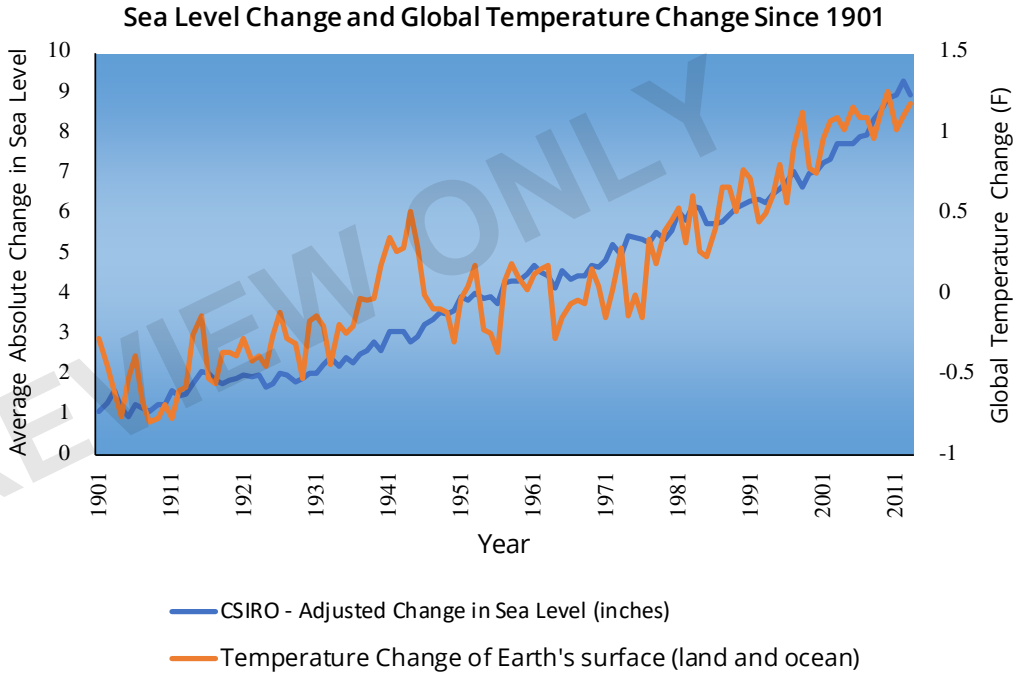


Figure 3.4.13

Geospatial Graphs

For centuries, humans have used maps to portray information about specific places or regions. Today, it is very common to see data that is associated with some type of geographic location such as zip codes, county codes, states, countries, or even specific longitudes and latitudes. This association allows us to layer the data values on top of a map of a certain area, or satellite image, ultimately providing us with a greater understanding of how different regions compare to one another based on a variety of different variables.

Choropleth maps are one type of geospatial graph that is very popular in modern society, although their use dates back to the early 19th century. Choropleth maps utilize shades of color to measure a variable across several pre-defined areas, and those areas can be anywhere on a map as long as they are pre-defined and have data associated with them. Choropleth maps are similar to heat maps; however, heat maps use the given data to define regions whereas choropleth maps come with pre-defined regions with which data is already associated. An example of a heat map would be weather radar, where the shaded region is dynamically defined by the amount and type of precipitation in an area rather than by state or county lines.

Suppose we are interested in visualizing obesity in different regions of the United States. The Center for Disease Control (CDC) provides a county level data set containing the number of people in each county who are above the obesity threshold. For our inquiry, we are only interested in two columns of the full data set: the county code, or FIPS Code, and the number of people in a county who are obese in the year 2016. A preview of the data we will use is in Table 3.4.4.[3]

Data

To download the full US County data set, please go to the web resource at **Data > US County Data.**

FOR REVIEW ONLY

Table 3.4.4 - Number of Obese Adults by US County	
FIPS Code	**Number of Obese Adults 2016**
1001	15,884
1003	47,117
1005	10,653
1007	7,611
1009	17,347
. . .	

We want to plot this data onto a map of the United States. Each FIPS Code can be used to determine the geographic boundaries of each county, and then we will shade each county region on a map of the U.S. based on the number of adults who are obese in that area. A lighter shade will indicate that the number of obese adults is smaller, and a darker shade will indicate that the number is high.

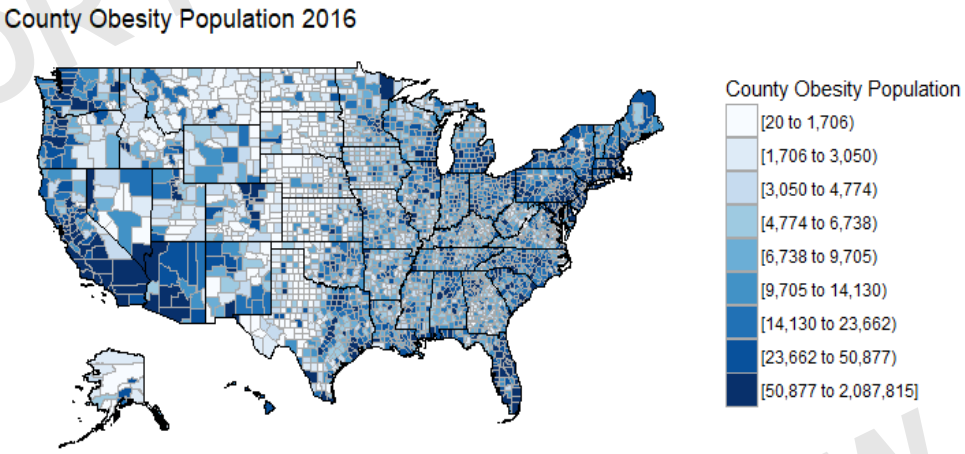


Figure 3.4.14

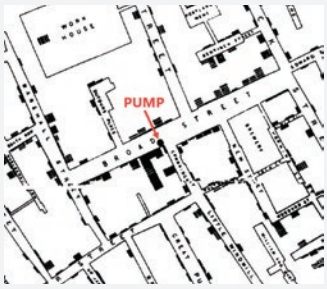
Figure 3.4.14 suggests that the majority of obese adults in the United States reside in the Southwest and Northeast, and that the Midwest and Central Plains areas don't have very many obese adults at all. However, we must now use our intuition to determine an explanation for this distribution. By looking at the legend to the right of the map, we can see that there is an extremely wide range from the lowest bin to the highest. This is because some counties contain millions of residents, while others contain as few as a couple hundred. So, for many of the rural counties in the center of the country, even if every single person in the county was obese, they still would not compare with, say, Los Angeles County which has a total population of nearly 10 million.

For the reasons explained above, it is inaccurate to rank the counties simply by the total number of obese adults, and doing so will lead to faulty or incomplete conclusions. In order to get a better understanding of how the counties truly compare to one another, we need to normalize the values being graphed. Since our variable of interest in Figure 3.4.14, number of obese adults, is a subset of the total population in each county, we can use that variable to find the percentage of each county population that is obese. This can be done by dividing the number of people in a county who are obese by the total county population, however, the US County data set provides

Technology

Choropleth maps can be usful when comparing locations. To learn how to create a county level choropleth map, please refer to the web resource at **Tech > Graphs > Choropleth Map (County).**

FOR REVIEW ONLY
FOR REVIEW ONLY



Absence of Evidence is Not Evidence of Absence

During the London cholera outbreaks of the mid-1800s, thousands of people died within a relatively short period. At the time, the prevailing theory regarding how cholera was spread was called the miasma theory. It stated that the disease was spread through "bad air" that emanated from rotting organic matter. However, Dr. John Snow suspected that unsanitary water from the River Thames was the true culprit. Unfortunately, germ theory had not been developed yet, so Dr. Snow didn't fully understand how the alternative transmission method worked. In 1854, Dr. Snow utilized sampling and data visualization to illustrate that most of the cholera outbreaks happening at the time were occuring in houses that were close to the water pump on Broad Street. Still, the sceptics endured. However, even though his examination of the water was absent of evidence for harmful microbes, that does not mean that the microbes themselves were absent. Over a decade later, Louis Pasteur would officially propose germ theory, vindicating the work of Dr. Snow and subsequently saving millions, if not billions, of lives in the years that were to follow.

the percentages for us. A preview of the updated data set, with the new variable of interest, is shown in Table 3.4.5.

Table 3.4.5 - Percentage of Obese Adults by US County

FIPS Code	Percentage of Obese Adults 2016
1001	30.5
1003	26.6
1005	37.3
1007	34.3
1009	30.4
. . .	

Using our new data, we generate the following map.

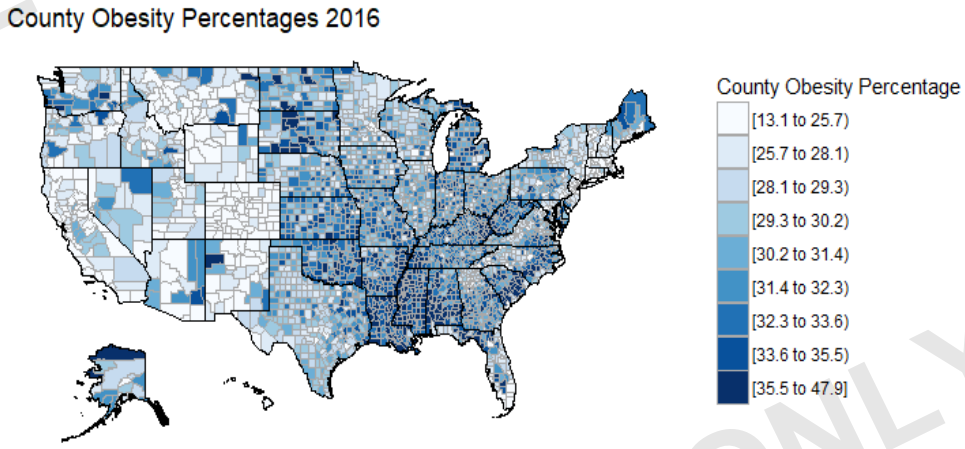


Figure 3.4.15

As you can see, Figure 3.4.15 is substantially different from Figure 3.4.14, and comes along with an entirely new set of conclusions. If we had used Figure 3.4.14 to come to a conclusion, we would have assumed that the Southwest and Northeast regions are the areas in the country that struggle with obesity the most. However, once we normalized our obesity variable by making it a ratio of total county population, Figure 3.4.15 seems to suggest that it is actually the Southeast that struggles with obesity the most. The data shows that the Southeast has a higher percent of adults who are obese relative to its total population than the Northeast and Southwest do. This is a prime example of why it is necessary to account for sample size when analyzing data.

Geospatial graphs allow us to gain new insights about the areas we live in, but they also drive us to ask new questions that nobody else has thought to ask before. Through this process of geospatial exploratory analysis, we are able to solve many problems that have eluded us in the past.

PREVIEW ONLY FOR REVIEW ONLY

3.4 Exercises

Basic Concepts

1. What is the main characteristic of data that a histogram reveals?
2. Describe the type of data that could be usefully described with a histogram.
3. True or false: A frequency distribution contains all of the information needed to construct a histogram.
4. List the important features to look for when studying a histogram.
5. Explain why the stem-and-leaf display is sometimes called a "hybrid graphical method."
6. Identify the advantages of a stem-and-leaf display.
7. Consider the following data value: 39. What would be the stem and leaf for this value if we identified the stem as the tens digit? What would be the steam and leaf if we identified the stem as the hundreds digit?
8. When constructing a stem-and-leaf display, how do you determine which part to make the stem and which part to make the leaf?
9. What is an ordered array?
10. What are some advantages of the ordered array?
11. What is a dot plot?
12. What are some advantages of using a dot plot?
13. How can the most frequently occurring value be identified by studying a dot plot?
14. Why is it important to plot time series data?
15. The time variable is always graphed on which axis?
16. Identify a variation on a time series plot that can make the data more visually interesting.

Exercises

17. The closing prices (in dollars) for selected stocks trading on the New York Stock Exchange and NASDAQ Exchange were as follows.

Closing Prices	
Stock	**Closing Price ($)**
Citigroup (C)	73.61
Pfizer (PFE)	36.16
Herbalife (HLF)	71.94
JP Morgan Chase (JPM)	101.01
Intel (INTC)	40.78
WalMart (WMT)	88.48

FOR REVIEW ONLY
FOR REVIEW ONLY

Closing Prices	
Stock	**Closing Price ($)**
Microsoft (MSFT)	78.63
PepsiCo (PEP)	110.07
General Motors (GM)	45.12
Verizon Communications (VZ)	48.64
Southwest Airlines (LUV)	57.15
Sprint (S)	7.10
Alibaba (BABA)	176.14
International Business Machines (IBM)	153.50

a. Construct a frequency distribution for the closing prices.

b. Construct a histogram for the closing prices.

18. Using the San Francisco Salaries - 2014 data set from the web resource, create a histogram and frequency distribution for the variable TotalPayBenefits and use them to answer the following questions. For the histogram and frequency distribution, use $30,000 for the minimum value and $480,000 for the maximum value. Use bin widths of $50,000.

Data

Data > San Francisco Salaries 2014

a. What is the level of measurement of the variable?

b. How many of the laborers earn more than $130,000 per year?

c. What percent of the laborers earn at most $130,000 per year?

d. What percent of the laborers earn more than $230,000 per year?

19. A chemist is interested in knowing the amount of alcohol contained in American-brewed beers. To study this, the chemist uses data containing information about several different kinds of American-brewed beers, and evaluates the alcohol by volume for each. Using the Beers and Breweries data set from the web resource, perform the following:

Data

Data > Beers and Breweries

a. Construct a frequency distribution for the alcohol by volume (ABV) variable. Use 0.001 for the minimum value and 0.130 for the maximum value. Use bin widths of 0.010.

b. Construct a relative frequency distribution for the ABV. Round the relative frequencies to four decimal places.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the alcohol by volume in American-brewed beers which you were able to ascertain by examining the distributions and the histogram.

20. Consider the assets (in billions of dollars) of the 10 largest commercial banks listed in the following table.

Assets (Billions of Dollars)				
216.9	138.9	115.5	110.3	103.5
98.2	76.4	64.0	53.5	49.0

FOR REVIEW ONLY
FOR REVIEW ONLY

a. Construct a frequency distribution for the assets (in billions of dollars) of the 10 largest commercial banks.

b. Construct a relative frequency distribution for the assets (in billions of dollars) of the 10 largest commercial banks.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the assets (in billions of dollars) of the 10 largest commercial banks which you were able to ascertain by examining the distributions and the histogram.

21. Fifty hospitals in a western state were polled as to their basic daily charges for a semi-private room. The results are listed in the following table, rounded to the nearest dollar.

Daily Charges for Semi-Private Rooms (Dollars)									
125	135	148	156	248	215	156	148	135	149
178	156	135	125	214	256	258	265	156	148
123	147	189	199	189	248	215	259	158	235
268	269	158	198	147	258	269	239	288	199
179	179	189	169	258	178	257	249	259	259

a. What level of measurement does the data possess?

b. Construct a stem-and-leaf display for the data using the tens digits as the stems.

c. Comment on the shape of the distribution.

22. The data in the following table are the toxic emissions (in thousands of tons) for 10 states in the United States.[4]

Toxic Emissions (Thousands of Tons)									
206	147	441	128	127	133	422	152	114	134

a. Construct a stem-and-leaf display for the data using the hundreds digits as the stems.

b. Comment on any information about the toxic emissions (in thousands of tons) of the 10 states that you were able to ascertain by examining the stem-and-leaf display.

23. Consider the following highway miles per gallon for 19 selected models of mini-compact, sub-compact, and compact cars.

Miles per Gallon									
26	46	36	31	28	28	27	38	42	36
37	33	23	29	37	34	29	40	28	

a. Construct a stem-and-leaf display for the data.

b. Comment on any information about the highway mpg of the selected models which you were able to ascertain by examining the stem-and-leaf display.

FOR REVIEW ONLY
FOR REVIEW ONLY

24. An instructor is interested in comparing exam scores for fraternity and non-fraternity males in her class. Meaningful comparisons between two sets of data can be made using a side-by-side stem-and-leaf display. To illustrate this, note the following display summarizing the scores.[5]

Leaf (Non-Fraternity)	Stem	Leaf (Fraternity)
	0	9
2	1	4 0 8
	2	5 7 9 4 5 5 1
3 9	3	2 6 6 9 7 7 3 2 1 6 0
	4	2 7 5
5 6 4 8 9 9 0 2	5	4 7 6 7
4 4 7 8 1 0 3 2 2 6 8 9	6	6 8 9 9 5
5 4 7 8 4 3 8 8 9 1	7	3 4 2 7 8 6 7 4 3
2 9 7 4	8	4 5 3 8 9 9 6 4 2 1 1 4 5
4 2	9	4 3 5 1 6 7 7 0 3

a. What level of measurement does the data possess?

b. Based upon the stem-and-leaf display, compare the two groups. Think of the several ways in which this can be done.

c. Suppose that 60% is considered a passing score on the exam. What percent of the fraternity students passed the exam? Non-fraternity students?

d. If someone scores 90 or higher on the exam, they will be exempt from taking the next exam. What percent of the fraternity students will be exempt from taking the next exam? Non-fraternity students?

25. Microsoft's consumer PC sales growth for the last 16 quarters are listed in the following table. Examine the data (sales growth, in percentages) and answer the following questions.

PC Sales Growth			
Quarter	Sales Growth (%)	Quarter	Sales Growth (%)
1	20	9	20
2	24	10	19
3	22	11	33
4	19	12	37
5	23	13	24
6	27	14	10
7	16	15	0
8	10	16	−4

a. Construct an ordered array of the data in rank order.

b. What conclusions can you make based on the ordered array?

FOR REVIEW ONLY
FOR REVIEW ONLY

26. *Fortune* magazine publishes a list of the top 100 best companies to work for. For the top 10 companies on this list, the average annual employee salaries are given in the following table (in thousands of dollars).

Average Salaries (Thousands of Dollars)									
121	122	136	74	118	101	114	61	95	132

a. Construct a stem-and-leaf display for the data using the tens digits as the stems.

b. Comment on any information about the average annual salaries (in thousands of dollars) of the top 10 companies which you were able to ascertain by examining the stem-and-leaf display.

c. Construct an ordered array of the average annual salaries in rank order.

d. Does the ordered array provide any additional insight into the nature of the data?

27. Use the table of data below to complete the following.

a. Construct a dot plot of the data.

b. Which data value occurs most often?

23	19	15	20	17
16	18	14	23	22
19	23	19	16	25
17	20	21	23	24

28. Listed in the following table is the number of passing attempts per game by super bowl champion Aaron Rodgers in the 2010 NFL season.

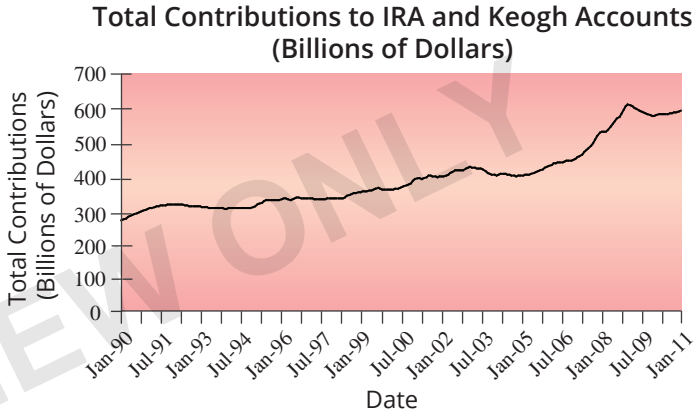
a. Construct a dot plot of the data.

b. Which data value occurs most often?

Passing Attempts by Aaron Rodgers				
31	29	45	17	46
33	34	34	34	31
35	30	11	37	28

FOR REVIEW ONLY
FOR REVIEW ONLY

29. The following line graph displays the total IRA and Keogh accounts (in billions of dollars) in the U.S., charted from June 1990 to June 2011.[7]

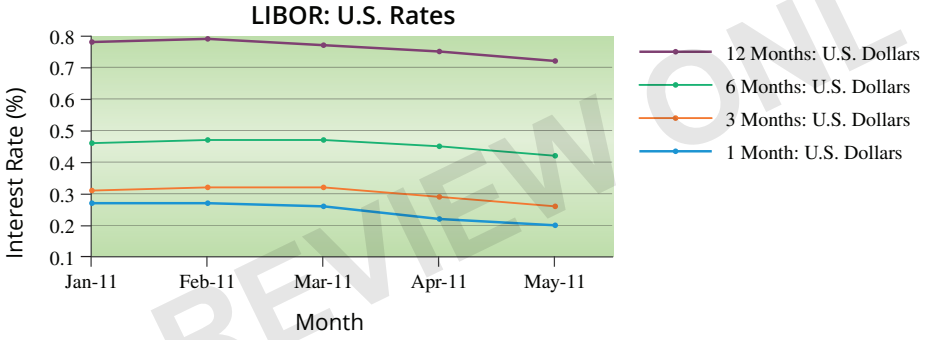


a. What conclusions can you make regarding the total contributed to the accounts?

b. Are the data time series data?

c. If the data are time series data, is the series stationary or nonstationary?

30. The following chart contains LIBOR (which stands for London Interbank Offered Rate) data for January 2011 through May 2011. LIBOR is the average interest rate that banks in London charge when lending funds to other banks. The line graphs in the figure represent 1 month, 3 month, 6 month, and 12 month interest rates.[8]



a. Examine the chart and discuss the data. What conclusions can you make?

b. If the data are time series data, is it a stationary or nonstationary time series? Explain your reasoning.

31. The Gallup Poll frequently obtains responses to the question, *At the present time, do you think religion as a whole is increasing its influence on American life or losing its influence?* The percent of the respondents who answered "increasing" is given below for various polls.

FOR REVIEW ONLY
FOR REVIEW ONLY

Survey Responses										
Year	2001	1995	1992	1991	1990	1988	1986	1984	1982	1980
Percent	71	38	27	27	33	36	48	42	41	35
Year	1978	1977	1975	1974	1970	1969	1968	1965	1962	1957
Percent	37	37	39	31	14	14	19	33	45	69

a. What level of measurement do the responses to the question possess?

b. Construct a time series plot for the data.

c. What conclusions can you make from the plot?

32. The following table gives the number of immigrants (in thousands) and the average annual immigration rate per 1000 people in the U.S. population for the decade ending in the year given.

Annual Immigration (per 1000 People)											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Number	3688	8795	5736	4107	528	1035	2515	3322	4493	7338	9095
Rate	5.3	10.4	5.7	3.5	0.4	0.7	1.5	1.7	2.1	2.9	3.2

a. What levels of measurement do the three variables in this exercise possess?

b. Construct a time series plot of the number of immigrants per decade.

c. Find the percent change in the number of immigrants from the decade ending in 1900 to the decade ending in 2000.

d. Find the percent change in the average annual immigration rate per 1000 people in the U.S. population from the decade ending in 1900 to the decade ending in 2000. Compare your answer to that which you obtained in part **c.** Can you explain why these answers are different?

33. For each set of data described below, discuss the most likely shape of its distribution.

a. The weights of the defensive linemen on football teams in the Big Ten Conference.

b. The math scores on the ACT of a large randomly selected group of high school seniors.

c. The lengths of the pregnancies of a group of gorillas being studied in the wild.

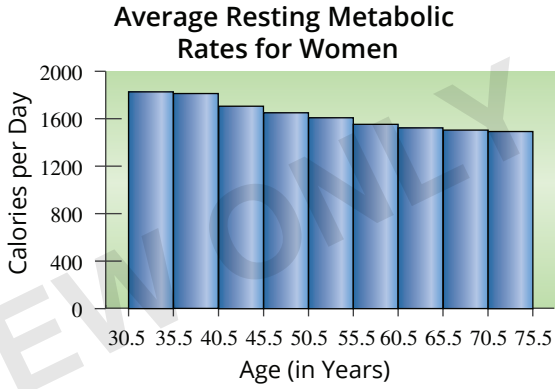
d. The last four digits used to generate telephone numbers.

e. The income levels of a group of professional baseball players.

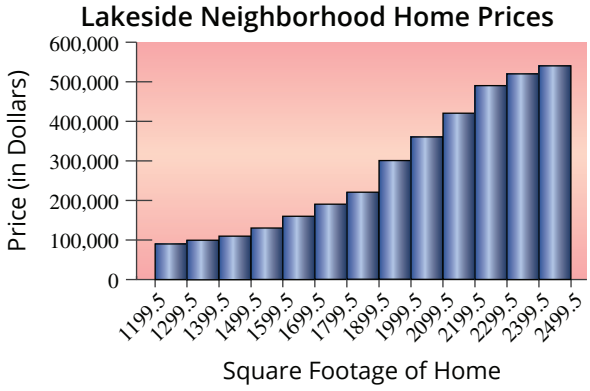
FOR REVIEW ONLY
FOR REVIEW ONLY

34. For each set of data displayed below, choose which of the four shapes defined in this section best describes the distribution.

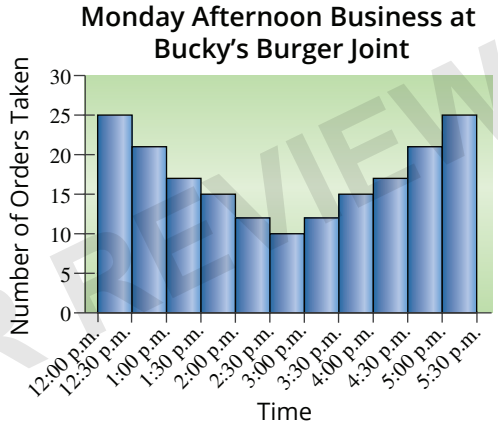
a.



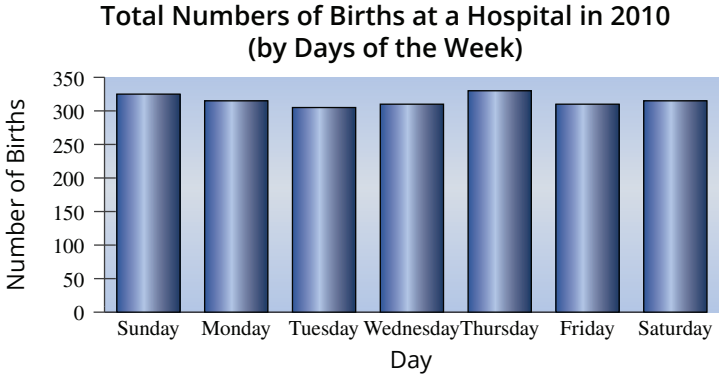
b.



c.



d.



FOR REVIEW ONLY
FOR REVIEW ONLY

3.5 Analyzing Graphs

Graphs that help us visualize data can either be enlightening, in the sense that they gives us insight and understanding of a set of data, or misleading, either intentionally or unintentionally. When you see graphs in the media, you need to be cautious to ensure the data has been accurately represented by the graph. This section will help you analyze graphs for accuracy and appropriate presentation of the given information. Here are a few key ideas to consider when interpreting information displayed in graphical form.

Graph Labeling

Every graph should be properly labeled with an appropriate title that tells you what type of information is being displayed. Also, if the graph has a horizontal and vertical axis, these should be labeled and should include the unit of measurement when necessary for the understanding of the data. For example, in the graph shown below, the title does not provide enough information about the data. Why were those countries chosen? Do they have relatively high or low prison populations compared to the rest of the world? Furthermore, we do not know whether this information is relevant to modern times. Is this data for a specific year? The countries are labeled along the horizontal axis, but note that the vertical axis is just labeled *Population*. We have no idea what the values along the vertical axis represent. Is the prisoner population in units of thousands, millions, or billions? In fact, this chart shows the countries with the top ten highest prisoner populations for the year 2016. The unit for the vertical axis should be thousands, which means that the United States had a prison population of approximately 2217 thousand, or 2.217 million, in the year 2016. Without these seemingly small pieces of information, the graph is not very informative. It is also good practice to use the largest possible unit for the scale of an axis, which in this case is correctly chosen to be thousands.

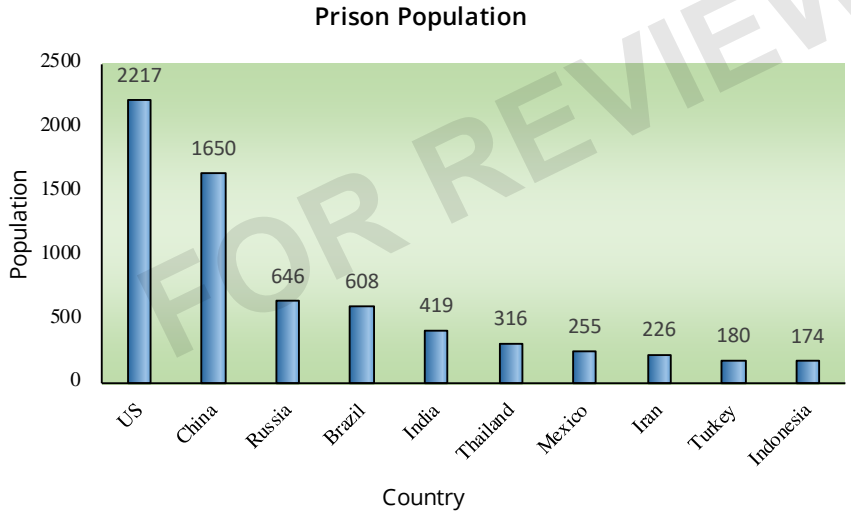


Figure 3.5.1

FOR REVIEW ONLY
FOR REVIEW ONLY

Sources

When examining graphs in the media it is very important to consider the source of the information, i.e., who is telling the story. What good is a graph, or any other piece of information, if it is not backed up by a reliable source? It certainly makes a difference whether the information comes from a reputable news network or a tabloid, doesn't it? The data on the U.S. population from 1790 to 2010 from the previous section came from census data collected by the U.S. Census Bureau, which is a highly reputable source of information.

Internet sources need to be evaluated carefully. The three letters at the end of a URL (Uniform Resource Locator) give you some indication as to the reliability of the information, both content and graphics, that you might obtain from the web site. Some of the more reliable sources end with the following:

.edu .gov .mil .org .com .net .biz

Stay away from web sites that allow anyone to publish or edit information on the site. You want to use a web site whose content is written by experts in the field and can be documented as to the source of the information. Also, check the date on the information to be sure that you are looking at the most recent data and graphics available.

Always evaluate the credentials of the author of any article, blog, or graphic that you find on the internet. What degrees or titles do they hold? Where was the research conducted? Did the results agree with other similar studies? If this level of detail is not provided, it is probably not a reliable source. You should also determine who owns or sponsors a particular web site to make sure it isn't biased in its opinions or favors a particular group's beliefs or views.

Appropriateness of a Graph

Throughout this chapter we have looked at a wide variety of graphs. What we want to consider now is the *appropriateness* of a graph. In other words, we want to be able to determine whether the type of graph being used is correct for the data being displayed. For example, let's contrast the different uses of line graphs and bar graphs. In the previous section we looked at the U.S. population since 1790 and chose to display that data using a line graph. We could just have easily used the following bar graph.

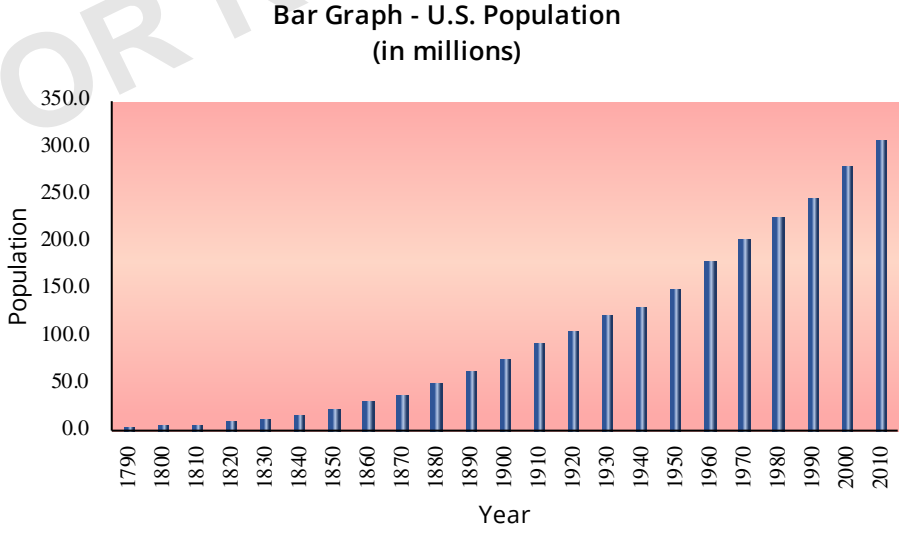


Figure 3.5.2

FOR REVIEW ONLY
FOR REVIEW ONLY

The bar graph also shows the increasing trend of the US population over time, but the trend is better seen using a line graph. The lengths of the bars in the bar graph are somewhat distracting, and obscure the change in population from one time period to the next.

Bar graphs and pie charts are often interchangeable when displaying qualitative data, but if you want to accurately display the categories as parts of a whole, the pie chart will give the best representation, as it allows you to visually compare the slices of the "pie" very quickly. However, the pie chart becomes less effective as the number of categories increases as was depicted in Section 3.2.

Scaling of Graphs

Another important feature to keep in mind when analyzing graphs is whether a graph is scaled appropriately. If you stretch or shrink the scale on either axis, the shape of the graph can change dramatically and thus affect the interpretation of the graph. For example, suppose that we change the scale for Figure 3.2.8 on federal government spending to range from 0 to 0.8 instead of 0 to 0.4.

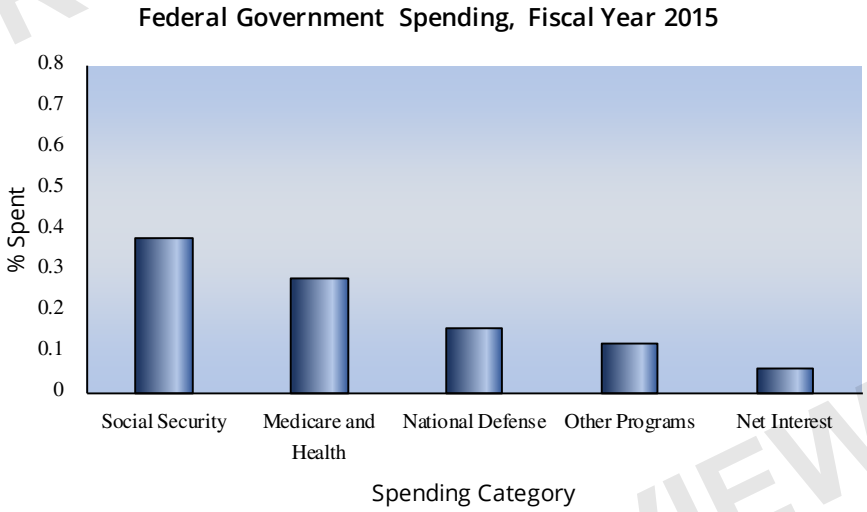


Figure 3.5.3

Note that the graph now minimizes the differences among the five spending categories compared to the original graph. The large amount of "white space" in the graph is a good indicator that the scale may not be appropriate for the data. Therefore, when analyzing a graph make sure that the scale represents the data well.

Data Transformations

It is often useful to transform data by replacing a variable in the data set by a function of that variable so that the distribution is easier to work with or interpret. For example, if the data set contains a variable x, we can transform the distribution of the data by replacing x with some function of x, such as the square root of x or the logarithm of x.

In fact, we have already applied a transformation to one of the geospatial graphs in the previous section. Recall that, due to the massive variance in US county populations, we could not accurately compare the obesity prevalence between counties when we measured the number of obese people in each county. To account

FOR REVIEW ONLY
FOR REVIEW ONLY

for this, we applied a function to our data that transformed the variable of interest, the *number* of obese people in a county, into the *percentage* of obese people in a county. The function divided the number of obese people in a county by the total county population, multiplied the quotient by 100, and resulted in the percentage of the county population that was obese. The alternate perspective of the data then allowed us to easily compare counties since the variable of interest was normalized by total county population.

In this section, we will focus on another transformation, the log transformation, which is one of the most common transformations in statistics. The log transformation is often used to help "unclutter" data points for visualization purposes. When data are tightly grouped together, it can be hard to visualize and make inferences about individual data points. Log transformations numerically "stretch" the portion of the axis closest to zero, and "compress" the portion of the axis farthest from zero. This allows us to better visualize each individual point while also maintaining the underlying relationship between the variables being graphed. The log transformation also allows us to visualize relative change as opposed to absolute change.

Data

Data > Amazon Stock

Figure 3.5.4 shows the price per share of Amazon stock since its initial public offering (IPO) in May 1997. No transformation is applied to the data which means that the graph shows absolute change.

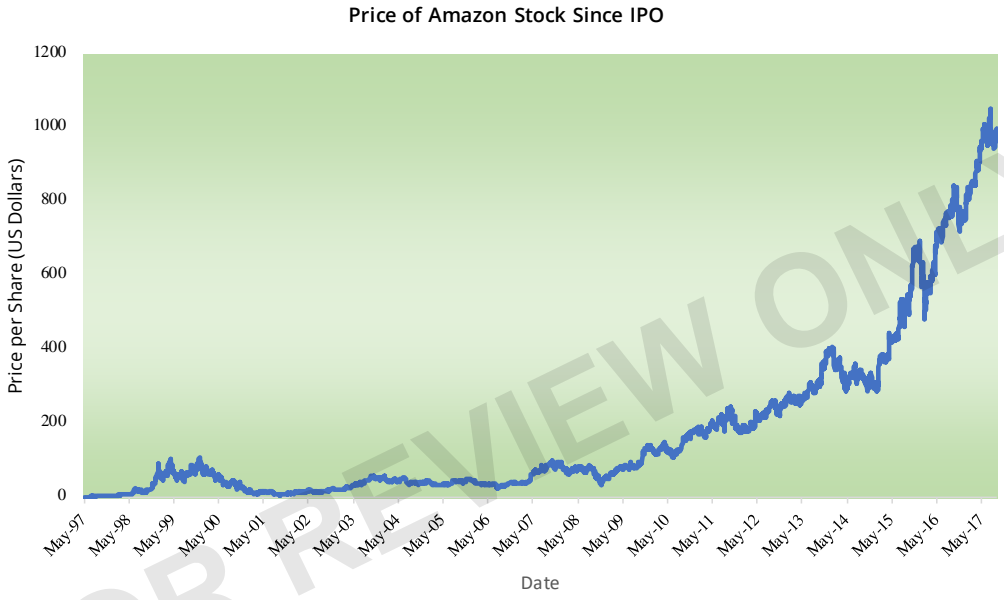


Figure 3.5.4

As you can see, Amazon has experienced massive success over the years. However, this graph seems to suggest that Amazon didn't really hit its stride until about 2010. How, then, was it possible for Amazon to survive the dot-com boom of the late 1990s? During such a tumultuous period, if an Internet company didn't catch the eye of the public, it quickly faded from existence; in many cases, even companies that did catch the public's eye were rapidly eclipsed by newer technologies, or simply better business models. There has to be more to Amazon's story that this graph isn't showing us.

In terms of data visualization, logarithms can be used to determine relative change between observations. As you can see in Figure 3.5.4, Amazon's stock price starts to become very substantial starting in 2010. In fact, the difference between the stock

FOR REVIEW ONLY
FOR REVIEW ONLY

price in May 2015 and the price in May 2016 is a larger number than the maximum stock price was in any year prior to 2011. When we compare every data point to the original base, the stock price in May 1997, it becomes very hard to compare relative performances from year to year; a good amount of growth in the late 1990s and early 2000s would not be considered very good after 2010. However, transforming our price variable by applying a log function will allow us to analyze Amazon's yearly performance relative to the stock price for each year. After applying the log function to the *y*-axis of the previous graph, we generate the following graph which tells the story in a slightly different way.

Figure 3.5.5

Notice that the *y*-axis of Figure 3.5.5 now portrays the variable in factors of 10 as opposed to a continuous scale. It turns out that Amazon experienced its most rapid period of relative growth in the years immediately following its IPO. However, notice that the stock price began to sharply decline in 1999. This was when the dot-com bubble "burst" which essentially means that too many mediocre companies were saturating the Internet market. When the bubble burst, thousands of Internet companies went out of business, which naturally meant that confidence in Amazon began to wane. Fortunately, Amazon's business model was built for long-term growth, and it allowed the company to survive the downward economic trend. In 2001, Amazon turned its first profit and restored the confidence of investors. Since then, Figure 3.5.5 reveals that Amazon continues to experience steady growth to this day. This knowledge aids in the explanation of how Amazon was able to survive the bursting of the dot-com bubble, but we might never have put all the pieces of the puzzle together had we not used the log transformation to gain an alternate perspective.

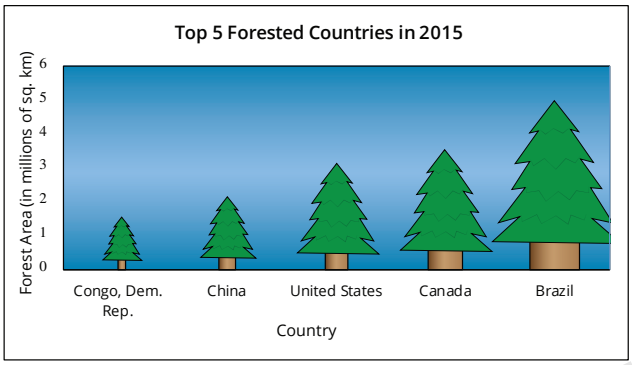
FOR REVIEW ONLY
FOR REVIEW ONLY

Misleading Graphs

An issue related to scaling that is often used to mislead readers is to start the vertical scale at some value other than zero. We saw this earlier with the sales performance data in Section 3.2. By starting the scale on the vertical axis at $180,000 and using increments of $5,000, the sales performance differences were exaggerated (see Figures 3.2.2 and 3.2.3).

One type of graph commonly used in the media is a pictograph, because it is visually appealing and simple to understand. A **pictograph** is basically a bar graph that uses pictures of objects in place of the bars. For example, the graph on the left in Figure 3.5.6 shows an inaccurate pictograph of the top five countries ranked in order by the amount of forest area they contain. The bars in this bar chart are represented by trees. Note how as the amount of forest area increases, the trees expand with regard to width and height, thus giving the illusion that the amount of increase is much larger than it is in reality. This is a common problem with pictographs found in the media. Often when the size of the object is increased or decreased, the change is not simply one-dimensional. So be very cautious when looking at data displayed with a pictograph to make sure the graph isn't misleading by representing increases or decreases along one dimension (height of the bar) using an object that is changing in area or volume. The graph on the right represents the correct way to scale a pictograph by only changing the heights of the trees.

Incorrect Pictograph



Correct Pictograph

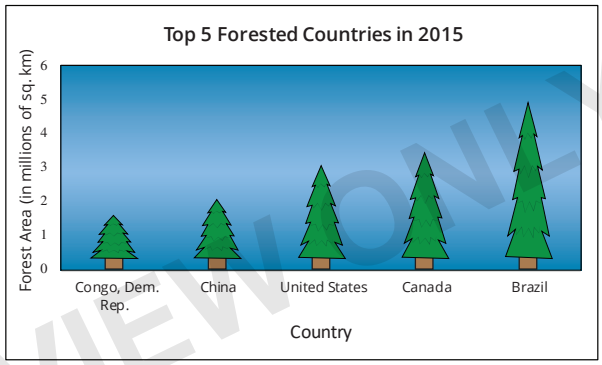


Figure 3.5.6

3.5 Exercises

Basic Concepts

1. Why is it important to label and title graphs properly?
2. List 3 URL extensions for reliable sources.
3. Why is the scaling of a graph important?
4. Why are data transformations useful?

FOR REVIEW ONLY
FOR REVIEW ONLY

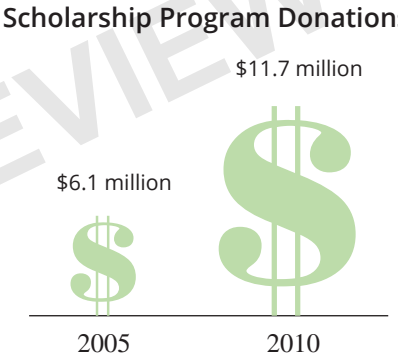
Exercises

1. Consider the following pictograph used to display the increase in funds donated to one university's scholarship program.

a. By what percentage did funds to the university scholarship program increase between 2005 and 2010?

b. Does the graph shown accurately depict the change in scholarship funds? Explain your answer.

c. What changes could be made to better display the given information?



2. Using the San Francisco Salaries 2014 data set from the web resource, create a histogram for the variable TotalPayBenefits and answer the following:

a. Is the distribution of the data in the histogram uniform, normal, skewed right, or skewed left?

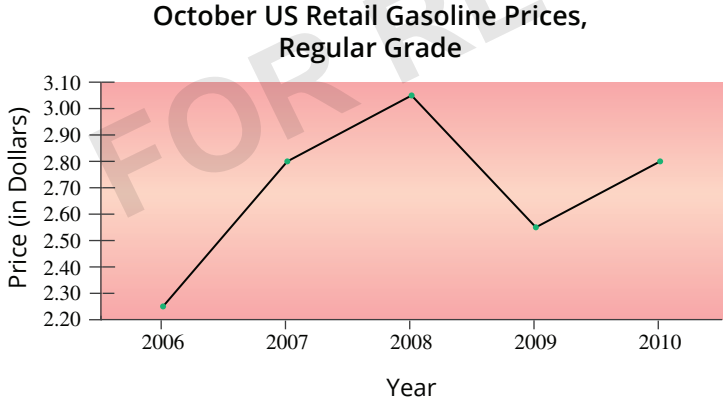
b. Construct a new histogram for the variable LogTotalPayBenefits, which is a log transformation of the variable TotalPayBenefits.

c. Is the distribution of the data in the log transformed histogram uniform, normal, skewed right, or skewed left?

Data

Data > San Francisco Salaries 2014

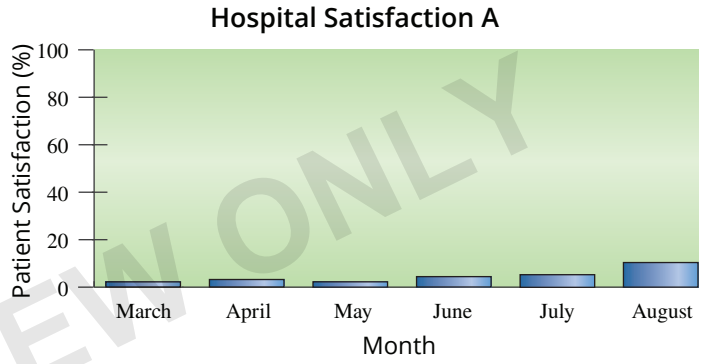
3. Does the scale on the following graph depict the situation accurately? Why or why not?

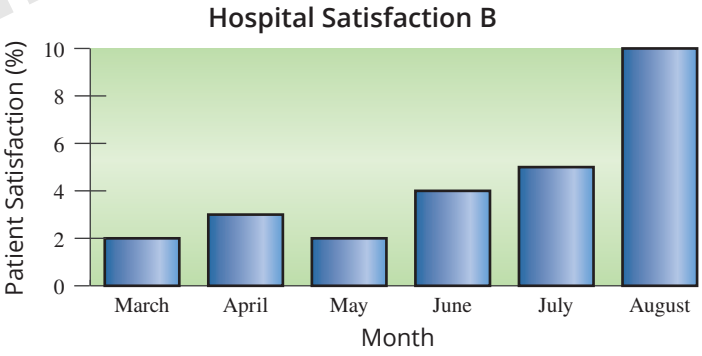


4. Look at the two graphs shown below depicting the *same* data on people's overall satisfaction level with the care they received at their local hospital. Which of these two graphs shows the more accurate picture of hospital satisfaction?

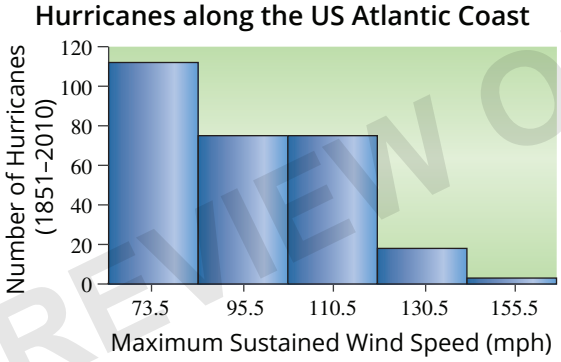
FOR REVIEW ONLY

Has hospital satisfaction increased? Are people satisfied with the care at their local hospital according to these graphs? How do you know?

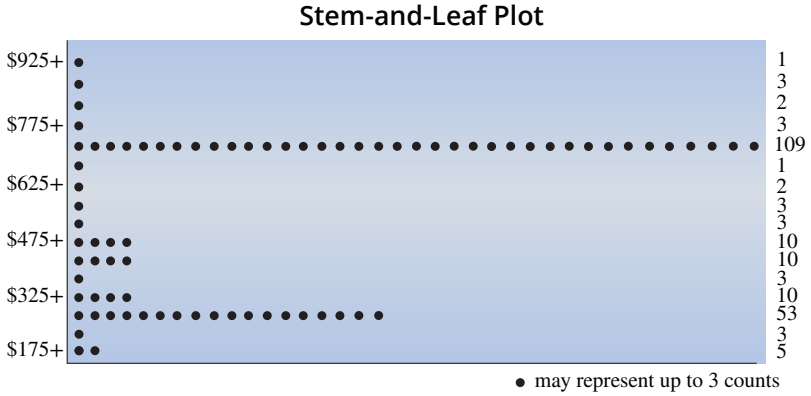




5. What errors occur in the following histogram?



6. Consider the following excerpt from an online publication. Is the graph correctly labeled? If not, identify the corrections needed.



FOR REVIEW ONLY
FOR REVIEW ONLY

4.1 Measures of Location

To your question concerning the whereabouts of Tom Stevens, a mutual friend responds, "In the library." The answer is informative in that it provides general information about Tom's whereabouts. Even though you don't know what part of the library he is in, the information does provide a focal point—a location of sorts. The information may also permit some inferences regarding what Tom is doing.

Statistically speaking, the idea of location is similar to knowing that Tom is in the library. If we think of a data set as a group of data values that cluster around some central value, then this central value provides a focal point for the data set—a location of sorts. Unfortunately, the notion of *central value* is a vague concept, which is as much defined by the way it is measured as by the notion itself. There are several statistical measures that can be used to define the notion of center: the arithmetic mean, weighted mean, trimmed mean, median, and mode.

The Arithmetic Mean

The **arithmetic mean** is one of the more commonly used statistical measures. It appears every day in newspapers, business publications, and frequently in conversation. For example, after receiving a grade on a test, you might be curious about how the rest of the class performed. You might ask the instructor, *What was the average on the test?* The term *average* is often associated with the arithmetic mean, which is more commonly referred to as just the **mean**.

Arithmetic Mean

Suppose there are n observations in a data set, consisting of the observations $x_1, x_2, ..., x_n$; then the **arithmetic mean** is defined to be

$$\frac{1}{n}\left(x_1 + x_2 + ... + x_n\right).$$

DEFINITION

If we use some common mathematical notation, the formula can be simplified to

$$\frac{\sum x_i}{n}$$

where x_i is the i^{th} data value in the data set and $\sum$ (pronounced sigma) is a mathematical notation for adding values. There are two symbols that are associated with the expression given above:

$\mu = \frac{1}{N}\left(x_1 + x_2 + ... + x_N\right)$ the **population mean**, and

$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + ... + x_n\right)$ the **sample mean**.

Here N refers to the size of the population and n refers to the size of the sample. Otherwise, the calculations are made in precisely the same way. The Greek letter μ, representing the population mean, is pronounced *mu*, and the symbol $\bar{x}$, representing the sample mean, is pronounced *x-bar*.

FOR REVIEW ONLY
FOR REVIEW ONLY

Technology

To find summary statistics, such as the sample mean, we can use the `1-Var Stats` option on the TI-83/84 calculator. The results are shown here. For instructions please refer to the web resource at **Tech > Descriptive Statistics - One Variable**.

```
1-Var Stats
 x̄=9
 Σx=36
 Σx²=390
 Sx=4.69041576
 σx=4.062019202
↓n=4
```

```
1-Var Stats
↑n=4
 minX=4
 Q1=5.5
 Med=8.5
 Q3=12.5
 maxX=15
```

Example 4.1.1

Calculate the sample mean of the following sample data values: 4, 10, 7, 15.

Solution

$$x_1 = 4,\ x_2 = 10,\ x_3 = 7,\ x_4 = 15, \text{ and } n = 4.$$

Note that

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4+10+7+15}{4} = \frac{36}{4} = 9.$$

The sample mean is 9. But why does adding up a group of numbers and dividing by the number of numbers measure central tendency? As unlikely as it sounds, the answer is related to balancing a scale.

Deviation

Given some point A and a data point x, then $x - A$ represents how far x **deviates** from A. This difference is also called a **deviation**.

DEFINITION

Let's calculate the deviations from the mean (9) for the data in Example 4.1.1. Examining the deviations from the mean in Table 4.1.1, we can see the deviations on the left side (−5 and −2) and right side (1 and 6) are in balance. In fact, the mean is considered a point of centrality because the deviations from the mean on the positive side and the negative side are equal (see Figure 4.1.1). The sample mean can be interpreted as a center of gravity.

Table 4.1.1 – Deviations from the Mean

Data (x_i)	Deviations from the Mean (x_i − 9)
4	−5
10	1
7	−2
15	6
TOTAL	$\sum(x_i - 9) = 0$

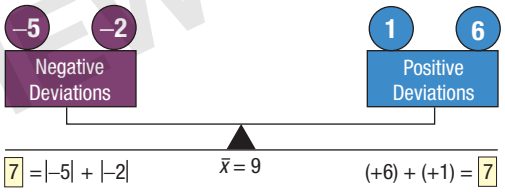


Figure 4.1.1

On the other hand, if we calculate the deviations about any other value the deviations do not balance. For example, assume the central value is 8. The deviation from the alleged central value, 8, for each data value is calculated in Table 4.1.2 and shown in Figure 4.1.2.

The positive deviations (+2 and +7) are not counterbalanced by the negative deviations (−4 and −1). A desirable characteristic of a central value would be to have the positive and negative deviations equal to each other in absolute value.

FOR REVIEW ONLY

Table 4.1.2 – Deviations from Some Other Value

Data (x_i)	Deviations from 8 ($x_i - 8$)
4	−4
10	2
7	−1
15	7
	TOTAL $\sum(x_i - 8) = 4$

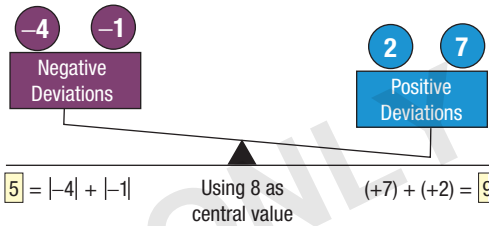


Figure 4.1.2

Although the arithmetic mean is frequently used, there are times when it should not be employed. *Since the mean requires that the data values be added, it should only be used for quantitative data*. Furthermore, if one of the data values is extremely large or small relative to the others, this data value could be considered an **outlier**. An outlier can have a dramatic impact on the value of the mean and dramatically affect its value as a measure of central tendency. (See Section 4.3 for further discussion of outliers.)

Outliers and Resistant Measures

An **outlier** is a data value that is extremely different from other measurements in the data set. Statistical measures which are not affected by outliers are said to be **resistant**.

DEFINITION

The mean is *not* a **resistant measure**.

Weighted Mean

The weighted mean is similar to the arithmetic mean except it allows you to give different weights (or importance) to each data value. The weighted mean gives you the flexibility to assign weights when you find it inappropriate to treat each observation the same. The weights are usually positive numbers that sum to one, with the largest weight being applied to the observation with the greatest importance. The weights can be determined in a variety of ways, such as the number of employees, the market value of a company, or some other objective or subjective method. There are occasions in which it is easier to assign the weights without worrying that they will sum to one. If you are concerned about your weights summing to one, you can make your weights sum to one by dividing each weight by the sum of all the weights.

Weighted Mean

The weighted mean of a data set with values $x_1, x_2, x_3, \ldots, x_n$ is given by

$$\bar{x} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum(w_i \cdot x_i)}{\sum w_i}$$

where w_i is the weight of observation x_i.

FORMULA

FOR REVIEW ONLY

Example 4.1.2

Meghan is a freshman in college and she received the following grades for her first semester.

Meghan's Grades

Course	Grade	Credit Hours
Psychology 101	B	3
Probability and Statistics	A	4
Anatomy I	C	5
English 101	A	3

A grade of A is worth 4 points on a 4-point scale. A grade of B is worth 3 points and a grade of C is worth 2 points.

a. Calculate Meghan's GPA using the credit hours as weights. Round to two decimal places.

b. If Meghan's goal was to have a GPA of 3.4, what grade did she need to make in the Anatomy I class to reach her goal?

Solution

a. To calculate Meghan's GPA we use the weighted mean formula with the numerical grade point values as the x-values and the credit hours as weights.

$$\bar{x} = \frac{\sum(w_i \cdot x_i)}{\sum w_i} = \frac{3\cdot 3 + 4\cdot 4 + 2\cdot 5 + 4\cdot 3}{3+4+5+3}$$

$$= \frac{9+16+10+12}{15} = \frac{47}{15}$$

$$\approx 3.13$$

b. To determine the grade that Meghan needed in the Anatomy I class to reach her goal of a 3.4 GPA, let x represent the grade in the weighted mean formula.

$$\bar{x} = 3.4 = \frac{3\cdot 3 + 4\cdot 4 + x\cdot 5 + 4\cdot 3}{15}$$

$$= \frac{9+16+5x+12}{15} = \frac{37+5x}{15}$$

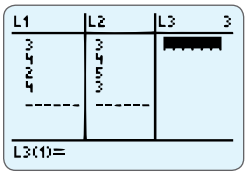
Solving this equation for x gives us the following result.

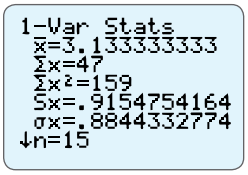
$$\begin{aligned} 3.4(15) &= 37 + 5x \\ 51 &= 37 + 5x \\ 51 - 37 &= 5x \\ 14 &= 5x \\ 2.8 &= x \end{aligned}$$

Since the grading scale only uses integers, Meghan would have had to make 3 points or a grade of B on her Anatomy I class for a GPA of 3.4.

Technology

To find the weighted mean we use the `1-Var Stats` option on the TI-83/84 calculator with the numerical grade point value in L1 and credit hours in L2. The results are shown below. For instructions please refer to the web resource at **Tech > Weighted Mean**.





FOR REVIEW ONLY
FOR REVIEW ONLY

The Trimmed Mean

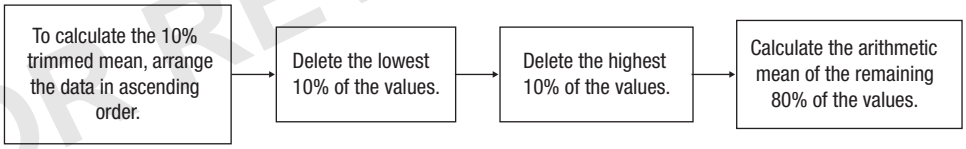
Since outliers can have an enormous effect on the value of the mean, the mean's usefulness as a typical measure of a set of data is diminished if the data contain outliers.

Trimmed Mean

The **trimmed mean** is a modification of the arithmetic mean which ignores an equal percentage of the highest and lowest data values in calculating the mean.

DEFINITION

Finding the 10% Trimmed Mean



Before calculating the trimmed mean, the data are arranged in ascending order of magnitude. A 10% trimmed mean uses the middle 80% of the data values. It is calculated by removing the top 10% *and* the bottom 10% of the data values, then finding the arithmetic average of the remaining values. If the data set does not contain any outliers, the mean and the trimmed mean will be similar. Unlike the arithmetic mean, the trimmed mean is not affected by outliers and is considered a resistant measure.

Example 4.1.3

Consider the following data taken from a poll on how many text messages a person sent in a day.

16 18 20 21 23 23 24 32 36 42

mean = 25.5

Find the 10% trimmed mean.

Solution

Since there are 10 observations, removing the highest 10% and lowest 10% means removing only one observation from each end of the data. That is,

$$10\% \text{ of } 10 = 0.1 \cdot 10 = 1.$$

Note that the data are already sorted. If the mean is calculated without including the values 16 and 42, the resultant measure is called the 10% trimmed mean.

~~16~~ 18 20 21 23 23 24 32 36 ~~42~~

$$10\% \text{ trimmed mean} = \frac{18+20+21+23+23+24+32+36}{8} = 24.625$$

Rounding Rule

A general rule of thumb regarding rounding for the statistics calculated in this chapter is to round to one more decimal place than the largest number of decimal places given in the data. Occasional exceptions to this rule can be made when the type of data lends itself to a more natural rounding scheme, such as rounding values of currency to two decimal places.

FOR REVIEW ONLY

Measuring Figure Skating Performances

Almost every figure skating competition has some scoring controversy. The Winter Olympics of 2002 were no exception. French judge, Marie-Reine Le Gougne, said she was "pressured to vote a certain way'' when she scored the Russian couple, Elena Berezhnaya and Anton Sikharulidze, over the Canadian pair, Jamie Sale and David Pelletier. In addition, very few people understood exactly how Sarah Hughes won the gold medal and how Michelle Kwan dropped to third after leading the event.

For almost a century figure skating has used a scoring method that is similar to the methodology of the trimmed mean in order to remove bias. Skaters were scored on a 0 to 6 scale. The highest and the lowest score are discarded (the data is trimmed) and the resulting "score" is computed. The intent of trimming the data is to avoid bias caused by judges with nationalistic or political agendas. The International Skating Union is replacing this scoring method with similar methods that attempt to remove biases in judge's scores.

If there had been 100 data values, the largest 10% and smallest 10% (a total of 20 data values) would have been removed before the mean was calculated.

Example 4.1.4

Consider the same data set, except the last data value is replaced with an outlier.

16 18 20 21 23 23 24 32 36 490

mean = 70.3

Find the 10% trimmed mean.

Solution

Since there are 10 observations, removing the highest 10% and lowest 10% means removing only one observation from each end of the data.

~~16~~ 18 20 21 23 23 24 32 36 ~~490~~

$$10\%\text{ trimmed mean} = \frac{18+20+21+23+23+24+32+36}{8} = 24.625$$

As expected, the trimmed mean is not affected by the addition of the outlier, while the mean increased dramatically. This is why the trimmed mean is considered to be a **resistant measure**.

The Median

The median of a set of data provides another measure of center that is different from a "mean". It is a simple idea. To find the median, place the data in ascending order and then find the observation that has an equal number of data values on either side. That is, half of the observations are less than the median and half of the observations are greater than the median. The median is the middle value.

Median

The **median** of a set of observations is the measure of center that is the middle value of the data when it is arranged in ascending order. The same number of data values lie on either side of the median.

DEFINITION

To determine the median of a set of data, we use the following steps.

PREVIEW ONLY FOR REVIEW ONLY

Finding the Median of a Data Set

1. Arrange the data in ascending order.
2. Determine the number of values in the data.
3. Find the data value in the middle of the data set.
4. If the number of data values is odd, then the median is the data value that is exactly in the middle of the data set.
5. If the number of data values is even, then the median is the mean of the two middle observations in the data set.

PROCEDURE

Example 4.1.5

Consider the following goal tallies from the last eleven games played by the Charleston Battery soccer team.

2, 3, 5, 4, 1, 7, 3, 3, 1, 2, 6

Find the median.

Solution

First, the data set must be ordered,

1, 1, 2, 2, 3, 3, 3, 4, 5, 6, 7

Since the data set contains an odd number of values, 11, the middle observation in the ordered array must be the sixth observation. Since the median is the sixth observation, the median value is 3.

1, 1, 2, 2, 3, **3**, 3, 4, 5, 6, 7

Example 4.1.6

Consider the following ten test scores from a student taking a high school calculus class.

65, 98, 76, 83, 94, 79, 88, 72, 90, 85

Find the median.

Solution

If there are an even number of observations, average the two center values in the ordered data set.

65, 72, 76, 79, **83**, **85**, 88, 90, 94, 98

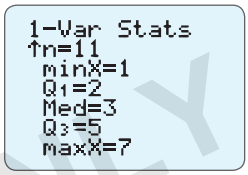
To find the median, average the fifth and sixth observations.

$$\frac{83+85}{2} = 84 \text{ (the median)}$$

Technology

To find summary statistics, such as the median, we can use the `1-Var Stats` option on the TI-83/84 calculator or use Excel. The results are shown here. For instructions please refer to the web resource at **Tech > Descriptive Statistics - One Variable**.

Column1	
Mean	3.363636364
Standard Error	0.591957113
Median	3
Mode	3
Standard Deviation	1.963299634
Sample Variance	3.854545455
Kurtosis	-0.466246885
Skewness	0.636683254
Range	6
Minimum	1
Maximum	7
Sum	37
Count	11



FOR REVIEW ONLY
FOR REVIEW ONLY

The median possesses a rather obvious notion of centrality, since it is defined as the central value in an ordered list. It is not affected by outliers and is therefore a resistant measure. For example, if we replaced 98 with 200,000,000 in the data set from the previous example, the median would not change at all. The median does possess one limitation: it cannot be applied to nominal data. In order to calculate the median, the data must be placed in order. To accomplish this task meaningfully, the level of measurement must be at least ordinal. Unless the data set is skewed or contains outliers, the median and the mean usually have similar values.

The Mode

The **mode** is another measure of location. It is not used as frequently as the mean or the median, and its relation to the "central tendency" concept and the values of the mean and median are not predictable. The mode is the only measure of location that can be used for nominal data. Of the three measures of location, the mode is used the least due to the limited information it provides. Sometimes sorting the data (in ascending or descending order) makes it easier to find the mode.

Mode

The **mode** of a data set is the most frequently occurring value.

DEFINITION

When reporting the mode for numerical data, do not round. The value of the mode should be the same as the original data value.

Example 4.1.7

Find the mode of the following data regarding the number of power outages reported over a period of eleven days.

0, 1, 4, 3, 9, 8, 10, 0, 1, 3, 0

Solution

Since the value of 0 occurs more than any other value, it is the mode. In this instance, as a measure of location, the modal value is not a particularly appealing choice. However, as noted previously, the mode does possess one very favorable property—it is the only measure of location that can be applied to nominal data. Therefore, for nominal measurements like color preferences, it would be perfectly reasonable to discuss the modal color.

Suppose we added one more value to the data set in Example 4.1.7. If this value were a 1, then both 0 and 1 would be repeated three times and there would be two modes. When this occurs, the data are said to be **bimodal**. Any time a data set has more than two modes, it is said to be **multimodal**. If all observations in a data set occur with the same frequency, the data set has **no mode**.

FOR REVIEW ONLY
FOR REVIEW ONLY

The Relationship between the Mean, Median, and Mode

Oftentimes, the shape of the data determines how the mean, median, and mode are related. For a bell-shaped distribution, the mean, median, and mode are identical.

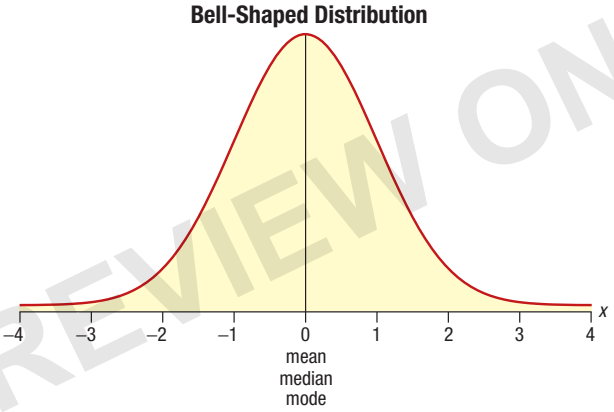


Figure 4.1.3

Certainly, not all data produce distributions that follow a bell-shaped curve. If the distribution of the data has a long tail on the right, it is said to be skewed to the right, or positively skewed. Conversely, if the distribution has a long tail on the left, it is said to be skewed to the left, or negatively skewed. If the data are positively skewed, the median will be smaller than the mean. If the data are negatively skewed, the mean will be smaller than the median.

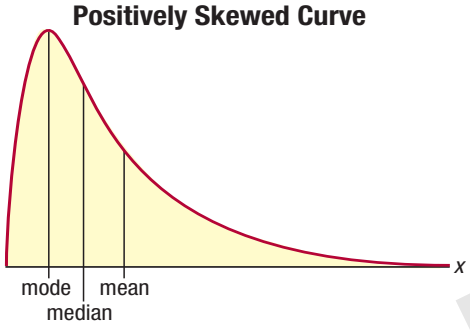


Figure 4.1.4

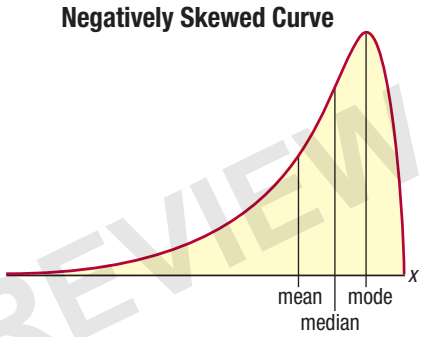
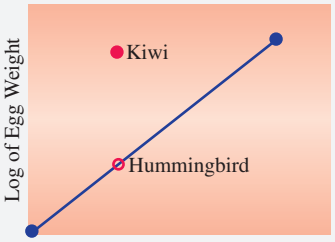


Figure 4.1.5

Where does the mode fall on these graphs? The area containing the greatest number of observations contains the mode. That area is represented by the large peak in the curve. In Figure 4.1.4 and Figure 4.1.5 the obvious peaks in the curves are the portions of the distributions that will contain the mode. The highest point on the curve will be the mode of the distribution. Notice that in a bell-shaped distribution (Figure 4.1.3), the mode is equal to the mean and median. If the distribution is positively skewed (Figure 4.1.4), the mode is less than the mean and median. If the distribution is negatively skewed (Figure 4.1.5), the mode is greater than the mean and median.



Kiwi Eggs Are Outliers!

Both animal and plant kingdoms offer spectacularly odd and beautiful sights. A kiwi bird (one of many interesting life forms from New Zealand) lays eggs that are close to 25% of their body weight and sometimes lays two or three such eggs at a time. For birds, eggs usually correspond to about 5% of the bird's weight among all species.

If you draw a graph relating (log) egg weights against (log) body weight you get a so-called hummingbird-moa curve (moa, an extinct ostrich like bird of the New Zealand area.) In this curve, the kiwi shows up as an outlier. Using the kiwi body weight (of about 5 lbs) one expects an egg weight of about 55 to 100 grams while the real weight of kiwi eggs is about 400 to 435 grams matching an expected body weight of about 40 lbs. Why? And what accounts for such an anomaly?

The most reasonable explanation provided by biologists is that kiwis and moa birds are members of the same species except the kiwis have dwarfed through their evolutionary history. A subarea of biology called "allometry" states that as body size decreases the internal organs decrease relatively slowly which supports the dwarfism hypothesis. The kiwis have lost body weight but not their internal womb structure which still holds large eggs. Outliers are important because they force you to think about your data more seriously.

Source: Gould, S. J. (1991). "Of Kiwi Eggs and the Liberty Bell" in Bully For Brontosaurus, W.W. Norton and Company, New York.

FOR REVIEW ONLY

Selecting a Measure of Location

The objective of using descriptive statistics is to provide measures that convey useful summary information about the data. When selecting a statistic to represent the central value of a data set, the first thing to consider is the type of data being analyzed.

The arithmetic mean is used so frequently that its computation is almost a knee-jerk reaction to analyzing data. Unfortunately, it is not always a reasonable measure of centrality or location. Table 4.1.3 defines the applicable levels of measurement for each measure of location. Table 4.1.4 defines the sensitivity to outliers for each measure of location. When the data are qualitative (nominal or ordinal), the mean should not be calculated and, if the data are quantitative and contain outliers, the mean does not convey the notion of typical value as well as some other measures. The only time in which the mean should be used without any explanation is when the distribution of the data is symmetrical or nearly so. In that event, the mean and median should be approximately the same value.

Table 4.1.3 – Applicable Level of Measurement

	Qualitative Nominal	Ordinal	Quantitative Interval	Ratio
Mean			✔	✔
Median		✔	✔	✔
Mode	✔	✔	✔	✔
Trimmed Mean			✔	✔

Table 4.1.4 – Sensitivity to Outliers

	Not Sensitive	Very Sensitive
Mean		✔
Median	✔	
Mode	✔	
Trimmed Mean	✔	

The median is also a good measure of central tendency. It is not sensitive to outliers and can be applied to data gathered from all levels of measurements except nominal.

If the level of measurement of the data is interval or ratio and there are no outliers, the mean is a reasonable choice. If the data set appears to have any unusual values, then the trimmed mean or the median would be more appropriate.

If the level of measurement is nominal or ordinal (the data are qualitative), appropriate measures of center are limited. If the data are ordinal, then the median is the best choice. If the data are nominal, there is only one choice, the mode. The mode is applicable to all data types, although it is not very useful for quantitative data.

Time Series Data and Measures of Centrality

We discussed two types of time series data in Chapter 2, stationary and nonstationary. Stationary time series wobbled around some central value, so calculating a central value is perfectly reasonable, and the methods we previously discussed are applicable. A nonstationary time series is another story. Nonstationary time series possess trend. That means there is no central value for the time series. Instead, the series trends in one direction or another. Computing a central value using the methods discussed earlier would be inappropriate for such data.

Table 4.1.5 shows the first six rows of the data on average US gas price from 1991 to 2015. In this nonstationary time series, the central value of the process is trending upward as shown in Figure 4.1.6. One way to capture this movement is with a **moving average**.

FOR REVIEW ONLY
FOR REVIEW ONLY

Table 4.1.5 – Average US Gas Price 1991-2015 (in dollars per gallon)			
Year	**Average US Gas Price**	**2 Period Moving Average**	**3 Period Moving Average**
1991	1.14		
1992	1.13	1.135	
1993	1.11	1.120	1.127
1994	1.11	1.110	1.117
1995	1.15	1.130	1.123
1996	1.23	1.190	1.163
		. . .	

Data

The complete data set can be found on the web resource at **Data > US Gas Price**.

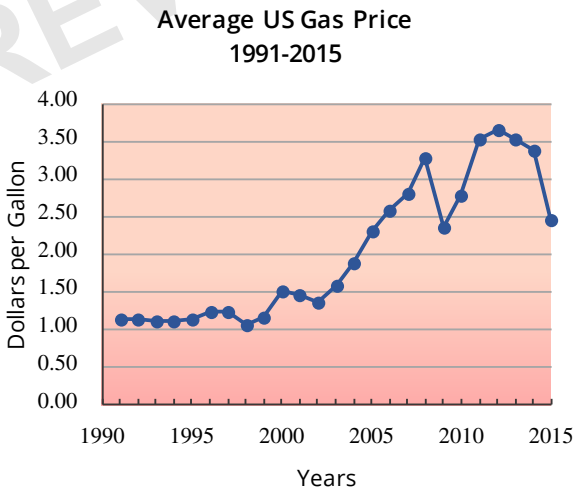


Figure 4.1.6

Moving Average

A **moving average** is obtained by adding consecutive observations for a number of periods and dividing the result by the number of periods included in the average.

DEFINITION

A moving average can be used to forecast the new level of a series over time or as a descriptive method. By averaging just two or three periods at a time we can still see long-term trends, but at the same time smooth out some of the short-term variability in the time series. How the average is associated with a specific period is dependent on its purpose. We will assume the moving average is to be used as a method of forecasting the next level of the time series. Suppose a two-period moving average is calculated for the gas price data and is used to specify the level of the series at a given point in time. The two-period moving average for 1992 averages the values of the time series in 1991 and 1992.

$$\frac{1.14+1.13}{2}=1.135$$

FOR REVIEW ONLY
FOR REVIEW ONLY

Similarly, the two-period moving average for 1993 would be the average of the time series values in 1992 and 1993.

$$\frac{1.13+1.11}{2}=1.120$$

Since data are not available for 1989 or 1990, the three-period moving average for 1991 cannot be calculated. The three-period moving average associated with 1993 is the average of the time series values in 1991, 1992, 1993.

$$\frac{1.14+1.13+1.11}{3}=1.127$$

And, the three-period moving average for 1994 would be the average of the time series values in 1992, 1993, and 1994.

$$\frac{1.13+1.11+1.11}{3}=1.117$$

The chart in Figure 4.1.7 displays the time series, the two-period moving average, and three-period moving average. Both of the averages follow the time series quite closely. However, notice that the two-period moving average follows the actual data values more closely than the three-period moving average.

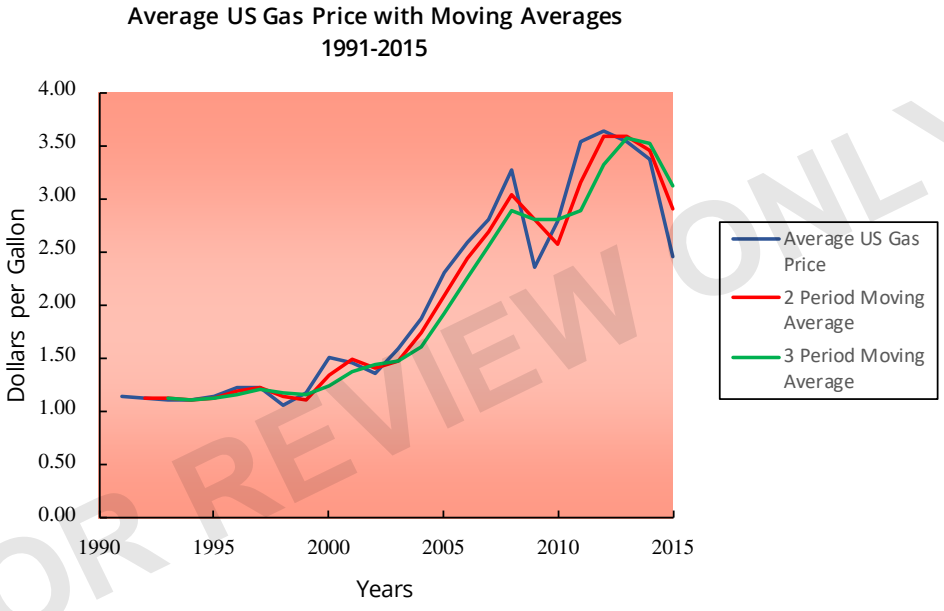


Figure 4.1.7

FOR REVIEW ONLY
FOR REVIEW ONLY

4.1 Exercises

Basic Concepts

1. Describe the difference between statistics and parameters.
2. Describe three major attributes used in summarizing a data set.
3. What are numerical descriptive statistics and why are they important?
4. Identify and describe five measures of location. List the advantages and disadvantages of each.
5. What is a resistant measure?
6. Describe a situation in which using the weighted mean as a measure of location would be appropriate.
7. What does it mean if we say that a data set is positively skewed? Negatively skewed?
8. Explain why the mean should not be calculated for a nonstationary time series.
9. What is a moving average? When is it useful?

Exercises

10. Calculate the mean, median, 10% trimmed mean, and the mode for the following data.

90.25	93.83	91.41	92.27	90.89	99.12	92.88
97.74	96.28	95.33	91.16	94.30	95.51	92.27
97.63	95.94	90.95	94.76	92.27	92.88	

11. Using the US violent crime rate data by state found on the web resource, calculate the mean, median, and 20% trimmed mean for the year 2014.

Data

Data > US Violent Crime by State

12. Using the US violent crime rate data from the previous problem, remove Washington D.C. (District of Columbia) from the data and then calculate the mean, median, and the 20% trimmed mean for the year 2014.

13. Calculate the mean, median, the 10% trimmed mean, and the mode for the following data.

2	22	6	18	10	14	12	12	16	8

14. Discuss the usefulness of each of the measures of central tendency with respect to the following situations.
    a. A company is considering a move into a regional market for specialty soft drinks. In analyzing the size of the containers that his competitors

FOR REVIEW ONLY
FOR REVIEW ONLY

are currently offering, would the company be more interested in the mean, median, or mode of their containers?

b. The creative director for an advertising agency is trying to target an ad campaign that will be shown in one city only. Would he be more interested in the mean or median family income in the city?

c. A young economist was assigned the task of comparing the interest rates on ninety day certificates of deposit (CDs) in three major cities. Should she compare the mean, median, or modal interest for the banks in the three cities?

d. A telephone company is interested in knowing how customers rate their service: excellent, good, average, or poor. Would the company be more interested in studying the mean, median, or mode of the customer service ratings?

15. Discuss the usefulness of each of the measures of central tendency with respect to the following situations.

a. A doctor is interested in analyzing the increase in systolic blood pressure caused by a certain antibiotic. Would the doctor be more interested in studying the mean, median or mode of the systolic blood pressures?

b. A car manufacturer is trying to decide in what colors it should offer its new sports coupe. In analyzing the preferred colors of other sports coupes, would the manufacturer be more interested in the mean, median, or mode of the colors?

c. A manufacturer of chocolate bars is interested in knowing how people rate its chocolate: the best, above average, average, below average, or the worst. Would the company be more interested in the mean, median, or mode of the ratings?

d. A realtor is interested in studying the prices of recent home sales in an area which has many diverse neighborhoods. Would the mean, median, or mode of the prices of recent home sales be the best measure of central tendency?

16. Using the Amazon stock price data set from the web resource, perform the following calculations.

Data

Data > Amazon Stock Price

a. Calculate the mean daily closing stock price in the year 2016.

b. Calculate the median daily closing stock price in the year 2016.

c. Calculate the mode of the daily closing stock price in the year 2016.

d. Calculate the 10% trimmed mean of the daily closing stock price in the year 2016.

e. Which measure of central tendency do you think best describes the center of the data set? Why?

17. A tour guide informs his group that the "average" temperature at their destination is 60 degrees Fahrenheit. Once they arrive, they discover that the daytime highs are about 120 degrees Fahrenheit and the nighttime lows

FOR REVIEW ONLY
FOR REVIEW ONLY

are about 0 degrees Fahrenheit. Do you feel that the tour guide accurately described the temperatures to the group? Discuss.

18. A worker is participating in a test on a new machine. Her daily production, measured in numbers of units, for the twenty-day test is listed below. On days 4 and 5, the worker was ill and went home shortly after coming to work.

Daily Production										
Day	1	2	3	4	5	6	7	8	9	10
Units	100	104	117	20	20	111	105	106	115	101
Day	11	12	13	14	15	16	17	18	19	20
Units	101	102	115	116	113	103	104	119	118	108

a. What level of measurement do the data possess?

b. Compute the mean, 10% trimmed mean and the 20% trimmed mean.

c. Considering the worker's illness, which measure computed in part **b.** best describes the production capability of the machine? Discuss.

19. Using the CO_2 emissions data set from the web resource, answer the following questions.

Data

Data > CO_2 Emissions

a. What level of measurement do the data possess?

b. For the United States, compute the mean, 10% trimmed mean, and the 20% trimmed mean for CO_2 emissions between the years 1960 and 2014. Round your answers to three decimal places.

20. Consider the following monthly sales for a small clothing store in a resort community.

Monthly Sales			
Month	**Sales ($)**	**Month**	**Sales ($)**
January	100,500	**July**	200,000
February	120,000	**August**	185,000
March	133,000	**September**	175,000
April	145,000	**October**	120,000
May	160,000	**November**	180,000
June	180,000	**December**	330,000

a. Draw a line graph of the data.

b. Calculate the two-period moving averages for the data.

c. Calculate the three-period moving averages for the data.

d. Add line graphs for the two-period moving averages and three-period moving averages to the graph which you constructed in part **a.**

e. Which series of data (the original sales data, the two-period moving averages, or the three-period moving averages) do you think best represents sales for the year? Why?

FOR REVIEW ONLY
FOR REVIEW ONLY

21. A student earned scores of 95, 97, 88, 92, and 100 on their first five homework assignments of the semester. The student then missed an assignment and received a 0. The student needs an average of 90 or above to have an overall homework grade of A. The maximum grade for any homework assignment is 100 and there is one homework assignment remaining in the course.

a. Is it still possible for the student to earn an A average for homework?

b. What score would the student need on the final homework assignment to do so?

c. What is the best overall grade the student can earn for their homework?

22. A course is set up such that attendance counts for 5% of the final grade, homework counts for 15%, quizzes count for 20%, and there are midterm and final exams that each count for 30% of the final grade. Before the final exam, a student has an attendance score of 100, a homework average of 80, a quiz average of 75, and a score of 77 for the midterm exam. The student wants to make a B in the course, which would require an overall average of at least an 82. To get a B in the course, what score does the student need to make on the final exam? Round your answer up to the nearest integer.

23. Consider the following average monthly balances for one bank customer for January through March. Calculate the weighted average balance for the three-month period. Note that each average monthly balance must be weighted by the number of days in that month on a non-leap year. Round to the nearest cent.

Average Monthly Balances for a Bank Customer (January through March)

Month	Average Monthly Balance
January	$1885.67
February	$1312.92
March	$2001.53

Data

Data > US County Data

24. Using the US County data set on the web resource, calculate the percentage of the total US population that has at least a high school diploma by utilizing the weighted mean. Use the At.Least.High.School.Diploma variable, which is the percentage of a county population with at least a high school diploma, as the data values, and use the Total.Population variable as the weights.

FOR REVIEW ONLY
FOR REVIEW ONLY

4.3 Measures of Relative Position, Boxplots, and Outliers

Suppose you want to know where an observation stands in relation to other values in a data set. For example, on many standardized tests such as the SAT, GMAT, and ACT, the test scores themselves are rather meaningless unless they are associated with some measure that tells you how well you did relative to others taking the same test. There are two principal methods of communicating relative position: **percentiles** and ***z*-scores**. Both of these methods are data transformations which change the scale of the data in some way.

Percentiles

The most commonly used measure of relative position is the percentile. In fact, we have already discussed the 50th percentile; it is the median. For example, in data sets that do not contain significant quantities of identical data, the 30th percentile is a value such that about 30 percent of the values are below it, and around 70 percent are above it.

P^{th} Percentile

Given a set of data $x_1, x_2, \ldots, x_n$, the **P^{th} percentile** is a value, say x, such that approximately P percent of the data is less than or equal to x and approximately $(100 - P)$ percent of the data is greater than or equal to x.

DEFINITION

To determine the P^{th} percentile, perform the following steps.

Finding the P^{th} Percentile

To determine the P^{th} percentile:

1. Form an ordered array by placing the data in order from smallest to largest.
2. To find the location of the P^{th} percentile in the ordered array, let
$$\ell = n\left(\frac{P}{100}\right)$$
where n is the number of observations in the ordered data.
3. If ℓ is not an integer, then round ℓ up to the next greatest integer. For example, if $\ell = 7.1$, then round ℓ up to 8 and find the data value in the ℓ^{th} location. If ℓ is an integer value, then average the data value in the ℓ^{th} location with the data value in the $(\ell + 1)^{th}$ location.

PROCEDURE

Caution

It is important to remember that when you find the value of ℓ, this result is not the percentile. It is the location of the percentile in the ordered array.

CAUTION

Finding the P^{th} Percentile

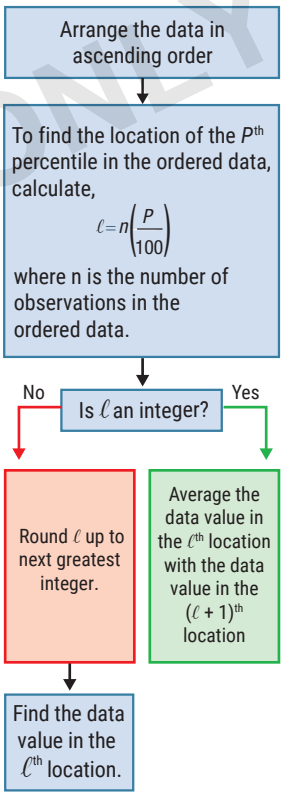


Figure 4.3.1

FOR REVIEW ONLY

For example, if the result of calculating (and rounding up) ℓ is 15, then the desired percentile would be the fifteenth value in the ordered list.

Example 4.3.1

Find the 50th percentile for the following data on the number of spelling errors found on 7 pages of a web site.

3, 5, 0, 1, 9, 2, 7

Solution

Number of observations, $n = 7$.

The percentile, $P = 50$.

The location of the percentile, $\ell = 7 \cdot \left(\frac{50}{100}\right) = 3.5$.

Since the location of the percentile is not an integer, the value is rounded up to 4.

Thus, the fourth observation in the ordered array is the 50th percentile.

0, 1, 2, 3, 5, 7, 9

fourth observation

Therefore, the median value (which is the 50th percentile) is 3.

Example 4.3.2

Suppose that 40 members of your company are given a screening test for a new position. These scores are reported in the first tables below. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10th and 88th percentiles for the test.

Test Scores			
67	45	18	82
45	54	61	55
63	47	21	31
58	46	43	49
34	71	69	56
54	80	73	77
27	70	41	29
66	32	44	33
21	64	52	81
48	55	57	62

Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

Solution

In order to calculate the percentiles, the data must be placed in an ordered array (Ordered Test Scores). To compute the 10th percentile, its position in the ordered

FOR REVIEW ONLY
FOR REVIEW ONLY

array must be determined. The number of observations is $n = 40$. The percentile is $P = 10$. The location of the percentile is found by

$$\ell = 40 \cdot \left(\frac{10}{100} \right) = 4.$$

Since ℓ is an integer, the 4th and 5th observations in the array must be averaged. Since the fourth data value is 27 and the fifth data value is 29, then the 10th percentile is calculated as follows.

$$10^{\text{th}} \text{ percentile} = \frac{27 + 29}{2} = 28$$

To determine the 88th percentile, first calculate its location in the ordered array.

$$\ell = 40 \cdot \left(\frac{88}{100} \right) = 35.2$$

Since the location is not an integer, its value is rounded up to 36. The 36th observation in the ordered array will correspond to the 88th percentile. The 36th value in the table of ordered scores is 73, so 73 is the 88th percentile.

Rounding Rule

When calculating the percentile that corresponds to a particular data value, always round to the nearest whole number.

A slightly different problem connected with percentiles involves taking a raw score and determining its corresponding percentile. Raw scores are usually not very meaningful. If someone scores a 50 on the screening test in the previous example, is that substantially less or about the same as someone who scored a 67 on the same test? To compare these two scores find the percentile of each.

Percentile

The **percentile** of some data value x is given by

$$\text{percentile of } x = \frac{\text{number of data values less than or equal to } x}{\text{total number of data values}} \cdot 100.$$

FORMULA

Note that when finding the percentile of a specific value, if there are multiple occurrences of that value in the data, they all need to be counted in the numerator in order to calculate the percentile. To determine the percentile for a score of 50, the number of data values less than or equal to 50 must be counted. Since there are 18 data values less than or equal to 50, the resulting percentile would be

$$\text{percentile of a score of 50} = \frac{18}{40} \cdot 100 = 45.$$

Hence a score of 50 on the screening test corresponds to the 45th percentile. Thus, approximately 45 percent of the scores are less than or equal to 50. Next, compute the percentile for a score of 67.

$$\text{percentile of a score of 67} = \frac{32}{40} \cdot 100 = 80$$

A score of 67 on the screening test corresponds to the 80th percentile. The score was better than or equal to 80 percent of all other scores on the test. By computing percentiles, we have changed the data's scaling. We see the data from a new

FOR REVIEW ONLY
FOR REVIEW ONLY

Interpreting Percentiles

When students take the SAT, they receive a copy of their scores as well as the percentile they fall into. This percentile can sometimes be confusing. If a student receives a score of 620 on the Critical Reading section, they would fall into the 84th percentile. This means that they received a higher score than 84 percent of the students. The same score in the Mathematics section would place the student into the 80th percentile. Receiving a score of 800 on Critical Reading or Mathematics will put the student in the 99th percentile. This means that less than 1 percent of the students taking the SAT had the same score.

perspective. Using percentiles, it is clear that a score of 67 is significantly better than a score of 50. The 17 point difference in raw score is translated into a 35 percent differential on the percentile scale.

Quartiles

Quartiles

The 25th, 50th, and 75th percentiles are known as **quartiles** and are denoted as Q_1, Q_2 , and Q_3 respectively.

DEFINITION

Quartiles serve as markers that divide a set of data into four equal parts. Q_1 separates the lowest 25 percent, Q_2 represents the median (50th percentile), and Q_3 marks the beginning of the top 25 percent of the data.

Since quartiles are nothing more than percentiles (25th, 50th and 75th), the same methods used to construct percentiles will also produce quartiles. For the screening test data in the previous example, the location of the 25th percentile would be

$$\ell = 40 \cdot \left(\frac{25}{100} \right) = 10.$$

Since the location is an integer, we average the 10th and 11th observation in the ordered data to find the 25th percentile.

$$Q_1 = 25^{\text{th}} \text{ percentile} = \frac{41 + 43}{2} = 42$$

Therefore, we would expect 25 percent of the data to be less than or equal to 42.

The location of the 50th percentile is given by

$$\ell = 40 \cdot \left(\frac{50}{100} \right) = 20.$$

Since the location is an integer, we must average the 20th and 21st observations in order to calculate the percentile.

$$Q_2 = 50^{\text{th}} \text{ percentile} = \frac{54 + 54}{2} = 54,$$

which means that approximately half the data are at or below 54.

The location of the 75th percentile is given by

$$\ell = 40 \cdot \left(\frac{75}{100} \right) = 30.$$

Since the location is an integer, we average the 30th and 31st observations in the ordered array.

$$Q_3 = 75^{\text{th}} \text{ percentile} = \frac{64 + 66}{2} = 65$$

FOR REVIEW ONLY
FOR REVIEW ONLY

This means that approximately 75 percent of the data are less than or equal to 65. The quartiles are useful descriptions of data. They provide a good idea of how the data vary. The **interquartile range** is a measure of dispersion that is calculated using the first and third quartiles.

Interquartile Range

The **interquartile range** is a measure of dispersion which describes the range of the middle fifty percent of the data. It is calculated as follows.

$$\text{IQR} = Q_3 - Q_1$$

FORMULA

For the screening test data, the interquartile range is $65 - 42 = 23$, indicating that the middle 50 percent of the data spans a 23-unit range.

Box Plots – Graphing with Quartiles

A very important use of quartiles is in the construction of box plots. As the name implies, box plots are graphical summaries of the data which, when constructed, have a box-like shape. They provide an alternative method to the histogram for displaying data. **Box plots** are a graphical summary of the central tendency, the spread, the skewness, and the potential existence of outliers in the data. Figure 4.3.2 displays a box plot of the screening test data from Example 4.3.2.

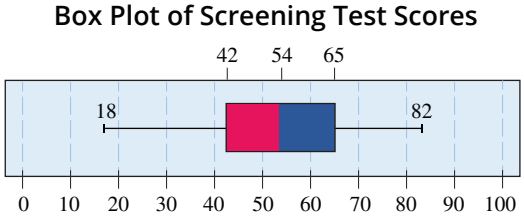


Figure 4.3.2

The box plot is constructed from five summary measures: the largest data value (maximum), the smallest data value (minimum), the 25th percentile (Q_1), the 75th percentile (Q_3), and the median (Q_2).

5-Number Summary

For a set of data, the 5-number summary consists of the following five values:

1. Minimum
2. First quartile, Q_1
3. Second quartile, Q_2, or the median
4. Third quartile, Q_3
5. Maximum

DEFINITION

The lower boundary of the box is the 25th percentile, which is 42 for the screening test data. The upper boundary of the box is the 75th percentile, which is 65 for the screening test data. The median is marked with a line through the box. The median

The Zen of Statistics

Douglas Hofstadler in his book *Godel, Escher, Bach* describes Zen as an attitude in which words and truth are incompatible, or at least that no words can capture truth. If we think of the collected data as truth, statistics is a language whose "words" are pictures and numerical measures which seek to describe that "truth." Despite our best efforts, the statistical language suffers from the same inadequacies as our own language. We ignore the totality of the data in order to summarize it. There is a tradeoff–the loss of the "truth" for a better understanding.

FOR REVIEW ONLY

of the test scores data is 54. Notice that the box itself represents the middle 50% of the data, and the length of the box is the interquartile range.

In Figure 4.3.2, a line is drawn from the 25th percentile to the smallest test score of 18, and another line is drawn from the 75th percentile to the largest score of 82. These lines are often referred to as "whiskers." Adding "whiskers" to the box plot creates a **box and whisker plot**. The box plot for the screening test scores shows that the test score data are slightly skewed to the left. Why? Because the whisker extending from Q_1 appears to be longer than the whisker that extends from Q_3.

Procedure for Constructing a Box Plot

1. Determine the 5-number summary for the data set.
2. Draw a scale that includes the minimum and maximum data values.
3. Construct a box extending from Q_1 to Q_3.
4. Draw a line through the box at the value of the median.
5. Draw lines extending from Q_1 to the minimum and from Q_3 to the maximum.

PROCEDURE

Caution

Different technologies may yield different box plots due to differences in the procedures used for determining quartiles.

CAUTION

Detecting Outliers

The concept of an outlier is an arbitrary concept. What you consider an outlier and what someone else considers an outlier may not be the same thing. However, one definition of an outlier which has gained some acceptance is developed in the context of a box plot.

Outlier

A data point is considered an **outlier** if it is 1.5 times the interquartile range above the 75th percentile or 1.5 times the interquartile range below the 25th percentile.

DEFINITION

If there is an outlier in the data set, the whiskers are drawn to the largest or smallest data point which is within 1.5 times the interquartile range from the box, and the outliers are marked with an a point. For example, suppose test scores of 110 and 2 were added to the screening test data in Example 4.3.2. For the screening test data, a point is considered an outlier if the data point is

- larger than the 75th percentile + 1.5 times the interquartile range = 66 + 1.5(25) = 103.5 or
- smaller than the 25th percentile − 1.5 times the interquartile range = 41 −1.5(25) = 3.5.

FOR REVIEW ONLY
FOR REVIEW ONLY

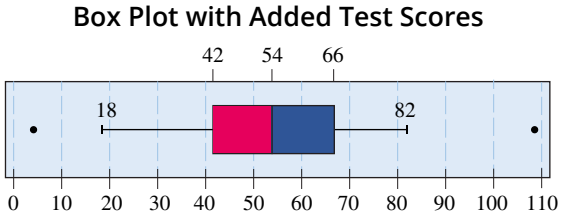


Figure 4.3.3

Since 110 is larger than 103.5, it is considered an outlier. Since 2 is smaller than 3.5, it is considered an outlier. Figure 4.3.3 shows the box plot of the screening test data with the outliers incorporated. Notice that the whiskers did not change because 82 and 18 are still the largest and smallest observations within 1.5 times the interquartile range from the box. This type of box plot is often referred to in statistical software packages as a **modified box plot**.

Although the box plot can be used to display data for a single data set, the histogram is probably more useful for this purpose. The real power of the box plot is the ease with which it allows the comparison of several data sets. Consider the data from the number of wins per baseball team given in Chapter 3. The four data sets are displayed by the box plots in Figure 4.3.4. It is easy to see from the box plots that the center of the Yankees' number of wins is higher than that for the Dodgers, which has a higher center than the center of both the Braves and the Cubs. Note the existence of outliers for both the Cubs and the Dodgers. Also, it appears that the spread of the data, or the variation within the observed values, is not the same for all four data sets. This type of comparison will be used in later chapters to help confirm assumptions which must be made about the data in order to perform statistical inference.

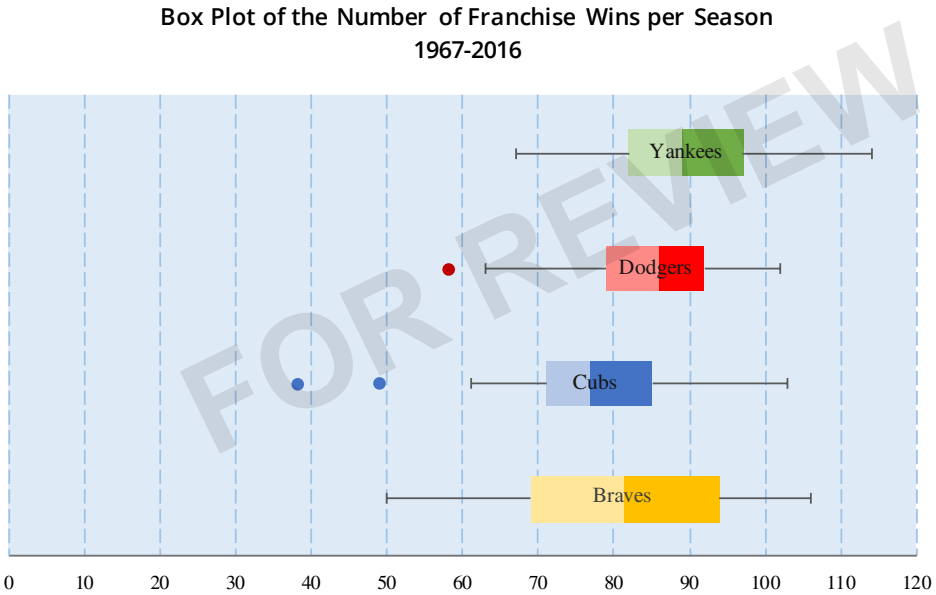


Figure 4.3.4

Technology

Excel is capable of creating a multitude of different charts. For instructions on how to create a box plot, or box-and-whisker plot, using Excel or other technologies, please refer to the web resource at **Tech > Box Plot.**

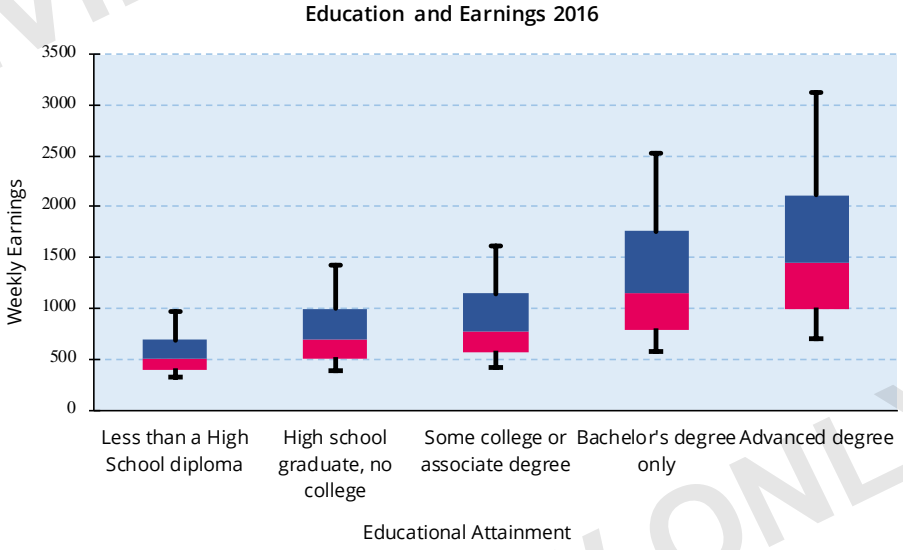
FOR REVIEW ONLY
FOR REVIEW ONLY

Example 4.3.3

Education attainment seems to have a positive association with a person's income.[1] The amount of education a person has can be divided into five categories.

1. Less than a high school diploma
2. High school graduate, no college
3. Some college or associate degree
4. Bachelor's degree only
5. Advanced degree

The categories are ordered according to educational attainment. The vertical box plots in the figure below show the distributions of personal income per each level of education, for individuals age 25 and older. Rather than extending the whiskers to the smallest and largest income values in each category, the whisker endpoints represent the 10^{th} and 90^{th} percentiles. What conclusions can be drawn from this figure?



Solution

It is rather obvious from this figure that more education equates to higher earnings. Also, there is a lot of spread in the last two box plots as shown by the longer lengths of the boxes. This is probably due in part to the market demands for degrees in scientific disciplines, especially computer science.

An interesting set of data is that of the violent crime rate in the United States. The crime rate is defined as the number of reported violent crimes per 100,000 residents. The data from 2010 to 2014 is shown in Figure 4.3.5. Note that there is an outlier for each year that is even more extreme than our definition of an outlier. In fact, for each year the outlier is over three times the median rate of the combined data set. The outlier for each year belongs to the same region of the United States – Washington D.C. Although it may seem that the box plots in the figure suggest a different level of criminality in Washington D.C., it should be noted that there are about 700,000 residents in the district, but commuters from the surrounding suburbs of Maryland and Virginia increase the city's population to over a million during the work week. The crime rate, therefore, may be confounded by the influx of people coming into the area on a daily basis. Usually, outliers that are very far away from the central grouping of data, such as the crime rate in Washington D.C., are removed from the

FOR REVIEW ONLY
FOR REVIEW ONLY

data before any analysis takes place so that the extreme values associated with the outliers do not skew the results.

Data

Data > US Violent Crime by State

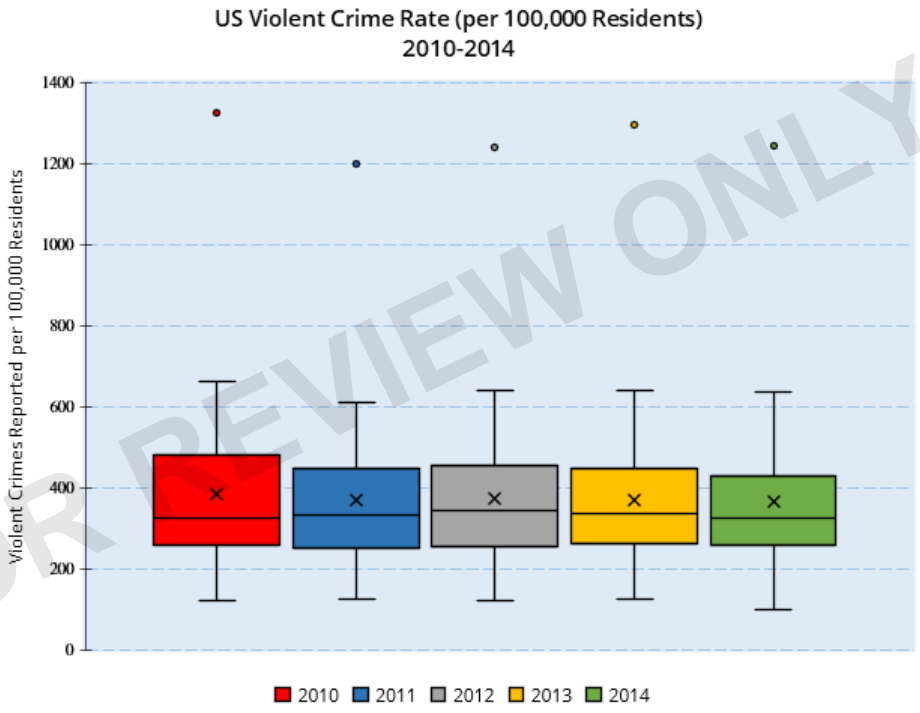


Figure 4.3.5

z-Scores

The ***z*-score** is a standardized measure of relative position with respect to the mean and variability (as measured by the standard deviation) of a data set.

z-Score

The ***z*-score** transforms a data value into the number of standard deviations that value is from the mean.

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

FORMULA

Describing a data value by its number of standard deviations from the mean is a fundamental concept in statistics that is found throughout this book. It is used as a standardization technique, a yardstick, to describe properties of data sets and to compare the relative values of data from different data sets.

Example 4.3.4

Suppose you scored an 86 on your biology test and a 94 on your psychology test. The mean and standard deviations of the two tests are given.

Test Scores		
Course	**Mean**	**Standard Deviation**
Biology	74	10

FOR REVIEW ONLY
FOR REVIEW ONLY

Test Scores		
Course	**Mean**	**Standard Deviation**
Psychology	82	11

What are the z-scores for your two tests? On which of the tests did you perform relatively better?

Solution

The z-score for the biology test is $z = \dfrac{86 - 74}{10} = 1.20.$

The z-score for the psychology test is $z = \dfrac{94 - 82}{11} = 1.09.$

On the biology test you scored 1.2 standard deviations above the mean, compared to only 1.09 standard deviations above the mean for the psychology test. Even though the raw score on the psychology test is larger than the raw score on the biology test, relative to the means and variability in the data sets, the performance on the biology test was slightly better. Once again, changing the scale of the data has beneficial effects. It enables the comparison of two measurements that are drawn from different populations.

Rounding Rule

When calculating a z-score, round to two decimal places.

Properties of a z-Score

- If a z-score is negative, the corresponding data value is less than the mean.
- Conversely, if a z-score is positive, the corresponding data value is greater than the mean.
- The z-score is a unit-free measure. That is, regardless of the original units of measurement (centimeters, meters, or kilometers), an observation's z-score will be the same.

PROPERTIES

FOR REVIEW ONLY
FOR REVIEW ONLY

4.3 Exercises

Basic Concepts

1. What are two methods for describing relative position?
2. If a data value is calculated to be the 72nd percentile, what does this mean?
3. Describe how to find the percentile of a particular value.
4. What are quartiles? Are they equivalent to percentiles? If so, how?
5. What is the interquartile range? What does it measure?
6. What are the advantages of using a box plot to display a data set?
7. What are the key calculations needed in order to construct a box plot?
8. What is an outlier? How can outliers be identified?
9. What is a z-score? Why is it useful?

Exercises

10. Using the "safety_score" variable in the OECD Better Life Index 2016 data set, answer the following questions.
    a. What level of measurement do the data possess?
    b. Calculate the 20th percentile.
    c. Calculate the 95th percentile.
    d. Interpret the meaning of each of these percentiles.

Data

Data > OECD Better Life Index 2016

11. Use the OECD data from the previous problem to find the following.
    a. Determine the percentile rank for Ireland's safety_score. Round to the nearest whole number.
    b. Determine the percentile rank for Chile's safety_score.

12. Using the "Adult.smoking" variable in the US County Data set, answer the following questions.
    a. What level of measurement do the data possess?
    b. Calculate the 20th percentile.
    c. Calculate the 95th percentile.
    d. Interpret the meaning of each of these percentiles.

Data

Data > US County Data

13. Use the US County Data from the previous problem to find the following.
    a. Determine the percentile rank for Lee County, Kentucky's Adult.smoking percentage. Round to the nearest whole number.
    b. Determine the percentile rank for Ozaukee County, Wisconsin's Adult. smoking percentage. Round to the nearest whole number.

FOR REVIEW ONLY
FOR REVIEW ONLY

14. Subjects in a marketing study were shown a film and at the end of the film were given a test to measure their recall. The scores are listed below.

97	31	61	49	61	85	35	57	31	26	27	40	86	78	28
61	87	62	92	58	38	95	81	68	64	72	45	57	84	100

a. What level of measurement do the data possess?

b. Calculate Q_1, the first quartile.

c. Calculate Q_2, the second quartile.

d. Calculate Q_3, the third quartile.

e. Explain the meaning of these percentiles in the context of the marketing study.

f. Calculate the interquartile range.

g. Construct a box plot for the test scores. Are there any outliers?

h. Compute the z-score for a test score of 81.

i. Compute the z-score for a test score of 62.

j. Explain what the z-scores in parts **h.** and **i.** are measuring.

15. Use the marketing study data from the previous problem to find the following.

a. Determine the percentile rank for the subject who scored 49.

b. Determine the percentile rank for the subject who scored 95.

16. Using the on-base percentage (OBP) variable from the Moneyball data set, answer the following questions.

Data

Data > Moneyball

a. What level of measurement do the data possess?

b. Calculate Q_1, the first quartile.

c. Calculate Q_2, the second quartile.

d. Calculate Q_3, the third quartile.

e. Explain the meaning of these percentiles in the context of the on-base percentages.

f. Calculate the interquartile range.

g. Construct a box plot for the on-base percentages. Are there any outliers?

h. Compute the z-score for an on-base percentage of .280.

i. Compute the z-score for an on-base percentage of .355.

j. Explain what the z-scores in parts **h.** and **i.** are measuring.

17. Using the on-base percentage (OBP) variable from the previous problem, find the following.

a. Determine the percentile rank for the Chicago Cubs in the year 2012. Round to the nearest whole number.

b. Determine the percentile rank for the New York Yankees in the year 2012. Round to the nearest whole number.

FOR REVIEW ONLY
FOR REVIEW ONLY

18. Consider a set of data in which the sample mean is 64 and the sample standard deviation is 21. For the following specific values, calculate the z-score and interpret the results.

a. $x = 80$ **b.** $x = 64$ **c.** $x = 40$

19. A statistics student scored a 75 on the first exam of the semester and an 82 on the second exam of the semester. The average score and standard deviation of scores for the two exams are given in the following table. On which exam did the student perform relatively better?

Test Scores		
Statistic	**First Exam**	**Second Exam**
μ	74	85
σ	10	7

20. A hospital measures babies' heights when they are born in both inches and centimeters. Eight baby girls are randomly selected and the following heights are recorded in both inches and centimeters.

Newborn Heights								
Baby	1	2	3	4	5	6	7	8
Inches	17.75	18.50	19.25	19.75	20.25	20.50	20.50	20.75
Centimeters	45.09	46.99	48.90	50.17	51.44	52.07	52.07	52.71

a. Calculate the mean height in inches and centimeters for the baby girls.

b. Calculate the standard deviation of the heights of baby girls in both inches and centimeters.

c. Calculate the z-score for the height of Baby Girl 3 measured in inches.

d. For Baby Girl 3, calculate the z-score for the heights measured in centimeters.

e. Consider the z-scores calculated in parts **c.** and **d.** Are the z-scores as you expected them to be? Explain.

FOR REVIEW ONLY
FOR REVIEW ONLY