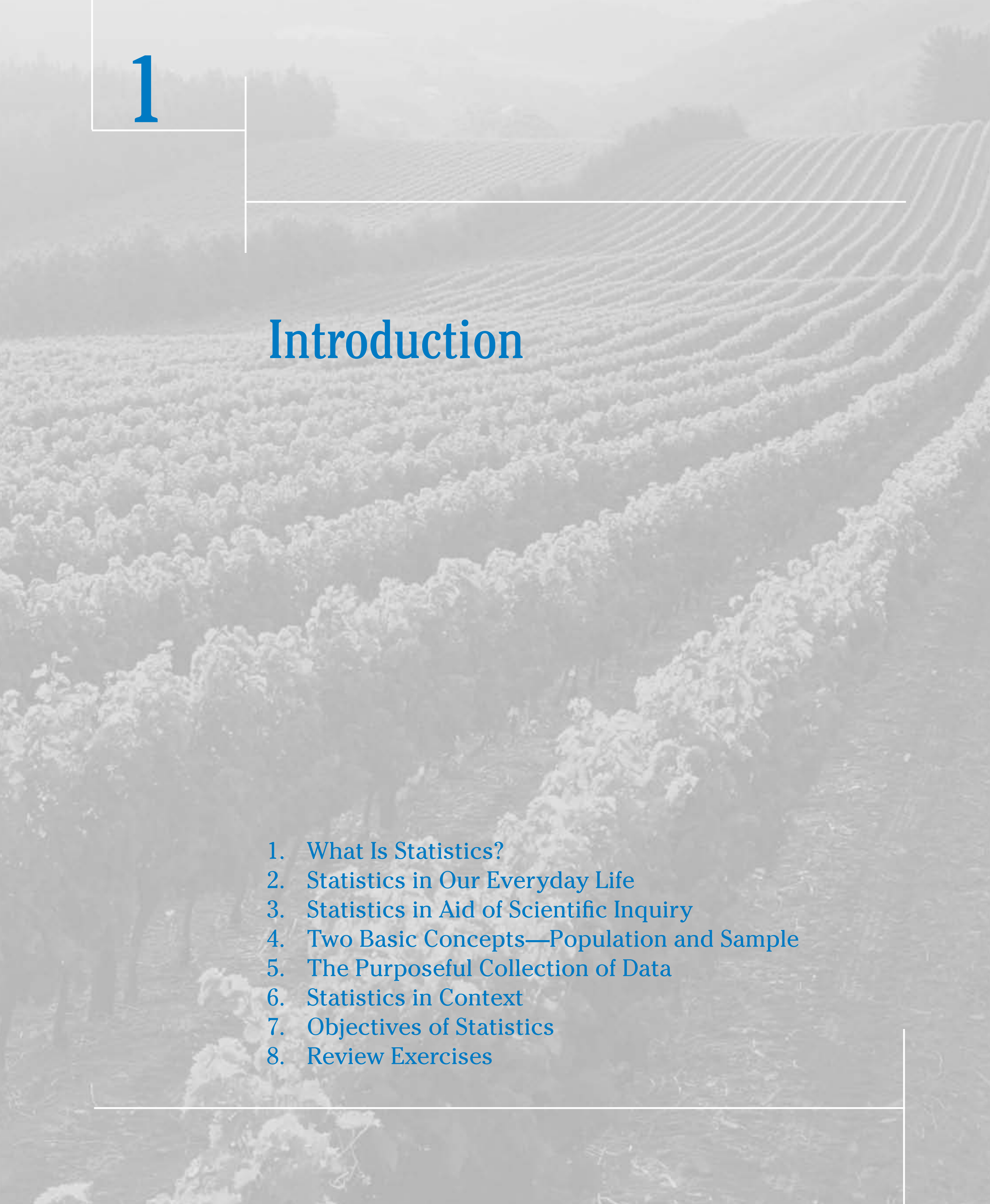


1

Introduction

1. What Is Statistics?
2. Statistics in Our Everyday Life
3. Statistics in Aid of Scientific Inquiry
4. Two Basic Concepts—Population and Sample
5. The Purposeful Collection of Data
6. Statistics in Context
7. Objectives of Statistics
8. Review Exercises

Surveys Provide Information About the Population

What is your favorite spectator sport?

Football	36.4%
Baseball	12.7%
Basketball	12.5%
Other	38.4%

College and professional sports are combined in our summary.[1] Clearly, football is the most popular spectator sport. Actually, the National Football League by itself is more popular than baseball.

Until the mid 1960s, baseball was most popular according to similar surveys. Surveys, repeated at different times, can detect trends in opinion.

Hometown fans attending today's game are but a sample of the population of all local football fans. A self-selected sample may not be entirely representative of the population on issues such as ticket price increases. Kiichiro Sato/ © AP/Wide World Photos

[1] These percentages are similar to those obtained by the ESPN Sports Poll, a service of TNS, in a 2007 poll of over 27,000 fans.

1. WHAT IS STATISTICS?

The word **statistics** originated from the Latin word “status,” meaning “state.” For a long time, it was identified solely with the displays of data and charts pertaining to the economic, demographic, and political situations prevailing in a country. Even today, a major segment of the general public thinks of statistics as synonymous with forbidding arrays of numbers and myriad graphs. This image is enhanced by numerous government reports that contain a massive compilation of numbers and carry the word statistics in their titles: “Statistics of Farm Production,” “Statistics of Trade and Shipping,” “Labor Statistics,” to name a few. However, gigantic advances during the twentieth century have enabled statistics to grow and assume its present importance as a discipline of data-based reasoning. Passive display of numbers and charts is now a minor aspect of statistics, and few, if any, of today’s statisticians are engaged in the routine activities of tabulation and charting.

What, then, are the role and principal objectives of statistics as a scientific discipline? Stretching well beyond the confines of data display, statistics deals with collecting informative data, interpreting these data, and drawing conclusions about a phenomenon under study. The scope of this subject naturally extends to all processes of acquiring knowledge that involve fact finding through collection and examination of data. Opinion polls (surveys of households to study sociological, economic, or health-related issues), agricultural field experiments (with new seeds, pesticides, or farming equipment), clinical studies of vaccines, and cloud seeding for artificial rain production are just a few examples. The principles and methodology of statistics are useful in answering questions such as, What kind and how much data need to be collected? How should we organize and interpret the data? How can we analyze the data and draw conclusions? How do we assess the strength of the conclusions and gauge their uncertainty?

> **Statistics** as a subject provides a body of principles and methodology for designing the process of data collection, summarizing and interpreting the data, and drawing conclusions or generalities.

2. STATISTICS IN OUR EVERYDAY LIFE

Fact finding through the collection and interpretation of data is not confined to professional researchers. In our attempts to understand issues of environmental protection, the state of unemployment, or the performance of competing football teams, numerical facts and figures need to be reviewed and interpreted. In our day-to-day life, learning takes place through an often implicit analysis of factual information.

We are all familiar to some extent with reports in the news media on important statistics.

Employment. Monthly, as part of the Current Population Survey, the Bureau of Census collects information about employment status from a sample of about 65,000 households. Households are contacted on a rotating basis with three-fourths of the sample remaining the same for any two consecutive months.

The survey data are analyzed by the Bureau of Labor Statistics, which reports monthly unemployment rates. □

Cost of Living. The consumer price index (CPI) measures the cost of a fixed market basket of over 400 goods and services. Each month, prices are obtained from a sample of over 18,000 retail stores that are distributed over 85 metropolitan areas. These prices are then combined taking into account the relative quantity of goods and services required by a hypothetical "1967 urban wage earner." Let us not be concerned with the details of the sampling method and calculations as these are quite intricate. They are, however, under close scrutiny because of the importance to the hundreds of thousands of Americans whose earnings or retirement benefits are tied to the CPI. □

Election time brings the pollsters into the limelight.

Gallup Poll. This, the best known of the national polls, produces estimates of the percentage of popular vote for each candidate based on interviews with a minimum of 1500 adults. Beginning several months before the presidential election, results are regularly published. These reports help predict winners and track changes in voter preferences. □

Our sources of factual information range from individual experience to reports in news media, government records, and articles in professional journals. As consumers of these reports, citizens need some idea of statistical reasoning to properly interpret the data and evaluate the conclusions. Statistical reasoning provides criteria for determining which conclusions are supported by the data and which are not. The credibility of conclusions also depends greatly on the use of statistical methods at the data collection stage. Statistics provides a key ingredient for any systematic approach to improve any type of process from manufacturing to service.

Quality and Productivity Improvement. In the past 30 years, the United States has faced increasing competition in the world marketplace. An international revolution in quality and productivity improvement has heightened the pressure on the U.S. economy. The ideas and teaching of W. Edwards Deming helped rejuvenate Japan's industry in the late 1940s and 1950s. In the 1980s and 1990s, Deming stressed to American executives that, in order to survive, they must mobilize their work force to make a continuing commitment to quality improvement. His ideas have also been applied to government. The city of Madison, WI, has implemented quality improvement projects in the police department and in bus repair and scheduling. In each case, the project goal was better service at less cost. Treating citizens as the customers of government services, the first step was to collect information from them in order to identify situations that needed improvement. One end result was the strategic placement of a new police substation and a subsequent increase in the number of foot patrol persons to interact with the community.

Statistical reasoning can guide the purposeful collection and analysis of data toward the continuous improvement of any process. © Andrew Sacks/Stone/Getty Images

Once a candidate project is selected for improvement, data must be collected to assess the current status and then more data collected on the effects of possible changes. At this stage, statistical skills in the collection and presentation of summaries are not only valuable but necessary for all participants.

In an industrial setting, statistical training for all employees—production line and office workers, supervisors, and managers—is vital to the quality transformation of American industry. □

3. STATISTICS IN AID OF SCIENTIFIC INQUIRY

The phrase scientific inquiry refers to a systematic process of learning. A scientist sets the goal of an investigation, collects relevant factual information (or data), analyzes the data, draws conclusions, and decides further courses of action. We briefly outline a few illustrative scenarios.

Training Programs. Training or teaching programs in many fields designed for a specific type of clientele (college students, industrial workers, minority groups, physically handicapped people, retarded children, etc.) are continually monitored, evaluated, and modified to improve their usefulness to society. To learn about the comparative effectiveness of different programs, it is essential to collect data on the achievement or growth of skill of the trainees at the completion of each program. □

Monitoring Advertising Claims. The public is constantly bombarded with commercials that claim the superiority of one product brand in comparison to others. When such comparisons are founded on sound experimental evidence, they serve to educate the consumer. Not infrequently, however, misleading advertising claims are made due to insufficient experimentation, faulty analysis of data, or even blatant manipulation of experimental results. Government agencies and consumer groups must be prepared to verify the comparative quality of products by using adequate data collection procedures and proper methods of statistical analysis. □

Plant Breeding. To increase food production, agricultural scientists develop new hybrids by cross-fertilizing different plant species. Promising new strains need to be compared with the current best ones. Their relative productivity is assessed by planting some of each variety at a number of sites. Yields are recorded and then analyzed for apparent differences. The strains may also be compared on the basis of disease resistance or fertilizer requirements. □

Genomics. This century's most exciting scientific advances are occurring in biology and genetics. Scientists can now study the genome, or sum total of all of a living organism's genes. The human DNA sequence is now known along with the DNA sequences of hundreds of other organisms.

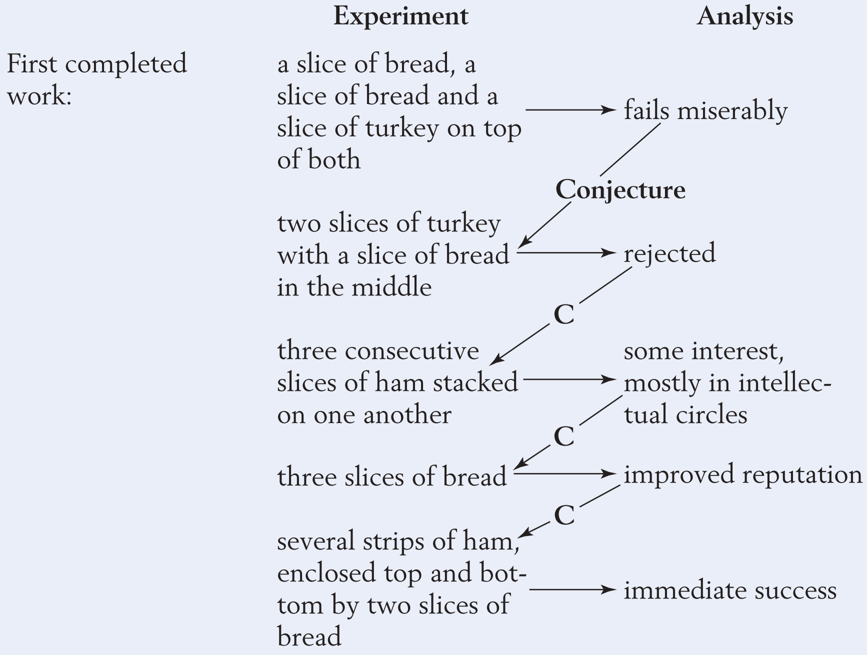
A primary goal of many studies is to identify the specific genes and related genetic states that give rise to complex traits (e.g., diabetes, heart disease, cancer). New instruments for measuring genes and their products are continually being developed. One popular technology is the microarray, a rectangular array of tens of thousands of genes. The power of microarray technologies derives from the ability to compare, for instance, healthy and diseased tissue. Two-color microarrays have two kinds of DNA material deposited at each site in the array. Due to the impact

Statistically designed experiments are needed to document the advantages of the new hybrid versus the old species. © Mitch Wojnarowicz/The Image Works of the disease and the availability of human tumor specimens, many early microarray studies focused on human cancer. Significant advances have been made in cancer classification, knowledge of cancer biology, and prognostic prediction. A hallmark example of the power of microarrays used in prognostic prediction is Mammaprint approved by the FDA in 2007. This, the first approved microarray based test, classifies a breast cancer patient as low or high risk for recurrence.

This is clearly only the beginning, as numerous groups are employing microarrays and other high-throughput technologies in their research studies. Typically, genomics experiments feature the simultaneous measurement of a great number of responses. As more and more data are collected, there is a growing need for novel statistical methods for analyzing data and thereby addressing critical scientific questions. Statisticians and other computational scientists are playing a major role in these efforts to better human health. □

Factual information is crucial to any investigation. The branch of statistics called **experimental design** can guide the investigator in planning the manner and extent of data collection.

The Conjecture-Experiment-Analysis Learning Cycle Invention of the Sandwich by the Earl of Sandwich (According to Woody Allen, Humorist)*



Experiment

Analysis

First completed work:

a slice of bread, a slice of bread and a slice of turkey on top of both → fails miserably

Conjecture

two slices of turkey with a slice of bread in the middle → rejected

C

three consecutive slices of ham stacked on one another → some interest, mostly in intellectual circles

C

three slices of bread → improved reputation

C

several strips of ham, enclosed top and bottom by two slices of bread → immediate success

*Copyright © 1966 by Woody Allen. Adapted by permission of Random House, Inc. from *Getting Even*, by Woody Allen.

After the data are collected, statistical methods are available that summarize and describe the prominent features of data. These are commonly known as **descriptive statistics.** Today, a major thrust of the subject is the evaluation of information present in data and the assessment of the new learning gained from this information. This is the area of **inferential statistics** and its associated methods are known as the methods of **statistical inference.**

It must be realized that a scientific investigation is typically a process of trial and error. Rarely, if ever, can a phenomenon be completely understood or a theory perfected by means of a single, definitive experiment. It is too much to expect to get it all right in one shot. Even after his first success with the electric light bulb, Thomas Edison had to continue to experiment with numerous materials for the filament before it was perfected. Data obtained from an experiment provide new knowledge. This knowledge often suggests a revision of an existing theory, and this itself may require further investigation through more experiments and analysis of data. Humorous as it may appear, the excerpt boxed above from a Woody Allen writing captures the vital point that a scientific process of learning is essentially iterative in nature.

4. TWO BASIC CONCEPTS—POPULATION AND SAMPLE

In the preceding sections, we cited a few examples of situations where evaluation of factual information is essential for acquiring new knowledge. Although these examples are drawn from widely differing fields and only sketchy descriptions of the scope and objectives of the studies are provided, a few common characteristics are readily discernible.

First, in order to acquire new knowledge, relevant data must be collected. Second, some amount of variability in the data is unavoidable even though observations are made under the same or closely similar conditions. For instance, the treatment for an allergy may provide long-lasting relief for some individuals whereas it may bring only transient relief or even none at all to others. Likewise, it is unrealistic to expect that college freshmen whose high school records were alike would perform equally well in college. Nature does not follow such a rigid law.

A third notable feature is that access to a complete set of data is either physically impossible or from a practical standpoint not feasible. When data are obtained from laboratory experiments or field trials, no matter how much experimentation has been performed, more can always be done. In public opinion or consumer expenditure studies, a complete body of information would emerge only if data were gathered from every individual in the nation—undoubtedly a monumental if not an impossible task. To collect an exhaustive set of data related to the damage sustained by all cars of a particular model under collision at a specified speed, every car of that model coming off the production lines would have to be subjected to a collision! Thus, the limitations of time, resources, and facilities, and sometimes the destructive nature of the testing, mean that we must work with incomplete information—the data that are actually collected in the course of an experimental study.

The preceding discussions highlight a distinction between the data set that is actually acquired through the process of observation and the vast collection of all potential observations that can be conceived in a given context. The statistical name for the former is **sample;** for the latter, it is **population,** or **statistical population.** To further elucidate these concepts, we observe that each measurement in a data set originates from a distinct source which may be a patient, tree, farm, household, or some other entity depending on the object of a study. The source of each measurement is called a **sampling unit,** or simply, a **unit.**

To emphasize population as the entire collection of units, we term it the **population of units.**

> A **unit** is a single entity, usually a person or an object, whose characteristics are of interest.
>
> The **population of units** is the complete collection of units about which information is sought.

There is another aspect to any population and that is the value, for each unit, of a characteristic or variable of interest. There can be several characteristics of interest for a given population of units, as indicated in Table 1.

TABLE 1 Populations, Units, and Variables

Population	Unit	Variables/Characteristics
Registered voters in your state	Voter	Political party Voted or not in last election Age Sex Conservative/liberal
All rental apartments near campus	Apartment	Rent Size in square feet Number of bedrooms Number of bathrooms TV and Internet connections
All campus fast food restaurants	Restaurant	Number of employees Seating capacity Hiring/not hiring
All computers owned by students at your school	Computer	Speed of processor Size of hard disk Speed of Internet connection Screen size

For a given variable or characteristic of interest, we call the collection of values, evaluated for every unit in the population, the **statistical population** or just the **population.** We refer to the collection of units as the **population of units** when there is a need to differentiate it from the collection of values.

A statistical **population** is the set of measurements (or record of some qualitative trait) corresponding to the entire collection of units about which information is sought.

The population represents the target of an investigation. We learn about the population by taking a sample from the population. A **sample** or **sample data set** then consists of measurements recorded for those units that are actually observed. It constitutes a part of a far larger collection about which we wish to make inferences—the set of measurements that would result if all the units in the population could be observed.

A **sample** from a statistical population is the subset of measurements that are actually collected in the course of an investigation.

Example 1 Identifying the Population and Sample

Questions concerning the effect on health of two or fewer cups of coffee a day are still largely unresolved. Current studies seek to find physiological changes that could prove harmful. An article carried the headline CAFFEINE DECREASES CEREBRAL BLOOD FLOW. It describes a study[2] which establishes a physiological side effect—a substantial decrease in cerebral blood flow for persons drinking two to three cups of coffee daily.

The cerebral blood flow was measured twice on each of 20 subjects. It was measured once after taking an oral dose of caffeine equivalent to two to three cups of coffee and then, on another day, after taking a look-alike dose but without caffeine. The order of the two tests was random and subjects were not told which dose they received. The measured decrease in cerebral blood flow was significant.

Identify the population and sample.

SOLUTION As the article implies, the conclusion should apply to you and me. The population could well be the potential decreases in cerebral blood flow for all adults living in the United States. It might even apply to all the decrease in blood flow for all caffeine users in the world, although the cultural customs

[2] A. Field et al. "Dietary Caffeine Consumption and Withdrawal: Confounding Variables in Quantitative Cerebral Perfusion Studies?" *Radiology* **227** (2003), pp. 129–135.

may vary the type of caffeine consumption from coffee breaks to tea time to kola nut chewing.

The sample consists of the decreases in blood flow for the 20 subjects who agreed to participate in the study.

Example 2 A Misleading Sample

A host of a radio music show announced that she wants to know which singer is the favorite among city residents. Listeners were then asked to call in and name their favorite singer.

Identify the population and sample. Comment on how to get a sample that is more representative of the city's population.

SOLUTION The population is the collection of singer preferences of all city residents and the purported goal was to learn who was the favorite singer. Because it would be nearly impossible to question all the residents in a large city, one must necessarily settle for taking a sample.

Having residents make a local call is certainly a low-cost method of getting a sample. The sample would then consist of the singers named by each person who calls the radio station. Unfortunately, with this selection procedure, the sample is not very representative of the responses from all city residents. Those who listen to the particular radio station are already a special subgroup with similar listening tastes. Furthermore, those listeners who take the time and effort to call are usually those who feel strongest about their opinions. The resulting responses could well be much stronger in favor of a particular country western or rock singer than is the case for preference among the total population of city residents or even those who listen to the station.

If the purpose of asking the question is really to determine the favorite singer of the city's residents, we have to proceed otherwise. One procedure commonly employed is a phone survey where the phone numbers are chosen at random. For instance, one can imagine that the numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 are written on separate pieces of paper and placed in a hat. Slips are then drawn one at a time and replaced between drawings. Later, we will see that computers can mimic this selection quickly and easily. Four draws will produce a random telephone number within a three-digit exchange. Telephone numbers chosen in this manner will certainly produce a much more representative sample than the self-selected sample of persons who call the station.

Self-selected samples consisting of responses to call-in or write-in requests will, in general, not be representative of the population. They arise primarily from subjects who feel strongly about the issue in question. To their credit, many TV news and entertainment programs now state that their call-in polls are nonscientific and merely reflect the opinions of those persons who responded.

USING A RANDOM NUMBER TABLE TO SELECT A SAMPLE

The choice of which population units to include in a sample must be impartial and objective. When the total number of units is finite, the name or number of each population unit could be written on a separate slip of paper and the slips placed in a box. Slips could be drawn one at a time without replacement and the corresponding units selected as the sample of units. Unfortunately, this simple and intuitive procedure is cumbersome to implement. Also, it is difficult to mix the slips well enough to ensure impartiality.

Alternatively, a better method is to take 10 identical marbles, number them 0 through 9, and place them in an urn. After shuffling, select 1 marble. After replacing the marble, shuffle and draw again. Continuing in this way, we create a sequence of random digits. Each digit has an equal chance of appearing in any given position, all pairs have the same chance of appearing in any two given positions, and so on. Further, any digit or collection of digits is unrelated to any other disjoint subset of digits. For convenience of use, these digits can be placed in a table called a **random number table.**

The digits in Table 1 of Appendix B were actually generated using computer software that closely mimics the drawing of marbles. A portion of this table is shown here as Table 2.

To obtain a random sample of units from a population of size *N*, we first number the units from 1 to *N*. Then numbers are read from the table of random digits until enough different numbers in the appropriate range are selected.

TABLE 2 Random Digits: A Portion of Table 1, Appendix B

Row										
1	0695	7741	8254	4297	0000	5277	6563	9265	1023	5925
2	0437	5434	8503	3928	6979	9393	8936	9088	5744	4790
3	6242	2998	0205	5469	3365	7950	7256	3716	8385	0253
4	7090	4074	1257	7175	3310	0712	4748	4226	0604	3804
5	0683	6999	4828	7888	0087	9288	7855	2678	3315	6718
6	7013	4300	3768	2572	6473	2411	6285	0069	5422	6175
7	8808	2786	5369	9571	3412	2465	6419	3990	0294	0896
8	9876	3602	5812	0124	1997	6445	3176	2682	1259	1728
9	1873	1065	8976	1295	9434	3178	0602	0732	6616	7972
10	2581	3075	4622	2974	7069	5605	0420	2949	4387	7679
11	3785	6401	0540	5077	7132	4135	4646	3834	6753	1593
12	8626	4017	1544	4202	8986	1432	2810	2418	8052	2710
13	6253	0726	9483	6753	4732	2284	0421	3010	7885	8436
14	0113	4546	2212	9829	2351	1370	2707	3329	6574	7002
15	4646	6474	9983	8738	1603	8671	0489	9588	3309	5860

Example 3 Using the Table of Random Digits to Select Items for a Price Check

One week, the advertisement for a large grocery store contains 72 special sale items. Five items will be selected with the intention of comparing the sales price with the scan price at the checkout counter. Select the five items at random to avoid partiality.

SOLUTION The 72 sale items are first numbered from 1 to 72. Since the population size $N = 72$ has two digits, we will select random digits two at a time from Table 2. Arbitrarily, we decide to start in row 7 and columns 19 and 20. Starting with the two digits in columns 19 and 20 and reading down, we obtain

12 97 34 69 32 86 32 51

We ignore 97 and 86 because they are larger than the population size 72. We also ignore any number when it appears a second time as 32 does here. Consequently, the sale items numbered

12 34 69 32 51

are selected for the price check.

For large sample size situations or frequent applications, it is often more convenient to use computer software to choose the random numbers.

Example 4 Selecting a Sample by Random Digit Dialing

A major Internet service provider wants to learn about the proportion of people in one target area who are aware of its latest product. Suppose there is a single three-digit telephone exchange that covers the target area. Use Table 1, in Appendix B, to select six telephone numbers for a phone survey.

SOLUTION We arbitrarily decide to start at row 31 and columns 25 to 28. Proceeding upward, we obtain

7566 0766 1619 9320 1307 6435

Together with the three-digit exchange, these six numbers form the phone numbers called in the survey. Every phone number, listed or unlisted, has the same chance of being selected. The same holds for every pair, every triplet, and so on. Commercial phones may have to be discarded and another four digits selected. If there are two exchanges in the area, separate selections could be done for each exchange.

For large sample sizes, it is better to use computer-generated random digits or even computer-dialed random phone numbers.

Data collected with a clear-cut purpose in mind are very different from **anecdotal data.** Most of us have heard people say they won money at a casino, but certainly most people cannot win most of the time as casinos are not in the business of giving away money. People tend to tell good things about themselves. In a similar vein, some drivers' lives are saved when they are thrown free of car wrecks because they were not wearing seat belts. Although such stories are told and retold, you must remember that there is really no opportunity to hear from those who would have lived if they had worn their seat belts. Anecdotal information is usually repeated because it has some striking feature that may not be representative of the mass of cases in the population. Consequently, it is not apt to provide reliable answers to questions.

5. THE PURPOSEFUL COLLECTION OF DATA

Many poor decisions are made, in both business and everyday activities, because of the failure to understand and account for variability. Certainly, the purchasing habits of one person may not represent those of the population, or the reaction of one mouse, on exposure to a potentially toxic chemical compound, may not represent that of a large population of mice. However, despite diversity among the purchasing habits of individuals, we can obtain accurate information about the purchasing habits of the population by collecting data on a large number of persons. By the same token, much can be learned about the toxicity of a chemical if many mice are exposed.

Just making the decision to collect data to answer a question, to provide the basis for taking action, or to improve a process is a key step. Once that decision has been made, an important next step is to develop a **statement of purpose** that is both specific and unambiguous. If the subject of the study is public transportation being behind schedule, you must carefully specify what is meant by late. Is it 1 minute, 5 minutes, or more than 10 minutes behind scheduled times that should result in calling a bus or commuter train late? Words like soft or uncomfortable in a statement are even harder to quantify. One common approach, for a quality like comfort, is to ask passengers to rate the ride on public transportation on the five-point scale

1	2	3	4	5
Very uncomfortable		Neutral		Very comfortable

where the numbers 1 through 5 are attached to the scale, with 1 for very uncomfortable and so on through 5 for very comfortable.

We might conclude that the ride is comfortable if the majority of persons in the sample check either of the top two boxes.

Example 5 A Clear Statement of Purpose Concerning Water Quality

Each day, a city must sample the lake water in and around a swimming beach to determine if the water is safe for swimming. During late summer, the primary difficulty is algae growth and the safe limit has been set in terms of water clarity.

SOLUTION The problem is already well defined so the statement of purpose is straightforward.

> ***PURPOSE:*** Determine whether or not the water clarity at the beach is below the safe limit.

The city has already decided to take measurements of clarity at three separated locations. In Chapter 8, we will learn how to decide if the water is safe despite the variation in the three sample values.

The overall purpose can be quite general but a specific statement of purpose is required at each step to guide the collection of data. For instance:

> ***GENERAL PURPOSE:*** Design a data collection and monitoring program at a completely automated plant that handles radioactive materials.

One issue is to ensure that the production plant will shut down quickly if materials start accumulating anywhere along the production line. More specifically, the weight of materials could be measured at critical positions. A quick shutdown will be implemented if any of these exceed a safe limit. For this step, a statement of purpose could be:

> ***PURPOSE:*** Implement a fast shutdown if the weight at any critical position exceeds 1.2 kilograms.

The safe limit 1.2 kilograms should be obtained from experts; preferrably it would be a consensus of expert opinion.

There still remain statistical issues of how many critical positions to choose and how often to measure the weight. These are followed with questions on how to analyze data and specify a rule for implementing a fast shutdown.

A clearly specified statement of purpose will guide the choice of what data to collect and help ensure that it will be relevant to the purpose. Without a clearly specified purpose, or terms unambiguously defined, much effort can be wasted in collecting data that will not answer the question of interest.

6. STATISTICS IN CONTEXT

A primary health facility became aware that sometimes it was taking too long to return patients' phone calls. That is, patients would phone in with requests for information. These requests, in turn, had to be turned over to doctors or nurses who would collect the information and return the call. The overall objective was to understand the current procedure and then improve on it. As a good first step, it was decided to find how long it was taking to return calls under the current procedure. Variation in times from call to call is expected, so the purpose of the initial investigation is to benchmark the variability with the current procedure by collecting a sample of times.

> ***PURPOSE:*** Obtain a reference or benchmark for the current procedure by collecting a sample of times to return a patient's call under the current procedure.

For a sample of incoming calls collected during the week, the time received was noted along with the request. When the return call was completed, the elapsed time, in minutes, was recorded. Each of these times is represented as a dot in Figure 1. Notice that over one-third of the calls took over 120 minutes, or over two hours, to return. This could be a long time to wait for information if it concerns a child with a high fever or an adult with acute symptoms. If the purpose was to determine what proportion of calls took too long to return, we would need to agree on a more precise definition of "too long" in terms of number of minutes. Instead, these data clearly indicate that the process needs improvement and the next step is to proceed in that direction.

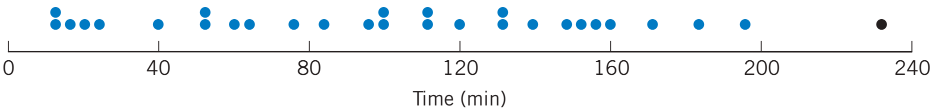


Figure 1 Time in minutes to return call.

In any context, to pursue potential improvements of a process, one needs to focus more closely on particulars. Three questions

When Where Who

should always be asked before gathering further data. More specifically, data should be sought that will answer the following questions.

When do the difficulties arise? Is it during certain hours, certain days of the week or month, or in coincidence with some other activities?

Where do the difficulties arise? Try to identify the locations of bottlenecks and unnecessary delays.

Who was performing the activity and who was supervising? The idea is not to pin blame, but to understand the roles of participants with the goal of making improvements.

It is often helpful to construct a **cause-and-effect diagram** or **fishbone diagram.** The main centerline represents the problem or the effect. A somewhat simplified fishbone chart is shown in Figure 2 for the *where* question regarding the location of delays when returning patients' phone calls. The main centerline represents the problem: Where are delays occurring? Calls come to the reception desk, but when these lines are busy, the calls go directly to nurses on the third or fourth floor. The main diagonal arms in Figure 2 represent the floors and the smaller horizontal lines more specific locations on the floor where the delay could occur. For instance, the horizontal line representing a delay in retrieving a patient's medical record connects to the second floor diagonal line. The resulting figure resembles the skeleton of a fish. Consideration of the diagram can help guide the choice of what new data to collect.

Fortunately, the quality team conducting this study had already given preliminary consideration to the *When*, *Where*, and *Who* questions and recorded not only the time of day but also the day and person receiving the call. That is, their current data gave them a start on determining if the time to return calls depends on when or where the call is received.

Although we go no further with this application here, the quality team next developed more detailed diagrams to study the flow of paper between the time the call is received and when it is returned. They then identified bottlenecks in the flow of information that were removed and the process was improved. In later chapters, you will learn how to compare and display data from two locations or old and new processes, but the key idea emphasized here is the purposeful collection of relevant data.

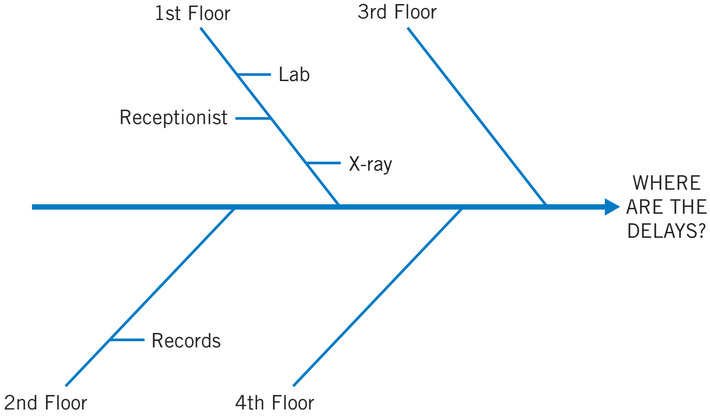


Figure 2 A cause-and-effect diagram for the location of delays.

7. OBJECTIVES OF STATISTICS

The subject of statistics provides the methodology to make **inferences** about the population from the collection and analysis of sample data. These methods enable one to derive plausible generalizations and then assess the extent of uncertainty underlying these generalizations. Statistical concepts are also essential during the planning stage of an investigation when decisions must be made as to the mode and extent of the sampling process.

The major objectives of statistics are:

1. To make **inferences** about a population from an analysis of information contained in sample data. This includes assessments of the extent of uncertainty involved in these inferences.
2. To **design the process and the extent of sampling** so that the observations form a basis for drawing valid inferences.

The design of the sampling process is an important step. A good design for the process of data collection permits efficient inferences to be made, often with a straightforward analysis. Unfortunately, even the most sophisticated methods of data analysis cannot, in themselves, salvage much information from data that are produced by a poorly planned experiment or survey.

The early use of statistics in the compilation and passive presentation of data has been largely superseded by the modern role of providing analytical tools with which data can be efficiently gathered, understood, and interpreted. Statistical concepts and methods make it possible to draw valid conclusions about the population on the basis of a sample. Given its extended goal, the subject of statistics has penetrated all fields of human endeavor in which the evaluation of information must be grounded in data-based evidence.

The basic statistical concepts and methods described in this book form the core in all areas of application. We present examples drawn from a wide range of applications to help develop an appreciation of various statistical methods, their potential uses, and their vulnerabilities to misuse.

USING STATISTICS WISELY

1. Compose a clear statement of purpose and use it to help decide upon which variables to observe.
2. Carefully define the population of interest.
3. Whenever possible, select samples using a random device or random number table.
4. Do not unquestionably accept conclusions based on self-selected samples.
5. Remember that conclusions reached in TV, magazine, or newspaper reports might not be as obvious as reported. When reading or listening to reports, you must be aware that the advocate, often a politician or advertiser, may only be presenting statistics that emphasize positive features.

KEY IDEAS

Before gathering data, on a characteristic of interest, identify a **unit** or **sampling unit.** This is usually a person or object. The **population of units** is the complete collection of units. In statistics we concentrate on the collection of values of the characteristic, or record of a qualitative trait, evaluated for each unit in the population. We call this the **statistical population** or just the **population.**

A **sample** or **sample data set** from the population is the subset of measurements that are actually collected.

Statistics is a body of principles that helps to first design the process and extent of sampling and then guides the making of **inferences** about the population **(inferential statistics). Descriptive statistics** help summarize the sample. Procedures for **statistical inference** allow us to make generalizations about the population from the information in the sample.

A **statement of purpose** is a key step in designing the data collection process.

8. REVIEW EXERCISES

1.1 A newspaper headline reads,

U.S. TEENS TRUST, FEAR THEIR PEERS

and the article explains that a telephone poll was conducted of 1055 persons 13 to 17 years old. Identify a statistical population and the sample.

1.2 Consider the population of all students at your college. You want to learn about total monthly entertainment expenses for a student.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.3 Consider the population of persons living in Chicago. You want to learn about the proportion which are illegal aliens.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.4 A student is asked to estimate the mean height of all male students on campus. She decides to use the heights of members of the basketball team because they are conveniently printed in the game program.

(a) Identify the statistical population and the sample.

(b) Comment on the selection of the sample.

(c) How should a sample of males be selected?

1.5 Psychologists[3] asked 46 golfers, after they played a round, to estimate the diameter of the hole on the green by visually selecting one of nine holes cut in a board.

(a) Specify the population unit.

(b) Specify the statistical population and sample.

1.6 A phone survey in 2008[4] of 1010 adults included a response to the number of leisure hours per week. Identify the population unit, statistical population, and sample.

1.7 It is often easy to put off doing an unpleasant task. At a Web site,[5] persons can take a test and receive a score that determines if they have a serious problem with procrastination. Should the scores from people who take this test on-line be considered a random sample? Explain your reasoning.

1.8 A magazine that features the latest electronics and computer software for homes enclosed a short questionnaire on a postcard. Readers were asked to answer questions concerning their use and ownership of various software and hardware products, and to then send the card to the publisher. A summary of the results appeared in a later issue of the magazine that used the data to make statements such as 40% of readers have purchased program X. Identify a population and sample and comment on the representativeness of the sample. Are readers who have not purchased any new products mentioned in the questionnaire as likely to respond as those who have purchased?

1.9 Each year a local weekly newspaper gives out "Best of the City" awards in categories such as restaurant, deli, pastry shop, and so on. Readers are asked to fill in their favorites on a form enclosed in this free weekly paper and then send it to the publisher. The establishment receiving the most votes is declared the winner in its category. Identify the population and sample and comment on the representativeness of the sample.

1.10 Which of the following are anecdotal and which are based on sample?

(a) Out of 200 students questioned, 40 admitted they lied regularly.

(b) Bobbie says the produce at Market W is the freshest in the city.

(c) Out of 50 persons interviewed at a shopping mall, 18 had made a purchase that day.

1.11 Which of the following are anecdotal and which are based on a sample?

(a) Tom says he gets the best prices on electronics at the www.bestelc.com Internet site.

[3] J. Witt et al. "Putting to a bigger hole: Golf performance relates to perceived size," *Psychonomic Bulletin and Review* **15**(3) (2008), pp. 581–586.

[4] Harris Interactive telephone survey (October 16–19, 2008).

[5] http://psychologytoday.psychtests.com/tests/procrastination_access.html

(b) Out of 22 students, 6 had multiple credit cards.

(c) Among 55 people checking in at the airport, 12 were going to destinations outside of the continental United States.

1.12 What is wrong with this statement of purpose?

PURPOSE: *Determine if a newly designed rollerball pen is comfortable to hold when writing.*

Give an improved statement of purpose.

1.13 What is wrong with this statement of purpose?

PURPOSE: *Determine if it takes too long to get cash from the automated teller machine during the lunch hour.*

Give an improved statement of purpose.

1.14 Give a statement of purpose for determining the amount of time it takes to make hotel reservations in San Francisco using the Internet.

1.15 Thirty-five classrooms on campus are equiped for multimedia instruction. Use Table 1, Appendix B, to select 4 of these classrooms to visit and check whether or not the instructor is using the equipment during that day's first hour lecture.

1.16 Fifty band members would like to ride the band bus to an out-of-town game. However, there is room for only 44. Use Table 1, Appendix B, to select the 44 persons who will go. Determine how to make your selection by taking only a few two-digit selections.

1.17 Eight young students need mentors. Of these, there are three whom you enjoy being with while you are indifferent about the others. Two of the students will be randomly assigned to you. Label the students you like by 0, 1, and 2 and the others by 3, 4, 5, 6, and 7. Then, the process of assigning two students at random is equivalent to choosing two different digits from the table of random digits and ignoring any 8 or 9. Repeat the experiment of assigning two students 20 times by using the table of random digits. Record the pairs of digits you draw for each experiment.

(a) What is the proportion of the 20 experiments that give two students that you like?

(b) What is the proportion of the 20 experiments that give one of the students you like and one other?

(c) What is the proportion of the 20 experiments that give none of the students you like?

1.18 According to the cause-and-effect diagram on page 17, where are the possible delays on the first floor?

1.19 Refer to the cause-and-effect diagram on page 17. The workers have now noticed that a delay could occur:

(i) On the fourth floor at the pharmacy

(ii) On the third floor at the practitioners' station

Redraw the diagram and include this added information.

1.20 The United States Environmental Protection Agency[6] reports that in 2006, each American generated 4.6 pounds of solid waste a day.

(a) Does this mean every single American produces the same amount of garbage? What do you think this statement means?

(b) Was the number 4.6 obtained from a sample? Explain.

(c) How would you select a sample?

1.21 As a very extreme case of self-selection, imagine a five-foot-high solid wood fence surrounding a collection of Great Danes and Miniature Poodles. You want to estimate the proportion of Great Danes inside and decide to collect your sample by observing the first seven dogs to jump high enough to be seen above the fence.

(a) Explain how this is a self-selected sample that is, of course, very misleading.

(b) How is this sample selection procedure like a call-in election poll?

[6] http://www.epa.gov/epawaste/nonhaz/index.htm

2

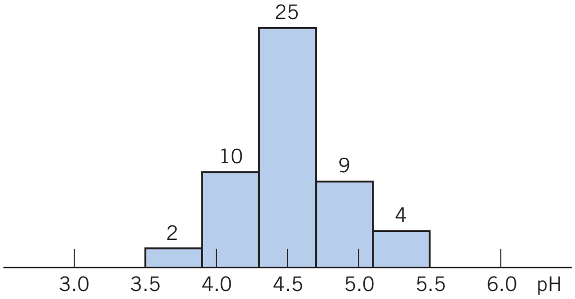
Organization and Description of Data

1. Introduction
2. Main Types of Data
3. Describing Data by Tables and Graphs
4. Measures of Center
5. Measures of Variation
6. Checking the Stability of the Observations over Time
7. More on Graphics
8. Statistics in Context
9. Review Exercises

Acid Rain Is Killing Our Lakes

© SuperStock, Inc.

Acid precipitation is linked to the disappearance of sport fish and other organisms from lakes. Sources of air pollution, including automobile emissions and the burning of fossil fuels, add to the natural acidity of precipitation. The Wisconsin Department of Natural Resources initiated a precipitation monitoring program with the goal of developing appropriate air pollution controls to reduce the problem. The acidity of the first 50 rains monitored, measured on a pH scale from 1 (very acidic) to 7 (basic), are summarized by the histogram.



Histogram of acid rain data

Notice that all the rains are more acidic than normal rain, which has a pH of 5.6. (As a comparison, apples are about pH 3 and milk is about pH 6.)

Researchers in Canada have established that lake water with a pH below 5.6 may severely affect the reproduction of game fish. More research will undoubtedly improve our understanding of the acid rain problem and lead, it is hoped, to an improved environment.

1. INTRODUCTION

In Chapter 1, we cited several examples of situations where the collection of data by appropriate processes of experimentation or observation is essential to acquire new knowledge. A data set may range in complexity from a few entries to hundreds or even thousands of them. Each entry corresponds to the observation of a specified characteristic of a sampling unit. For example, a nutritionist may provide an experimental diet to 30 undernourished children and record their weight gains after two months. Here, children are the sampling units, and the data set would consist of 30 measurements of weight gains. Once the data are collected, a primary step is to organize the information and extract a descriptive summary that highlights its salient features. In this chapter, we learn how to organize and describe a set of data by means of tables, graphs, and calculation of some numerical summary measures.

2. MAIN TYPES OF DATA

In discussing the methods for providing summary descriptions of data, it helps to distinguish between the two basic types:

1. **Qualitative** or **categorical** data
2. **Numerical** or **measurement** data

When the characteristic under study concerns a qualitative trait that is only classified in categories and not numerically measured, the resulting data are called categorical data. Hair color (blond, brown, red, black), employment status (employed, unemployed), and blood type (O, A, B, AB) are but some examples. If, on the other hand, the characteristic is measured on a numerical scale, the resulting data consist of a set of numbers and are called measurement data. We will use the term **numerical-valued variable** or just **variable** to refer to a characteristic that is measured on a numerical scale. The word "variable" signifies that the measurements vary over different sampling units. In this terminology, observations of a numerical-valued variable yield measurement data. A few examples of numerical-valued variables are the shoe size of an adult male, daily number of traffic fatalities in a state, intensity of an earthquake, height of a 1-year-old pine seedling, the time in line at an automated teller, and the number of offspring in an animal litter.

Although in all these examples the stated characteristic can be numerically measured, a close scrutiny reveals two distinct types of underlying scale of measurement. Shoe sizes are numbers such as 6, $6\frac{1}{2}$, 7, $7\frac{1}{2}$, . . . , which proceed in steps of $\frac{1}{2}$. The count of traffic fatalities can only be an integer and so is the number of offspring in an animal litter. These are examples of **discrete variables.** The name **discrete** draws from the fact that the scale is made up of distinct numbers with gaps in between. On the other hand, some variables such as height, weight, and survival time can ideally take any value in an interval. Since the measurement scale does not have gaps, such variables are called **continuous.**

We must admit that a truly continuous scale of measurement is an idealization. Measurements actually recorded in a data set are always rounded either for the sake of simplicity or because the measuring device has a limited accuracy. Still, even though weights may be recorded in the nearest pounds or time recorded in the whole hours, their actual values occur on a continuous scale so the data are referred to as continuous. Counts are inherently discrete and treated as such, provided that they take relatively few distinct values (e.g., the number of children in a family or the number of traffic violations of a driver). But when a count spans a wide range of values, it is often treated as a continuous variable. For example, the count of white blood cells, number of insects in a colony, and number of shares of stock traded per day are strictly discrete, but for practical purposes, they are viewed as continuous.

A summary description of categorical data is discussed in Section 3.1. The remainder of this chapter is devoted to a descriptive study of measurement data, both discrete and continuous. As in the case of summarization and commentary on a long, wordy document, it is difficult to prescribe concrete steps for summary descriptions that work well for all types of measurement data. However, a few important aspects that deserve special attention are outlined here to provide general guidelines for this process.

Describing a Data Set of Measurements

1. **Summarization and description of the overall pattern.**
   (a) Presentation of tables and graphs.
   (b) Noting important features of the graphed data including symmetry or departures from it.
   (c) Scanning the graphed data to detect any observations that seem to stick far out from the major mass of the data—the outliers.
2. **Computation of numerical measures.**
   (a) A typical or representative value that indicates the center of the data.
   (b) The amount of spread or variation present in the data.

3. DESCRIBING DATA BY TABLES AND GRAPHS

3.1 CATEGORICAL DATA

When a qualitative trait is observed for a sample of units, each observation is recorded as a member of one of several categories. Such data are readily organized in the form of a frequency table that shows the counts (**frequencies**) of the individual categories. Our understanding of the data is further enhanced by calculation of the proportion (also called **relative frequency**) of observations in each category.

$$\text{Relative frequency of a category} = \frac{\text{Frequency in the category}}{\text{Total number of observations}}$$

Example 1 Calculating Relative Frequencies to Summarize an Opinion Poll

A campus press polled a sample of 280 undergraduate students in order to study student attitude toward a proposed change in the dormitory regulations. Each student was to respond as support, oppose, or neutral in regard to the issue. The numbers were 152 support, 77 neutral, and 51 opposed. Tabulate the results and calculate the relative frequencies for the three response categories.

SOLUTION Table 1 records the frequencies in the second column, and the relative frequencies are calculated in the third column. The relative frequencies show that about 54% of the polled students supported the change, 18% opposed, and 28% were neutral.

TABLE 1 Summary Results of an Opinion Poll

Responses	Frequency	Relative Frequency
Support	152	$\frac{152}{280} = .543$
Neutral	77	$\frac{77}{280} = .275$
Oppose	51	$\frac{51}{280} = .182$
Total	280	1.000

Remark: The relative frequencies provide the most relevant information as to the pattern of the data. One should also state the sample size, which serves as an indicator of the credibility of the relative frequencies. (More on this in Chapter 8.)

Categorical data are often presented graphically as a **pie chart** in which the segments of a circle exhibit the relative frequencies of the categories. To obtain the angle for any category, we multiply the relative frequency by 360 degrees, which corresponds to the complete circle. Although laying out the angles by hand can be tedious, many software packages generate the chart with a single command. Figure 1 presents a pie chart for the data in Example 1.

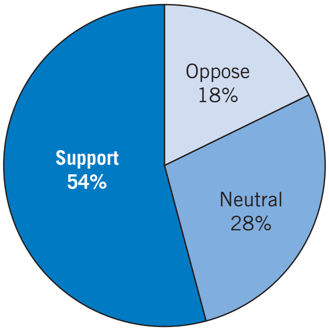


Figure 1 Pie chart of student opinion on change in dormitory regulations.

When questions arise that need answering but the decision makers lack precise knowledge of the state of nature or the full ramifications of their decisions, the best procedure is often to collect more data. In the context of quality improvement, if a problem is recognized, the first step is to collect data on the magnitude and possible causes. This information is most effectively communicated through graphical presentations.

A **Pareto diagram** is a powerful graphical technique for displaying events according to their frequency. According to Pareto's empirical law, any collection of events consists of only a few that are major in that they are the ones that occur most of the time.

Figure 2 gives a Pareto diagram for the type of defects found in a day's production of facial tissues. The cumulative frequency is 22 for the first cause and

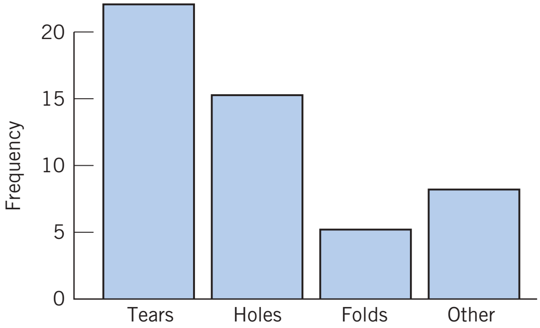


Figure 2 Pareto diagram of facial tissue defects.

22 + 15 = 37 for the first and second causes combined. This illustrates Pareto's rule, with two of the causes being responsible for 37 out of 50, or 74%, of the defects.

Example 2 A Pareto Diagram Clarifies Circumstances Needing Improvement

Graduate students in a counseling course were asked to choose one of their personal habits that needed improvement. In order to reduce the effect of this habit, they were asked to first gather data on the frequency of the occurrence and the circumstances. One student collected the following frequency data on fingernail biting over a two-week period.

Frequency	Activity
58	Watching television
21	Reading newspaper
14	Talking on phone
7	Driving a car
3	Grocery shopping
12	Other

Make a Pareto diagram showing the relationship between nail biting and type of activity.

SOLUTION The cumulative frequencies are 58, 58 + 21 = 79, and so on, out of 115. The Pareto diagram is shown in Figure 3, where watching TV accounts for 50.4% of the instances.

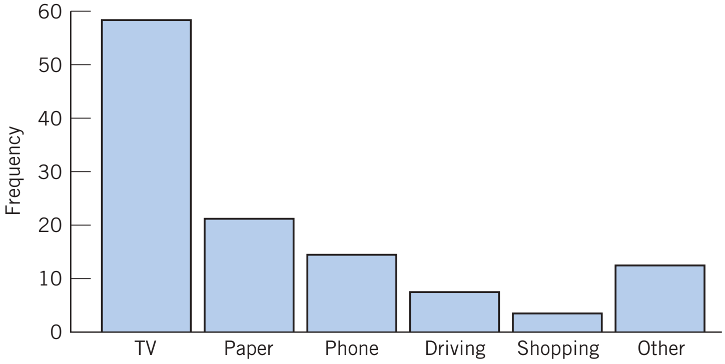


Figure 3 Pareto diagram for nail biting example.

The next step for this person would be to try and find a substitute for nail biting while watching television.

3.2 DISCRETE DATA

We next consider summary descriptions of measurement data and begin our discussion with discrete measurement scales. As explained in Section 2, a data set is identified as discrete when the underlying scale is discrete and the distinct values observed are not too numerous.

Similar to our description of categorical data, the information in a **discrete data set** can be summarized in a frequency table, or **frequency distribution** that includes a calculation of the **relative frequencies.** In place of the qualitative categories, we now list the distinct numerical measurements that appear in the data set and then count their frequencies.

Example 3 Creating a Frequency Distribution

Retail stores experience their heaviest returns on December 26 and December 27 each year. Most are gifts that, for some reason, did not please the recipient. The number of items returned, by a sample of 30 persons at a large discount department store, are observed and the data of Table 2 are obtained. Determine the frequency distribution.

TABLE 2 Number of items returned

1	4	3	2	3	4	5	1	2	1
2	5	1	4	2	1	3	2	4	1
2	3	2	3	2	1	4	3	2	5

SOLUTION The **frequency distribution** of these data is presented in Table 3. The values are paired with the frequency and the calculated relative frequency.

TABLE 3 Frequency Distribution for Number (x) of Items Returned

Value x	Frequency	Relative Frequency
1	7	.233
2	9	.300
3	6	.200
4	5	.167
5	3	.100
Total	30	1.000

The frequency distribution of a discrete variable can be presented pictorially by drawing either lines or rectangles to represent the relative frequencies. First, the distinct values of the variable are located on the horizontal axis. For a **line diagram,** we draw a vertical line at each value and make the height of the line equal to the relative frequency. A **histogram** employs vertical rectangles instead of lines. These rectangles are centered at the values and their areas represent relative frequencies. Typically, the values proceed in equal steps so the rectangles are all of the same width and their heights are proportional to the relative frequencies as well as frequencies. Figure 4(*a*) shows the line diagram and 4(*b*) the histogram of the frequency distribution of Table 3.

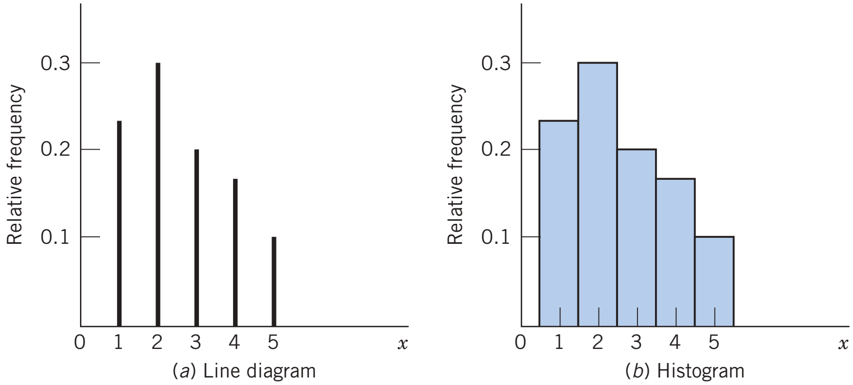


Figure 4 Graphic display of the frequency distribution of data in Table 3.

3.3 DATA ON A CONTINUOUS VARIABLE

We now consider tabular and graphical presentations of data sets that contain numerical measurements on a virtually continuous scale. Of course, the recorded measurements are always rounded. In contrast with the discrete case, a data set of measurements on a continuous variable may contain many distinct values. Then, a table or plot of all distinct values and their frequencies will not provide a condensed or informative summary of the data.

The two main graphical methods used to display a data set of measurements are the **dot diagram** and the **histogram.** Dot diagrams are employed when there are relatively few observations (say, less than 20 or 25); histograms are used with a larger number of observations.

Dot Diagram

When the data consist of a small set of numbers, they can be graphically represented by drawing a line with a scale covering the range of values of the measurements. Individual measurements are plotted above this line as prominent dots. The resulting diagram is called a **dot diagram.**

Example 4 A Dot Diagram Reveals an Unusual Observation

The number of days the first six heart transplant patients at Stanford survived after their operations were 15, 3, 46, 623, 126, 64. Make a dot diagram.

SOLUTION These survival times extended from 3 to 623 days. Drawing a line segment from 0 to 700, we can plot the data as shown in Figure 5. This dot diagram shows a cluster of small survival times and a single, rather large value.

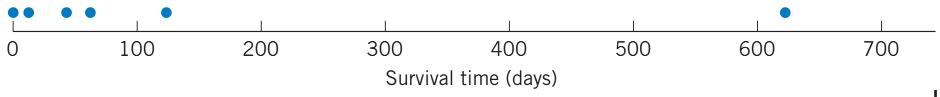


Figure 5 Dot diagram for the heart transplant data.

Frequency Distribution on Intervals

When the data consist of a large number of measurements, a dot diagram may be quite tedious to construct. More seriously, overcrowding of the dots will cause them to smear and mar the clarity of the diagram. In such cases, it is convenient to condense the data by grouping the observations according to intervals and recording the frequencies of the intervals. Unlike a discrete frequency distribution, where grouping naturally takes place on points, here we use intervals of values. The main steps in this process are outlined as follows.

Constructing a Frequency Distribution for a Continuous Variable

1. Find the minimum and the maximum values in the data set.
2. Choose intervals or cells of equal length that cover the range between the minimum and the maximum without overlapping. These are called class intervals, and their endpoints class boundaries.
3. Count the number of observations in the data that belong to each class interval. The count in each class is the **class frequency** or **cell frequency.**
4. Calculate the relative frequency of each class by dividing the class frequency by the total number of observations in the data:

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

The choice of the number and position of the class intervals is primarily a matter of judgment guided by the following considerations. The number of

Paying Attention

© Britt Erlanson/The Image Bank/Getty Images

Paying attention in class. Observations on 24 first-grade students.

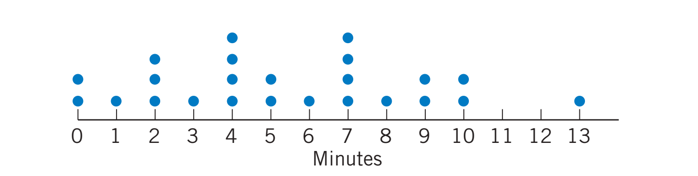


Figure 6 Time not concentrating on the mathematics assignment (out of 20 minutes).

First-grade teachers allot a portion of each day to mathematics. An educator, concerned about how students utilize this time, selected 24 students and observed them for a total of 20 minutes spread over several days. The number of minutes, out of 20, that the student was not on task was recorded (courtesy of T. Romberg). These lack-of-attention times are graphically portrayed in the dot diagram in Figure 6. The student with 13 out of 20 minutes off-task stands out enough to merit further consideration. Is this a student who finds the subject too difficult or might it be a very bright child who is bored?

classes usually ranges from 5 to 15, depending on the number of observations in the data. Grouping the observations sacrifices information concerning how the observations are distributed within each cell. With too few cells, the loss of information is serious. On the other hand, if one chooses too many cells and the data set is relatively small, the frequencies from one cell to the next would jump up and down in a chaotic manner and no overall pattern would emerge. As an initial step, frequencies may be determined with a large number of intervals that can later be combined as desired in order to obtain a smooth pattern of the distribution.

Computers conveniently order data from smallest to largest so that the observations in any cell can easily be counted. The construction of a frequency distribution is illustrated in Example 5.

Example 5 Creating a Frequency Distribution for Hours of Sleep

Students require different amounts of sleep. A sample of 59 students at a large midwest university reported the following hours of sleep the previous night.

TABLE 4 Hours of Sleep for Fifty-nine Students

4.5	4.7	5.0	5.0	5.3	5.5	5.5	5.7	5.7	5.7
6.0	6.0	6.0	6.0	6.3	6.3	6.3	6.5	6.5	6.5
6.7	6.7	6.7	6.7	7.0	7.0	7.0	7.0	7.3	7.3
7.3	7.3	7.5	7.5	7.5	7.5	7.7	7.7	7.7	7.7
8.0	8.0	8.0	8.0	8.3	8.3	8.3	8.5	8.5	8.5
8.5	8.7	8.7	9.0	9.0	9.0	9.3	9.3	10.0	

Construct a frequency distribution of the sleep data.

SOLUTION To construct a frequency distribution, we first notice that the minimum hours of sleep is 4.5 and the maximum is 10.0. We choose class intervals of length 1.2 hours as a matter of convenience.

The selection of class boundaries is a bit of fussy work. Because the data have one decimal place, we could add a second decimal to avoid the possibility of any observation falling exactly on the boundary. For example, we could end the first class interval at 5.45. Alternatively, and more neatly, we could write 4.3–5.5 and make the **endpoint convention** that the left-hand end point is included but not the right.

The first interval contains 5 observations so its frequency is 5 and its relative frequency is $\frac{5}{59} = .085$. Table 5 gives the frequency distribution. The relative frequencies add to 1, as they should (up to rounding error) for any frequency distribution. We see, for instance, that just about one-third of the students $.271 + .051 = .322$ got 7.9 hours or more of sleep.

Remark: The rule requiring equal class intervals is inconvenient when the data are spread over a wide range but are highly concentrated in a small part of the range with relatively few numbers elsewhere. Using smaller intervals where the data are highly concentrated and larger intervals where the data are sparse helps to reduce the loss of information due to grouping.

TABLE 5 Frequency Distribution for Hours of Sleep Data (left endpoints included but right endpoints excluded)

Class Interval	Frequency	Relative Frequency
4.3–5.5	5	$\frac{5}{59} = .085$
5.5–6.7	15	$\frac{15}{59} = .254$
6.7–7.9	20	$\frac{20}{59} = .339$
7.9–9.1	16	$\frac{16}{59} = .271$
9.1–10.3	3	$\frac{3}{59} = .051$
Total	59	1.000

In every application involving an **endpoint convention,** it is important that you clearly state which endpoint is included and which is excluded. This information should be presented in the title or in a footnote of any frequency distribution.

Histogram

A frequency distribution can be graphically presented as a histogram. To draw a histogram, we first mark the class intervals on the horizontal axis. On each interval, we then draw a vertical rectangle whose **area represents the relative frequency**—that is, the proportion of the observations occurring in that class interval.

To create rectangles whose area is equal to relative frequency, use the rule

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of interval}}$$

The total area of all rectangles equals 1, the sum of the relative frequencies.

The total area of a histogram is 1.

The histogram for Table 5 is shown in Figure 7. For example, the rectangle drawn on the class interval 4.3–5.5 has area $= .071 \times 1.2 = .085$, which is the relative frequency of this class. Actually, we determined the height .071 as

$$\text{Height} = \frac{\text{Relative frequency}}{\text{Width of interval}} = \frac{.085}{1.2} = .071$$

The units on the vertical axis can be viewed as relative frequencies per unit of the horizontal scale. For instance, .071 is the relative frequency per hour for the interval 4.3–5.5.

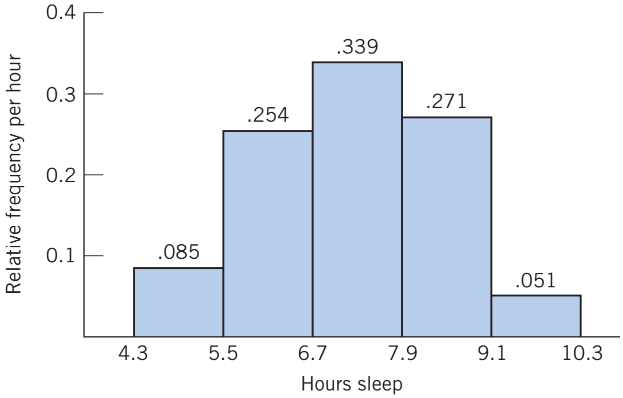


Figure 7 Histogram of the sleep data of Tables 4 and 5. Sample size = 59.

Visually, we note that the rectangle having largest area, or most frequent class interval, is 6.7–7.9. Also, proportion .085 + .254 = .339 of the students slept less than 6.7 hours.

> ***Remark:*** When all class intervals have equal widths, the heights of the rectangles are proportional to the relative frequencies that the areas represent. The formal calculation of height, as area divided by the width, is then redundant. Instead, one can mark the vertical scale according to the relative frequencies—that is, make the heights of the rectangles equal to the relative frequencies. The resulting picture also makes the areas represent the relative frequencies if we read the vertical scale as if it is in units of the class interval. This leeway when plotting the histogram is not permitted in the case of unequal class intervals.

Figure 8 shows one ingenious way of displaying two histograms for comparison. In spite of their complicated shapes, their back-to-back plot as a “tree” allows for easy visual comparison. Females are the clear majority in the last age groups of the male and female age distributions.

Stem-and-Leaf Display

A **stem-and-leaf display** provides a more efficient variant of the histogram for displaying data, especially when the observations are two-digit numbers. This plot is obtained by sorting the observations into rows according to their leading digit. The stem-and-leaf display for the data of Table 6 is shown in Table 7. To make this display:

1. List the digits 0 through 9 in a column and draw a vertical line. These correspond to the leading digit.
2. For each observation, record its second digit to the right of this vertical line in the row where the first digit appears.
3. Finally, arrange the second digits in each row so they are in increasing order.

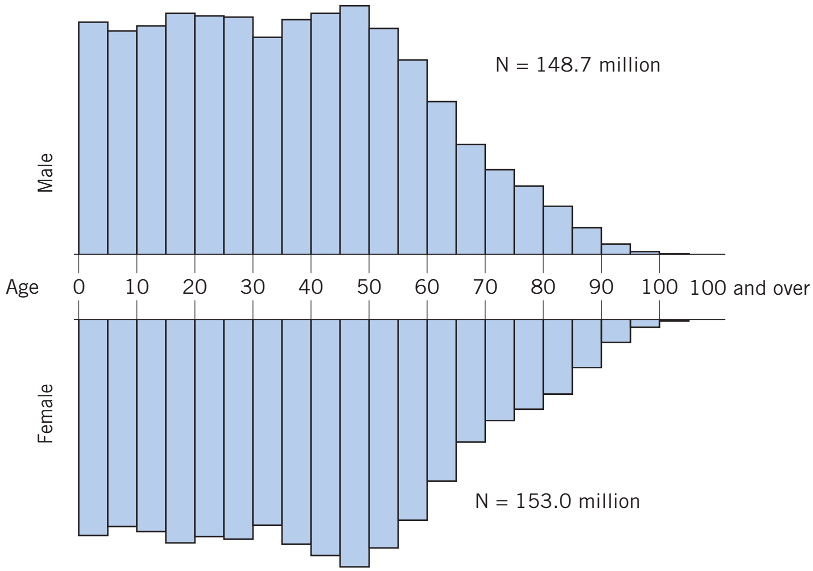


Figure 8 Population tree (histograms) of the male and female age distributions in the United States in 2007. (*Source:* U.S. Bureau of the Census.)

TABLE 6 Examination Scores of 50 Students

75	98	42	75	84	87	65	59	63
86	78	37	99	66	90	79	80	89
68	57	95	55	79	88	76	60	77
49	92	83	71	78	53	81	77	58
93	85	70	62	80	74	69	90	62
84	64	73	48	72				

TABLE 7 Stem-and-Leaf Display for the Examination Scores

0	
1	
2	
3	7
4	289
5	35789
6	022345689
7	01234556778899
8	00134456789
9	0023589

In the stem-and-leaf display, the column of first digits to the left of the vertical line is viewed as the stem, and the second digits as the leaves. Viewed sidewise, it looks like a histogram with a cell width equal to 10. However, it is more informative than a histogram because the actual data points are retained. In fact, every observation can be recovered exactly from this stem-and-leaf display.

A stem-and-leaf display retains all the information in the leading digits of the data. When the leaf unit = .01, 3.5 | 0 2 3 7 8 presents the data 3.50, 3.52, 3.53, 3.57, and 3.58. Leaves may also be two-digit at times. When the first leaf digit = .01, .4 | 07 13 82 90 presents the data 0.407, 0.413, 0.482, and 0.490.

Further variants of the stem-and-leaf display are described in Exercises 2.25 and 2.26. This versatile display is one of the most applicable techniques of exploratory data analysis.

When the sample size is small or moderate, no information is lost with the stem-and-leaf diagram because you can see every data point. The major disadvantage is that, when the sample size is large, diagrams with hundreds of numbers in a row cannot be constructed in a legible manner.

Exercises

2.1 Cities must find better ways to dispose of solid waste. According to the Environmental Protection Agency, the composition of the 254 million tons of solid municipal waste created in 2007 was

Paper and paperboard	32.7%
Yard waste	12.8%
Food waste	12.5%
Plastics	12.1%
Metals	8.2%
Other materials	

(a) Determine the percentage of other materials in the solid waste. This category includes glass, wood, rubber, and so on.

(b) Create a Pareto chart.

(c) What percentage of the total solid waste is paper or paperboard? What percentage is from the top two categories? What percentage is from the top five categories?

2.2 Recorded here are the blood types of 40 persons who have volunteered to donate blood at a plasma center. Summarize the data in a frequency table. Include calculations of the relative frequencies.

O O A B A O A A A O
B O B O O A O O A A
A A AB A B A A O O A
O O A A A O A O O AB

2.3 A student at the University of Wisconsin surveyed 40 students in her dorm concerning their participation in extracurricular activities during the past week. The data on number of activities are

1 5 0 1 4 3 0 2 1 6 1 1 0 0
2 0 0 3 1 2 1 2 2 2 2 2 1 0
2 2 3 4 2 7 2 2 3 3 1 1

Present these data in a frequency table and in a relative frequency bar chart.

2.4 The number of automobile accidents reported per month helps to identify intersections that require improvement. Beginning January 2004 and ending November 2008, the number of crashes per month reported at an intersection near a university campus in Madison, Wisconsin, are

1 3 3 3 2 2 3 1 2 4 1 4
1 3 1 1 1 0 1 2 2 5 5 2
5 5 4 3 3 6 1 2 3 2 4 3
4 4 3 5 3 3 3 5 1 5 5 3
4 2 2 0 0 1 4 1 0 2 0

Present these data in a frequency table and in a relative frequency bar chart.

2.5 The following table shows how workers in one department get to work.

Mode of Transportation	Frequency
Drive alone	25
Car pool	3
Ride bus	7
Other	5

(a) Calculate the relative frequency of each mode of transportation.

(b) Construct a pie chart.

2.6 Of the $207 million raised by a major university's fund drive, $117 million came from individuals and bequests, $24 million from industry and business, and $66 million from foundations and associations. Present this information in the form of a pie chart.

2.7 Data from one campus dorm on the number of burglaries are collected each week of the semester. These data are to be grouped into the classes 0–1, 2–3, 3–5, 6 or more. Both endpoints included. Explain where a difficulty might arise.

2.8 Data from one campus dorm, on the number of complaints about the dorm food are collected each week of the semester. These weekly counts are to be grouped into the classes 0–1, 2–3, 4–5, 7 or more. Both endpoints are included. Explain where a difficulty might arise.

2.9 A sample of persons will each be asked to give the number of their close friends. The responses are to be grouped into the following classes: 0, 1–3, 3–5, 6 or more. Left endpoint is included. Explain where difficulties might arise.

2.10 The weights of the players on the university football team (to the nearest pound) are to be grouped into the following classes: 160–175, 175–190, 190–205, 205–220, 220–235, 235 or more. The left endpoint is included but not the right endpoint. Explain where difficulties might arise.

2.11 On flights from San Francisco to Chicago, the number of empty seats are to be grouped into the following classes: 0–4, 5–9, 10–14, 15–19, more than 19.

Is it possible to determine from this frequency distribution the exact number of flights on which there were:

(a) Fewer than 10 empty seats?

(b) More than 14 empty seats?

(c) At least 5 empty seats?

(d) Exactly 9 empty seats?

(e) Between 5 and 15 empty seats inclusively?

2.12 A major West Coast power company surveyed 50 customers who were asked to respond to the statement, "People should rely mainly on themselves to solve problems caused by power outages" with one of the following responses.

1. Definitely agree.
2. Somewhat agree.
3. Somewhat disagree.
4. Definitely disagree.

The responses are as follows:

```
4  2  1  3  3  2  4  2  1  1  2  2  2  2  1  3  4
1  4  4  1  3  2  4  1  4  3  3  1  1  1  2  1  1
4  4  4  4  4  1  2  2  2  4  4  4  1  3  4  2
```

Construct a frequency table.

2.13 A sample of 50 departing airline passengers at the main check-in counter produced the following number of bags checked through to final destinations.

```
0  1  2  2  1  2  1  2  3  0  1  0
1  1  0  1  3  0  1  2  1  1  1  2
1  2  2  1  2  0  0  2  2  1  1  1
1  1  1  1  2  0  1  3  0  1  2  1
1  3
```

(a) Make a relative frequency line diagram.

(b) Comment on the pattern.

(c) What proportion of passengers who check in at the main counter fail to check any bags?

2.14 A person with asthma took measurements by blowing into a peak-flow meter on seven consecutive days.

429 425 471 422 432 444 454

Display the data in a dot diagram.

2.15 Before microwave ovens are sold, the manufacturer must check to ensure that the radiation coming through the door is below a specified safe limit. The amounts of radiation leakage (mW/cm^2) with the door closed from 25 ovens are as follows (courtesy of John Cryer):

15	9	18	10	5	12	8
5	8	10	7	2	1	
5	3	5	15	10	15	
9	8	18	1	2	11	

Display the data in a dot diagram.

2.16 A campus area merchant recorded the number of bad checks received per month, for five months

4 5 4 7 6

Display the data in a dot diagram.

2.17 The city of Madison regularly checks the water quality at swimming beaches located on area lakes. The concentration of fecal coliforms, in number of colony forming units (CFU) per 100 ml of water, was measured on fifteen days during the summer at one beach.

180	1600	90	140	50	260	400	90
380	110	10	60	20	340	80	

(a) Make a dot diagram.

(b) Comment on the pattern and any unusual features.

(c) The city closes any swimming beach if a count is over 1350. What proportion of days, among the fifteen, was this beach closed?

2.18 Tornadoes kill many people every year in the United States. The yearly number of lives lost during the 59 years 1950 through 2008 are summarized in the following table.

Number of Deaths	Frequency
24 or less	2
25–49	20
50–74	18
75–99	7
100–149	6
150–199	2
200–249	1
250 or more	3
Total	59

(a) Calculate the relative frequency for the intervals [0, 25), [25, 50) and so on where the right-hand endpoint is excluded. Take the last interval to be [250, 550).

(b) Plot the relative frequency histogram. (*Hint:* Since the intervals have unequal widths, make the height of each rectangle equal to the relative frequency divided by the width of the interval.)

(c) What proportion of the years had 49 or fewer deaths due to tornadoes?

(d) Comment on the shape of the distribution.

2.19 A zoologist collected wild lizards in the Southwestern United States. Thirty lizards from the genus *Phrynosoma* were placed on a treadmill and their speed measured. The recorded speed (meters/second) is the fastest time to run a half meter. (Courtesy of K. Bonine.)

1.28	1.36	1.24	2.47	1.94	2.52	2.67	1.29
1.56	2.66	2.17	1.57	2.10	2.54	1.63	2.11
2.57	1.72	0.76	1.02	1.78	0.50	1.49	1.57
1.04	1.92	1.55	1.78	1.70	1.20		

(a) Construct a frequency distribution using the class intervals 0.45–0.90, 0.90–1.35, and so on, with the endpoint convention that the left endpoint is included and the right endpoint is excluded. Calculate the relative frequencies.

(b) Make a histogram.

2.20 The United States Geological Survey maintains data on large earthquakes including those of magnitude greater than 6.0 in California. Through 2008, the ordered magnitudes of the 55 quakes are

6.1	6.1	6.1	6.1	6.1	6.2	6.2	6.2	6.2	6.3	6.3
6.3	6.4	6.4	6.4	6.4	6.4	6.4	6.4	6.5	6.5	6.5
6.5	6.5	6.6	6.6	6.6	6.7	6.7	6.7	6.8	6.8	6.8
6.8	6.8	6.9	6.9	7.0	7.0	7.0	7.1	7.1	7.1	7.2
7.2	7.2	7.2	7.3	7.3	7.3	7.3	7.4	7.8	7.8	7.9

Construct a histogram using equal-length intervals starting with (6.0, 6.3] where the right-hand endpoint is included but not the left-hand endpoint.

2.21 Referring to Exercise 2.20, construct a density histogram using the intervals (6.0, 6.3], (6.3, 6.6], (6.6, 6.9], (6.9, 7.2], and (7.2, 7.9].

2.22 The following data represent the scores of 40 students on a college qualification test (courtesy of R. W. Johnson).

162	171	138	145	144	126	145	162	174	178
167	98	161	152	182	136	165	137	133	143
184	166	115	115	95	190	119	144	176	135
194	147	160	158	178	162	131	106	157	154

Make a stem-and-leaf display.

2.23 A federal government study of the oil reserves in Elk Hills, CA, included a study of the amount of iron present in the oil.

Amount of Iron (percent ash)				
20	18	25	26	17
14	20	14	18	15
22	15	17	25	22
12	52	27	24	41
34	20	17	20	19
20	16	20	15	34
22	29	29	34	27
13	6	24	47	32
12	17	36	35	41
36	32	46	30	51

Make a stem-and-leaf display.

2.24 The following is a stem-and-leaf display with two-digit leaves. (The leading leaf digit = 10.0.)

1					
2	46	68	93		
3	19	44	71	82	97
4	05	26	43	90	
5	04	68			
6	13				

List the corresponding measurements.

2.25 If there are too many leaves on some stems in a stem-and-leaf display, we might double the number of stems. The leaves 0–4 could hang on one stem and 5–9 on the repeated stem. For the observations

193 198 200 202 203 203 205 205 206 207
207 208 212 213 214 217 219 220 222 226 237

we would get the **double-stem display**

19	3
19	8
20	0233
20	556778
21	234
21	79
22	02
22	6
23	
23	7

Construct a double-stem display with one-digit leaves for the data of Exercise 2.22.

2.26 If the double-stem display still has too few stems, we may wish to construct a stem-and-leaf display with a separate stem to hold leaves 0 and 1, 2 and 3, 4 and 5, 6 and 7, and a stem to hold 8 and 9. The resulting stem-and-leaf display is called a **five-stem display.** The following is a five-digit stem-and-leaf display. (Leaf unit = 1.0)

1	8
2	001
2	2233
2	444555
2	667
2	9
3	0

List the corresponding measurements.

2.27 The following table lists values of the Consumer Price Index for 24 selected areas both for 2007 and 2001. Construct a five-stem display for the consumer price index in 2007.

	2007	2001
Anchorage	181	155
Atlanta	198	176
Boston	227	191
Chicago	198	178
Cincinnati	188	168
Cleveland	186	173
Dallas	195	170
Denver	194	181
Detroit	195	174
Honolulu	219	178
Houston	182	159
Kansas City	186	172
Los Angeles	210	177
Miami	210	173
Milwaukee	198	172
Minneapolis	195	177
New York	221	187
Philadelphia	216	181
Pittsburgh	194	173
Portland	203	182
St. Louis	192	167
San Diego	218	191
San Francisco	211	190
Seattle	210	186

4. MEASURES OF CENTER

The graphic procedures described in Section 3 help us to visualize the pattern of a data set of measurements. To obtain a more objective summary description and a comparison of data sets, we must go one step further and obtain numerical values for the location or center of the data and the amount of variability present. Because data are normally obtained by sampling from a large population, our discussion of numerical measures is restricted to data arising in this context. Moreover, when the population is finite and completely sampled, the same arithmetic operations can be carried out to obtain numerical measures for the population.

To effectively present the ideas and associated calculations, it is convenient to represent a data set by symbols to prevent the discussion from becoming anchored to a specific set of numbers. A data set consists of a number of measurements which are symbolically represented by $x_1, x_2, \ldots, x_n$. The last subscript n denotes the number of measurements in the data, and $x_1, x_2, \ldots$ represent the first observation, the second observation, and so on. For instance, a data set consisting of the five measurements 2.1, 3.2, 4.1, 5.6, and 3.7 is represented in symbols by x_1, x_2, x_3, x_4, x_5, where $x_1 = 2.1$, $x_2 = 3.2$, $x_3 = 4.1$, $x_4 = 5.6$, and $x_5 = 3.7$.

The most important aspect of studying the distribution of a sample of measurements is locating the position of a central value about which the measurements are distributed. The two most commonly used indicators of center are the **mean** and the **median.**

The **mean,** or **average,** of a set of measurements is the sum of the measurements divided by their number. For instance, the mean of the five measurements 2.1, 3.2, 4.1, 5.6, and 3.7 is

$$\frac{2.1 + 3.2 + 4.1 + 5.6 + 3.7}{5} = \frac{18.7}{5} = 3.74$$

To state this idea in general terms, we use symbols. If a sample consists of n measurements $x_1, x_2, \ldots, x_n$, the mean of the sample is

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\text{sum of the } n \text{ measurements}}{n}$$

The notation $\bar{x}$ will be used to represent a sample mean. To further simplify the writing of a sum, the Greek capital letter $\sum$ (sigma) is used as a statistical shorthand. With this symbol:

The sum $x_1 + x_2 + \cdots + x_n$ is denoted as $\sum_{i=1}^{n} x_i$.

Read this as "the sum of all x_i with i ranging from 1 to n."

For example, $\sum_{i=1}^{5} x_i$ represents the sum $x_1 + x_2 + x_3 + x_4 + x_5$.

Remark: When the number of terms being summed is understood from the context, we often simplify to $\sum x_i$, instead of $\sum_{i=1}^{n} x_i$. Some further operations with the $\sum$ notation are discussed in Appendix A1.

We are now ready to formally define the sample mean.

The **sample mean** of a set of n measurements $x_1, x_2, \ldots, x_n$ is the sum of these measurements divided by n. The sample mean is denoted by $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{or} \qquad \frac{\sum x_i}{n}$$

According to the concept of "average," the mean represents a center of a data set. If we picture the dot diagram of a data set as a thin weightless horizontal bar on which balls of equal size and weight are placed at the positions of the data points, then the mean $\overline{x}$ represents the point on which the bar will balance. The computation of the sample mean and its physical interpretation are illustrated in Example 6.

Example 6 Calculating and Interpreting the Sample Mean

The birth weights in pounds of five babies born in a hospital on a certain day are 9.2, 6.4, 10.5, 8.1, and 7.8. Obtain the sample mean and create a dot diagram.

SOLUTION The mean birth weight for these data is

$$\overline{x} = \frac{9.2 + 6.4 + 10.5 + 8.1 + 7.8}{5} = \frac{42.0}{5} = 8.4 \text{ pounds}$$

The dot diagram of the data appears in Figure 9, where the sample mean (marked by Δ) is the balancing point or center of the picture.

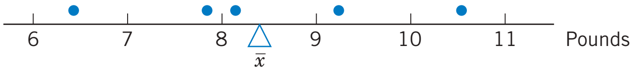


Figure 9 Dot diagram and the sample mean for the birth-weight data.

Another measure of center is the middle value.

The sample median of a set of n measurements $x_1, \ldots, x_n$ is the middle value when the measurements are arranged from smallest to largest.

Roughly speaking, the median is the value that divides the data into two equal halves. In other words, 50% of the data lie below the median and 50% above it. If n is an odd number, there is a unique middle value and it is the median. If n is an even number, there are two middle values and the median is defined as their average. For instance, the ordered data 3, 5, 7, 8 have two middle values 5 and 7, so the median $= (5 + 7)/2 = 6$.

Example 7 Calculating the Sample Median

Find the median of the birth-weight data given in Example 6.

SOLUTION The measurements, ordered from smallest to largest, are

$$6.4 \quad 7.8 \quad \boxed{8.1} \quad 9.2 \quad 10.5$$

The middle value is 8.1, and the median is therefore 8.1 pounds.

Example 8 Choosing between the Mean and Median

Calculate the median of the survival times given in Example 4. Also calculate the mean and compare.

SOLUTION To find the median, first we order the data. The ordered values are

$$3 \quad 15 \quad 46 \quad 64 \quad 126 \quad 623$$

There are two middle values, so

$$\text{Median} = \frac{46 + 64}{2} = 55 \text{ days}$$

The sample mean is

$$\bar{x} = \frac{3 + 15 + 46 + 64 + 126 + 623}{6} = \frac{877}{6} = 146.2 \text{ days}$$

Note that one large survival time greatly inflates the mean. Only 1 out of the 6 patients survived longer than $\bar{x} = 146.2$ days. Here the median of 55 days appears to be a better indicator of the center than the mean.

Example 8 demonstrates that the median is not affected by a few very small or very large observations, whereas the presence of such extremes can have a considerable effect on the mean. For extremely asymmetrical distributions, the median is likely to be a more sensible measure of center than the mean. That is why government reports on income distribution quote the median income as a summary, rather than the mean. A relatively small number of very highly paid persons can have a great effect on the mean salary.

If the number of observations is quite large (greater than, say, 25 or 30), it is sometimes useful to extend the notion of the median and divide the **ordered data** set into quarters. Just as the point for division into halves is called the median, the points for division into quarters are called **quartiles.** The points of division into more general fractions are called **percentiles.**

The sample 100 *p*-th percentile is a value such that after the data are ordered from smallest to largest, at least $100\,p\,\%$ of the observations are at or below this value and at least $100\,(1 - p)\,\%$ are at or above this value.

If we take $p = .5$, the above conceptual description of the sample $100(.5) = 50$th percentile specifies that at least half the observations are equal or smaller and at least half are equal or larger. If we take $p = .25$, the sample $100(.25) =$ 25th percentile has proportion one-fourth of the observations that are the same or smaller and proportion three-fourths that are the same or larger.

We adopt the convention of taking an observed value for the sample percentile except when two adjacent values satisfy the definition, in which case their average is taken as the percentile. This coincides with the way the median is defined when the sample size is even. When all values in an interval satisfy the definition of a percentile, the particular convention used to locate a point in the interval does not appreciably alter the results in large data sets, except perhaps for the determination of extreme percentiles (those before the 5th or after the 95th percentile).

The following operating rule will simplify the calculation of the sample percentile.

Calculating the Sample 100*p*-th Percentile

1. Order the data from smallest to largest.
2. Determine the product (*sample size*) × (*proportion*) = np.

If np is not an integer, round it up to the next integer and find the corresponding ordered value.

If np is an integer, say k, calculate the average of the kth and $(k + 1)$st ordered values.

The quartiles are simply the 25th, 50th, and 75th percentiles.

Sample Quartiles

Lower (first) quartile	Q_1	= 25th percentile
Second quartile (or median)	Q_2	= 50th percentile
Upper (third) quartile	Q_3	= 75th percentile

Example 9 Calculating Quartiles to Summarize Length of Phone Calls

An administrator wanted to study the utilization of long-distance telephone service by a department. One variable of interest is the length, in minutes, of long-distance calls made during one month. There were 38 calls that resulted in a connection. The lengths of calls, already ordered from smallest to largest, are presented in Table 8. Locate the quartiles and also determine the 90th percentile.

Table 8 The Lengths of Long-Distance Phone Calls in Minutes

1.6	1.7	1.8	1.8	1.9	2.1	2.5	3.0	3.0	4.4
4.5	4.5	5.9	7.1	7.4	7.5	7.7	8.6	9.3	9.5
12.7	15.3	15.5	15.9	15.9	16.1	16.5	17.3	17.5	19.0
19.4	22.5	23.5	24.0	31.7	32.8	43.5	53.3		

SOLUTION To determine the first quartile, we take $p = .25$ and calculate the product $38 \times .25 = 9.5$. Because 9.5 is not an integer, we take the next largest integer, 10. In Table 8, we see that the 10th ordered observation is 4.4 so the first quartile is $Q_1 = 4.4$ minutes.

We confirm that this observation has 10 values *at or below* it and 29 values *at or above* so that it does satisfy the conceptual definition of the first quartile.

For the median, we take $p = .5$ and calculate $38 \times .5 = 19$. Because this is an integer, we average the 19th and 20th smallest observations to obtain the median, $(9.3 + 9.5)/2 = 9.4$ minutes.

Next, to determine the third quartile, we take $p = .75$ and calculate $38 \times .75 = 28.5$. The next largest integer is 29, so the 29th ordered observation is the third quartile $Q_3 = 17.5$ minutes. More simply, we could mimic the calculation of the first quartile but now count down 10 observations starting with the largest value.

For the 90th percentile, we determine $38 \times .90 = 34.2$, which we increase to 35. The 90th percentile is 31.7 minutes. Only 10% of calls last 31.7 minutes or longer.

Exercises

2.28 Calculate the mean and median for each of the following data sets.

(a) 3 7 4 11 5

(b) 3 1 7 3 1

2.29 Calculate the mean and median for each of the following data sets.

(a) 2 5 1 4 3

(b) 26 30 38 32 26 31

(c) −1 2 0 1 4 −1 2

2.30 The height that bread rises may be one indicator of how light it will be. As a first step, before modifying her existing recipe, a student cook measured the raise height (cm) on eight occasions:

6.3 6.9 5.7 5.4 5.6 5.5 6.6 6.5

Find the mean and median of the raised heights.

2.31 With reference to the water quality in Exercise 2.17:

(a) Find the sample mean.

(b) Does the sample mean or the median give a better indication of the water quality of a "typical" day? Why?

2.32 The monthly income in dollars for seven sales persons at a car dealership are

2450 2275 2425 4700 2650 2350 2475

(a) Calculate the mean and median salary.

(b) Which of the two is preferable as a measure of center and why?

2.33 Records show that in Las Vegas, NV, the normal daily maximum temperature (°F) for each month starting in January is

56 62 68 77 87 99 105 102 95 82 66 57

Verify that the mean of these figures is 79.67. Comment on the claim that the daily maximum temperature in Las Vegas averages a pleasant 79.67.

2.34 A major wine producer reported sales (in hundreds of cases) for two-week periods during one summer:

85 82 77 83 80 77 94

Obtain the sample mean and median.

2.35 With reference to the radiation leakage data given in Exercise 2.15:

(a) Calculate the sample mean.

(b) Which gives a better indication of the amount of radiation leakage, the sample mean or the median?

2.36 Recent crime reports on the number of aggravated assaults at each of the 27 largest universities reporting for the year are summarized in the computer output

```
Descriptive Statistics: AggAslt

Variable  N   Mean  Median  StDev
AggAslt  27  10.30   10.00   7.61
```

Locate two measures of center tendency, or location, and interpret the values.

2.37 The weights (oz) of nineteen babies born in Madison, Wisconsin, are summarized in the computer output

```
Descriptive Statistics: Weight

Variable  N    Mean  Median  StDev
Weight   19  118.05  117.00  15.47
```

Locate two measures of center tendency, or location, and interpret the values.

2.38 With reference to the extracurricular activities data in Exercise 2.3, obtain the

(a) sample mean.

(b) sample median.

(c) Comment on the effect of a large observation.

2.39 With reference to the number of returns in Example 3, obtain the sample (a) mean and (b) median.

2.40 Old Faithful, the most famous geyser in Yellowstone Park, had the following durations (measured in seconds) in six consecutive eruptions:

240 248 113 268 117 253

(a) Find the sample median.

(b) Find the sample mean.

2.41 Loss of calcium is a serious problem for older women. To investigate the amount of loss, a researcher measured the initial amount of bone mineral content in the radius bone of the dominant hand of elderly women and then the amount remaining after one year. The differences, representing the loss of bone mineral content, are given in the following table (courtesy of E. Smith).

8	7	13	3	6
4	8	6	3	4
0	1	11	7	1
8	6	12	13	10
9	11	3	2	9
7	1	16	3	2
10	15	2	5	8
17	8	2	5	5

(a) Find the sample mean.

(b) Does the sample mean or the median give a better indication of the amount of mineral loss?

2.42 Physical education researchers interested in the development of the overarm throw measured the horizontal velocity of a thrown ball at the time of release. The results for first-grade children (in feet/second) (courtesy of L. Halverson and M. Roberton) are

Males

54.2	39.6	52.3	48.4	35.9	30.4	25.2	45.4	48.9	48.9
45.8	44.0	52.5	48.3	59.9	51.7	38.6	39.1	49.9	38.3

Females

30.3	43.0	25.7	26.7	27.3	31.9	53.7	32.9	19.4	23.7
23.3	23.3	37.8	39.5	33.5	30.4	28.5			

(a) Find the sample median for males.

(b) Find the sample median for females.

(c) Find the sample median for the combined set of males and females.

2.43 On opening day one season, 10 major league baseball games were played and they lasted the following numbers of minutes.

167 211 187 176 170 158 198 218 145 232

Find the sample median.

2.44 If you were to use the data on the length of major league baseball games in Exercise 2.43 to estimate the total amount of digital memory needed to film another 10 major league baseball games, which is the more meaningful description, the sample mean or the sample median? Explain.

2.45 The following measurements of the diameters (in feet) of Indian mounds in southern Wisconsin were gathered by examining reports in the *Wisconsin Archeologist* (courtesy of J. Williams).

22 24 24 30 22 20 28 30 24 34 36 15 37

(a) Create a dot diagram.

(b) Calculate the mean and median and then mark these on the dot diagram.

(c) Calculate the quartiles.

2.46 With reference to Exercise 2.3, calculate the quartiles.

2.47 Refer to the data of college qualification test scores given in Exercise 2.22.

(a) Find the median.

(b) Find Q_1 and Q_3.

2.48 A large mail-order firm employs numerous persons to take phone orders. Computers on which orders are entered also automatically collect data on phone activity. One variable useful for planning staffing levels is the number of calls per shift handled by each employee. From the data collected on 25 workers, calls per shift were (courtesy of Land's End)

118	118	57	92	127	109	96	68	73
69	106	91	93	94	102	105	100	104
80	50	96	82	72	108	73		

Calculate the sample mean.

2.49 With reference to Exercise 2.48, calculate the quartiles.

2.50 The speedy lizard data, from Exercise 2.19, are

1.28	1.36	1.24	2.47	1.94	2.52	2.67	1.29
1.56	2.66	2.17	1.57	2.10	2.54	1.63	2.11
2.57	1.72	0.76	1.02	1.78	0.50	1.49	1.57
1.04	1.92	1.55	1.78	1.70	1.20		

(a) Find the sample median, first quartile, and third quartile.

(b) Find the sample 90th percentile.

2.51 With reference to the water quality data in Exercise 2.17:

(a) Find the sample median, first quartile, and third quartile.

(b) Find the sample 90th percentile.

2.52 ***Some properties of the mean and median.***

1. If a fixed number c is added to all measurements in a data set, then the mean of the new measurements is

   c + (the original mean).

2. If all measurements in a data set are multiplied by a fixed number d, then the mean of the new measurements is

   d × (the original mean).

(a) Verify these properties for the data set

$$4 \quad 8 \quad 8 \quad 7 \quad 9 \quad 6$$

taking $c = 4$ in property (1) and $d = 2$ in (2).

(b) The same properties also hold for the median. Verify these for the data set and the numbers c and d given in part (a).

2.53 On a day, the noon temperature measurements (in °F) reported by five weather stations in a state were

$$74 \quad 80 \quad 76 \quad 76 \quad 73$$

(a) Find the mean and median temperature in °F.

(b) The Celsius (°C) scale is related to the Farenheit (°F) scale by $C = \frac{5}{9}(F - 32)$. What are the mean and median temperatures in °C? (Answer without converting each temperature measurement to °C. Use the properties stated in Exercise 2.52.)

2.54 Given here are the mean and median salaries of machinists employed by two competing companies *A* and *B*.

	Company	
	A	*B*
Mean salary	$70,000	$65,500
Median salary	$56,000	$59,000

Assume that the salaries are set in accordance with job competence and the overall quality of workers is about the same in the two companies.

(a) Which company offers a better prospect to a machinist having superior ability? Explain your answer.

(b) Where can a medium-quality machinist expect to earn more? Explain your answer.

2.55 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on testosterone x_4 for male alligators:

(a) Make separate dot plots for the Lake Apopka and Lake Woodruff alligators.

(b) Calculate the sample means for each group.

(c) Do the concentrations of testosterone appear to differ between the two groups? What does this suggest the contamination has done to male alligators in the Lake Apopka habitat?

2.56 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on testosterone x_4 from Lake Apopka:

(a) Make separate dot plots for the male and female alligators.

(b) Calculate the sample means for each group.

(c) Do the concentrations of testosterone appear to differ between the two groups? We would expect differences. What does your graph suggest the contamination has done to alligators in the Lake Apopka habitat?

5. MEASURES OF VARIATION

Besides locating the center of the data, any descriptive study of data must numerically measure the extent of variation around the center. Two data sets may exhibit similar positions of center but may be remarkably different with respect to variability. For example, the dots in Figure 10*b* are more scattered than those in Figure 10*a*.

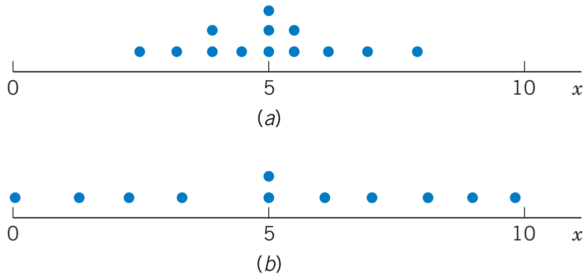


Figure 10 Dot diagrams with similar center values but different amounts of variation.

Because the sample mean $\overline{x}$ is a measure of center, the variation of the individual data points about this center is reflected in their deviation from the mean

$$\begin{aligned}\text{Deviation} &= \text{Observation} - (\text{Sample mean})\\ &= x - \overline{x}\end{aligned}$$

For instance, the data set 3, 5, 7, 7, 8 has mean $\overline{x} = (3 + 5 + 7 + 7 + 8)/5 = 30/5 = 6$, so the deviations are calculated by subtracting 6 from each observation. See Table 9.

TABLE 9 Calculation of Deviations

Observation x	Deviation $x - \overline{x}$
3	−3
5	−1
7	1
7	1
8	2

One might feel that the average of the deviations would provide a numerical measure of spread. However, some deviations are positive and some negative, and the total of the positive deviations exactly cancels the total of the negative ones. In the foregoing example, we see that the positive deviations add to 4 and the negative ones add to −4, so the total deviation is 0. With a little reflection on the definition of the sample mean, the reader will realize that this was not just an accident. For any data set, the total deviation is 0 (for a formal proof of this fact, see Appendix A1).

$$\sum (\text{Deviations}) = \sum (x_i - \overline{x}) = 0$$

To obtain a measure of spread, we must eliminate the signs of the deviations before averaging. One way of removing the interference of signs is to square the numbers. A measure of spread, called the **sample variance,** is constructed by adding the squared deviations and dividing the total by the number of observations minus one.

Sample variance of n observations:

$$s^2 = \frac{\text{sum of squared deviations}}{n - 1}$$

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

Example 10 Calculating Sample Variance

Calculate the sample variance of the data 3 5 7 7 8.

SOLUTION For this data set, $n = 5$. To find the variance, we first calculate the mean, then the deviations and the squared deviations. See Table 10.

TABLE 10 Calculation of Variance

	Observation x	Deviation $x - \bar{x}$	(Deviation)^2 $(x - \bar{x})^2$
	3	-3	9
	5	-1	1
	7	1	1
	7	1	1
	8	2	4
Total	30 $\sum x$	0 $\sum (x - \bar{x})$	16 $\sum (x - \bar{x})^2$

$$\bar{x} = \frac{30}{5} = 6$$

$$\text{Sample variance} \quad s^2 = \frac{16}{5 - 1} = 4$$

Remark: Although the sample variance is conceptualized as the **average squared deviation,** notice that the divisor is $n - 1$ rather than n. The divisor, $n - 1$, is called the degrees of freedom[1] associated with s^2.

Because the variance involves a sum of squares, its unit is the square of the unit in which the measurements are expressed. For example, if the data pertain to measurements of weight in pounds, the variance is expressed in $(\text{pounds})^2$. To obtain a measure of variability in the same unit as the data, we take the positive square root of the variance, called the **sample standard deviation.** The standard deviation rather than the variance serves as a basic measure of variability.

Sample Standard Deviation

$$s = \sqrt{\text{Variance}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Example 11 Calculating the Sample Standard Deviation

Calculate the standard deviation for the data of Example 10.

SOLUTION We already calculated the variance $s^2 = 4$ so the standard deviation is $s = \sqrt{4} = 2$.

To show that a larger spread of the data does indeed result in a larger numerical value of the standard deviation, we consider another data set in Example 12.

Example 12 Using Standard Deviations to Compare Variation in Two Data Sets

Calculate the standard deviation for the data 1, 4, 5, 9, 11. Plot the dot diagram of this data set and also the data set of Example 10.

SOLUTION The standard deviation is calculated in Table 11. The dot diagrams, given in Figure 11, show that the data points of Example 10 have less spread than those of Example 12. This visual comparison is confirmed by a smaller value of s for the first data set.

[1] The deviations add to 0 so a specification of any $n - 1$ deviations allows us to recover the one that is left out. For instance, the first four deviations in Example 10 add to -2, so to make the total 0, the last one must be $+2$, as it really is. In the definition of s^2, the divisor $n - 1$ represents the number of deviations that can be viewed as free quantities.

TABLE 11 Calculation of s

	x	$(x - \bar{x})$	$(x - \bar{x})^2$
	1	-5	25
	4	-2	4
	5	-1	1
	9	3	9
	11	5	25
Total	30	0	64

$$\bar{x} = 6 \qquad s^2 = \frac{64}{4} = 16$$

$$s = \sqrt{16} = 4$$

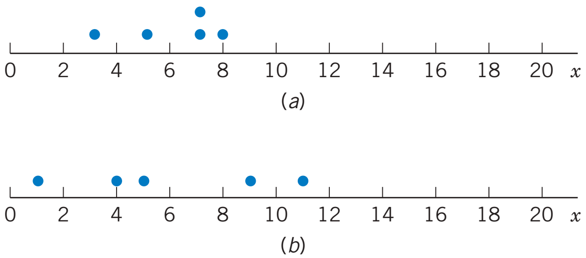


Figure 11 Dot diagrams of two data sets.

An alternative formula for the sample variance is

$$s^2 = \frac{1}{n - 1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$$

It does not require the calculation of the individual deviations. In hand calculation, the use of this alternative formula often reduces the arithmetic work, especially when $\bar{x}$ turns out to be a number with many decimal places. The equivalence of the two formulas is shown in Appendix A1.2.

Example 13 Calculating Sample Variance Using the Alternative Formula

In a psychological experiment a stimulating signal of fixed intensity was used on six experimental subjects. Their reaction times, recorded in seconds, were 4, 2, 3, 3, 6, 3. Calculate the standard deviation for the data by using the alternative formula.

SOLUTION These calculations can be conveniently carried out in tabular form:

	x	x^2
	4	16
	2	4
	3	9
	3	9
	6	36
	3	9
Total	21 $= \sum x$	83 $= \sum x^2$

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] = \frac{83 - (21)^2/6}{5} = \frac{83 - 73.5}{5}$$

$$= \frac{9.5}{5} = 1.9$$

$$s = \sqrt{1.9} = 1.38 \text{ seconds}$$

The reader may do the calculations with the first formula and verify that the same result is obtained.

In Example 12, we have seen that one data set with a visibly greater amount of variation yields a larger numerical value of s. The issue there surrounds a comparison between different data sets. In the context of a single data set, can we relate the numerical value of s to the physical closeness of the data points to the center $\bar{x}$? To this end, we view one standard deviation as a benchmark distance from the mean $\bar{x}$. For bell-shaped distributions, an empirical rule relates the standard deviation to the proportion of the data that lie in an interval around $\bar{x}$.

Empirical Guidelines for Symmetric Bell-Shaped Distributions

Approximately	68%	of the data lie within $\bar{x} \pm s$
	95%	of the data lie within $\bar{x} \pm 2s$
	99.7%	of the data lie within $\bar{x} \pm 3s$

Example 14 Comparing the Sleep Data with the Empirical Guidelines

Examine the 59 hours of sleep in Table 4 in the context of the empirical guideline.

SOLUTION Using a computer (see, for instance, Exercise 2.124), we obtain

$$\bar{x} = 7.18$$

$$s = 1.28 \qquad 2s = 2(1.28) = 2.56$$

Going two standard deviations either side of $\overline{x}$ results in the interval

$$7.18 - 2.56 = 4.62 \quad \text{to} \quad 9.74 = 7.18 + 2.56$$

By actual count, all the observations except 4.5 and 10.0 fall in this interval. We find that 57/59 = .966, or 96.6% of the observations lie within two standard deviations of $\overline{x}$. The empirical guidelines suggest 95% so they are close.

Other Measures of Variation

Another measure of variation that is sometimes employed is

Sample range = Largest observation − Smallest observation

The range gives the length of the interval spanned by the observations.

Example 15 Calculating the Sample Range

Calculate the range for the hours of sleep data given in Example 5.

SOLUTION The data given in Table 4 contained

$$\begin{aligned} \text{Smallest observation} &= 4.5 \\ \text{Largest observation} &= 10.0 \end{aligned}$$

Therefore, the length of the interval covered by these observations is

$$\text{Sample range} = 10.0 - 4.5 = 5.5 \text{ hours}$$

As a measure of spread, the range has two attractive features: It is extremely simple to compute and interpret. However, it suffers from the serious disadvantage that it is much too sensitive to the existence of a very large or very small observation in the data set. Also, it ignores the information present in the scatter of the intermediate points.

To circumvent the problem of using a measure that may be thrown far off the mark by one or two wild or unusual observations, a compromise is made by measuring the interval between the first and third quartiles.

Sample interquartile range = Third quartile − First quartile

The sample interquartile range represents the length of the interval covered by the center half of the observations. This measure of the amount of variation is not disturbed if a small fraction of the observations are very large or very small. The sample interquartile range is usually quoted in government reports on income and other distributions that have long tails in one direction, in preference to standard deviation as the measure of spread.

Example 16 Calculating the Interquartile Range

Calculate the sample interquartile range for the length of long distance phone calls data given in Table 8.

SOLUTION In Example 9, the quartiles were found to be $Q_1 = 4.4$ and $Q_3 = 17.5$. Therefore,

$$\begin{aligned}\text{Sample interquartile range} &= Q_3 - Q_1\\ &= 17.5 - 4.4\\ &= 13.1 \text{ minutes}\end{aligned}$$

Boxplots

A recently created graphic display, called a **boxplot,** highlights the summary information in the quartiles. Begin with the

Five-number summary: minimum, Q_1, Q_2, Q_3, maximum.

The center half of the data, from the first to the third quartile, is represented by a rectangle (box) with the median indicated by a bar. A line extends from Q_3 to the maximum value and another from Q_1 to the minimum. Figure 12 gives the boxplot for the length of phone calls data in Table 8. The long line to the right is a consequence of the largest value, 53.3 minutes, and, to some extent, the second largest value, 43.5 minutes.

Boxplots are particularly effective for displaying several samples alongside each other for the purpose of visual comparison.

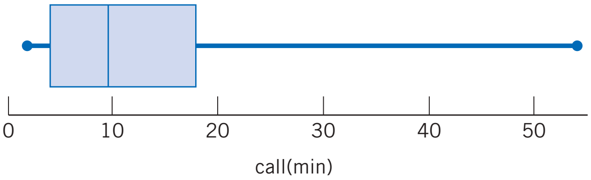


Figure 12 Boxplot of the length of phone call data in Table 8.

Figure 13 displays the amount of reflected light in the near-infrared band as recorded by satellite when flying over forest areas and urban areas, respectively. Because high readings tend to correspond to forest and low readings to urban areas, the readings have proven useful in classifying unknown areas.

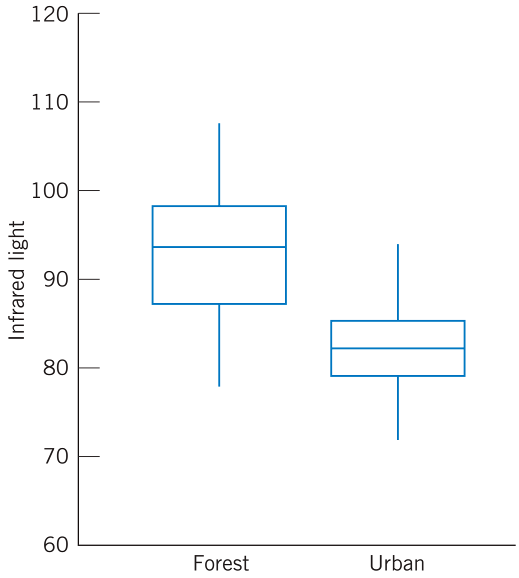


Figure 13 Boxplots of near-infrared light reflected from forest and urban areas.

Example 17 Comparing Boxplots for the Reflected Light Data

Refer to the boxplots for the amount of reflected near-infrared light in Figure 13.

(a) Do forests or urban areas produce the largest median reflected light?

(b) Which has the largest interquartile range, IQR?

SOLUTION

(a) It is clear that forests have the largest median. Its median is over 10 units higher than that of urban areas.

(b) The height of the box is the IQR. The IQR of the forest data is over twice that of the IQR for urban areas.

Exercises

2.57 For the data set

7 2 3

(a) Calculate the deviations $(x - \bar{x})$ and check to see that they add up to 0.

(b) Calculate the sample variance and the standard deviation.

2.58 Repeat (a) and (b) of Exercise 2.57 for the data set

4 9 2

2.59 For the data set 8 6 14 4:

(a) Calculate the deviations $(x - \bar{x})$ and check to see that they add up to 0.

(b) Calculate the variance and the standard deviation.

2.60 Repeat (a) and (b) of Exercise 2.59 for the data set

2.5 1.7 2.1 1.5 1.7

2.61 For the data of Exercise 2.57, calculate s^2 by using the alternative formula.

2.62 For the data of Exercise 2.59, calculate s^2 by using the alternative formula.

2.63 For each data set, calculate s^2.

(a) 1 4 3 2 2

(b) −2 1 −1 −3 0 −2

(c) 9 8 8 9 8 8 9

2.64 The monthly rents for 7 one-bedroom apartments located in one area of the city, are

625 740 805 670 705 740 870

(a) Give two possible factors that may contribute to variation in the monthly rents.

Calculate

(b) The sample variance.

(c) The sample standard deviation.

2.65 Find the standard deviation of the measurements of diameters given in Exercise 2.45.

2.66 A campus area merchant recorded the number of bad checks received per month, for five months

4 5 4 7 6

Calculate:

(a) The sample variance.

(b) The sample standard deviation.

2.67 The city of Madison regularly checks the quality of water at swimming beaches located on area lakes. Fifteen times the concentration of fecal coliforms, in number of colony forming units (CFU) per 100 ml of water, was measured during the summer at one beach.

180 1600 90 140 50 260 400 90
380 110 10 60 20 340 80

(a) Calculate the sample variance.

(b) Calculate the sample standard deviation.

(c) One day, the water quality was bad—the reading was 1600 CFU—and the beach was closed. Drop this value and calculate the sample standard deviation for the days where the water quality was suitable for swimming. Comment on the change.

2.68 With reference to the radiation leakage data given in Exercise 2.15, calculate:

(a) The sample variance.

(b) The sample standard deviation.

2.69 With reference to the data on the length of 10 major league baseball games in Exercise 2.43:

(a) Find the sample mean.

(b) Find the sample variance.

(c) Find the sample standard deviation.

2.70 With reference to checked bags in Exercise 2.13,

(a) Find the sample mean.

(b) Find the sample standard deviation.

2.71 A sample of seven compact discs at the music store stated the performance times as lasting the following numbers of minutes for Beethoven's Ninth Symphony.

66.9 66.2 71.0 68.6 65.4 68.4 71.9

(a) Find the sample median.

(b) Find the sample mean.

(c) Find the sample standard deviation.

2.72 Recent crime reports on the number of aggravated assaults at each of the 27 largest universities reporting for the year are summarized in the computer output.

```
Descriptive Statistics: AggAslt
Variable  N   Mean   Median   StDev
AggAslt   27  10.30  10.00     7.61

Variable Minimum Maximum  Q1    Q3
AggAslt    0.00   29.00  5.00 14.00
```

(a) Locate a measure of variation and also calculate the sample variance.

(b) Calculate the interquartile range and interpret this value.

(c) Give a value for a standard deviation that would correspond to greater variation in the numbers of aggravated assaults.

2.73 The weights (oz) of nineteen babies born in Madison, Wisconsin, are summarized in the computer output.

```
Descriptive Statistics: Weight

Variable      N    Mean   Median   StDev
Weight       19  118.05   117.00   15.47

Variable Minimum Maximum     Q1      Q3
Weight      89.00  144.00 106.00  131.00
```

(a) Locate a measure of variation and also calculate the sample variance.

(b) Calculate the interquartile range and interpret this value.

(c) Give a value for a standard deviation that would correspond to smaller variation in the weights.

2.74 ***Some properties of the standard deviation.***

1. If a fixed number c is added to all measurements in a data set, the deviations $(x - \bar{x})$ remain unchanged (see Exercise 2.52). Consequently, s^2 and s remain unchanged.
2. If all measurements in a data set are multiplied by a fixed number d, the deviations $(x - \bar{x})$ get multiplied by d. Consequently, s^2 gets multiplied by d^2, and s by $|d|$. (*Note:* The standard deviation is never negative.)

Verify these properties for the data set

5 9 9 8 10 7

taking $c = 4$ in property (1) and $d = 2$ in (2).

2.75 For the data set of Exercise 2.22, calculate the interquartile range.

2.76 For the extracurricular data of Exercise 2.3, calculate the interquartile range.

2.77 Should you be surprised if the range is larger than twice the interquartile range? Explain.

2.78 Calculations with the test scores data of Exercise 2.22 give $\bar{x} = 150.125$ and $s = 24.677$.

(a) Find the proportion of the observations in the intervals $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$.

(b) Compare your findings in part (a) with those suggested by the empirical guidelines for bell-shaped distributions.

2.79 Refer to the data on bone mineral content in Exercise 2.41.

(a) Calculate $\bar{x}$ and s.

(b) Find the proportion of the observations that are in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

(c) Compare the results of part (b) with the empirical guidelines.

2.80 Refer to the data on lizards in Exercise 2.19.

(a) Calculate $\bar{x}$ and s.

(b) Find the proportion of the observations that are in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

(c) Compare the results of part (b) with the empirical guidelines.

2.81 Refer to the data on number of returns in Example 3.

(a) Calculate $\bar{x}$ and s.

(b) Find the proportions of the observations that are in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

(c) Compare the results of part (b) with the empirical guidelines.

2.82 **Sample z score.** The z *scale* (or *standard scale*) measures the position of a data point relative to the mean and in units of the standard deviation. Specifically,

$$z \text{ value of a measurement } = \frac{\text{Measurement} - \bar{x}}{s}$$

When two measurements originate from different sources, converting them to the z scale helps to draw a sensible interpretation of their relative magnitudes. For instance, suppose a student scored 65 in a math course and 72 in a history course. These (raw) scores tell little about the student's performance. If the class averages and standard deviations were $\bar{x} = 60$, $s = 20$ in math and $\bar{x} = 78$, $s = 10$ in history, this student's

$$z \text{ score in math} = \frac{65 - 60}{20} = .25$$

$$z \text{ score in history} = \frac{72 - 78}{10} = -.60$$

Thus, the student was .25 standard deviations above the average in math and .6 standard deviations below the average in history.

(a) If $\bar{x} = 490$ and $s = 120$, find the z scores of 350 and 620.

(b) For a z score of 2.4, what is the raw score if $\bar{x} = 210$ and $s = 50$?

2.83 The weights (oz) of nineteen babies born in Madison, Wisconsin, are summarized in the computer output.

```
Descriptive Statistics: Weight
Variable N    Mean   Median   StDev
Weight   19 118.05   117.00   15.47
```

Referring to Exercise 2.82 obtain the z score for a baby weighing

(a) 102 oz
(b) 144 oz

2.84 Two cities provided the following information on public school teachers' salaries.

	Minimum	Q_1	Median	Q_3	Maximum
City *A*	38,400	44,000	48,300	50,400	56,300
City *B*	39,600	46,500	51,200	55,700	61,800

(a) Construct a boxplot for the salaries in City *A*.

(b) Construct a boxplot, on the same graph, for the salaries in City *B*.

(c) Are there larger differences at the lower or the higher salary levels? Explain.

2.85 Refer to the data on throwing speed in Exercise 2.42. Make separate boxplots to compare males and females.

2.86 Refer to Exercise 2.27 and the data on the consumer price index for various cities. Find the increase, for each city, by subtracting the 2001 value from the 2007 value.

(a) Obtain the five-number summary: minimum, Q_1, Q_2, Q_3, and maximum. Which city had the largest increase? Were there any decreases?

(b) Make a boxplot of the increases.

2.87 Refer to Exercise 2.27 and the data on the consumer price index for various cities. Find the increase, for each city, by subtracting the 2001 value from the 2007 value.

(a) Find the sample mean and standard deviation of these differences.

(b) What proportion of the increases lie between $\bar{x} \pm 2s$?

2.88 Refer to Example 5 and the data on hours of sleep

(a) Obtain the five-number summary: minimum, Q_1, Q_2, Q_3, and maximum.

(b) Make a boxplot of the hours of sleep.

2.89 Refer to Exercise 2.3 and the data on extracurricular activities. Find the sample mean and standard deviation.

2.90 Presidents also take midterms! After two years of the President's term, members of Congress are up for election. The following table gives the number of net seats lost, by the party of the President, in the House of Representatives since the end of World War II.

Net House Seats Lost in Midterm Elections

1950	Truman (D)	55
1954	Eisenhower (R)	16
1962	Kennedy (D)	4
1966	Johnson (D)	47
1970	Nixon (R)	12
1974	Nixon/Ford (R)	43
1978	Carter (D)	11
1982	Reagan (R)	26
1986	Reagan (R)	5
1990	Bush (R)	8
1994	Clinton (D)	52
1998	Clinton (D)	−5 (gain)
2002	Bush (R)	−8
2006	Bush (R)	30

For the data on the number of House seats lost:

(a) Calculate the sample mean.

(b) Calculate the standard deviation.

(c) Make a dot plot.

(d) What is one striking feature of these data that could be useful in predicting future midterm election results? (*Hint:* Would you have expected more elections to result in net gains?)

2.91 With reference to Exercise 2.90:

(a) Calculate the median number of lost House seats.

(b) Find the maximum and minimum losses and identify these with a President.

(c) Determine the range for the number of House seats lost.

6. CHECKING THE STABILITY OF THE OBSERVATIONS OVER TIME

The calculations for the sample mean and sample variance treat all the observations alike. The presumption is that there are no apparent trends in data over time and there are no unusual observations. Another way of saying this is that the process producing the observations is in **statistical control.** The concept of statistical control allows for variability in the observations but requires that the pattern of variability be the same over time. Variability should not increase or decrease with time and the center of the pattern should not change.

To check on the stability of the observations over time, observations should be plotted versus time, or at least the order in which they were taken. The resulting plot is called a **time plot** or sometimes a **time series plot.**

Example 18 A Time Plot of Overtime Hours

The Madison Police Department charts several important variables, one of which is the number of overtime hours due to extraordinary events. These events would include murders, major robberies, and so forth. Although any one event is not very predictable, there is some constancy when data are grouped into six-month periods.

The values of overtime hours for extraordinary events for eight recent years, beginning with 2200, 875, . . . , through 1223, are

2200	875	957	1758	868	398	1603	523
2034	1136	5326	1658	1945	344	807	1223

Is the extraordinary event overtime hours process in control? Construct a time plot and comment.

SOLUTION The time plot is shown in Figure 14. There does not appear to be any trend, but there is one large value of 5326 hours.

Example 19 A Time Plot of the Yen/Dollar Exchange Rate

The exchange rate between the United States and Japan can be stated as the number of yen that can be purchased with $1. Although this rate changes daily, we quote the official value for the year:

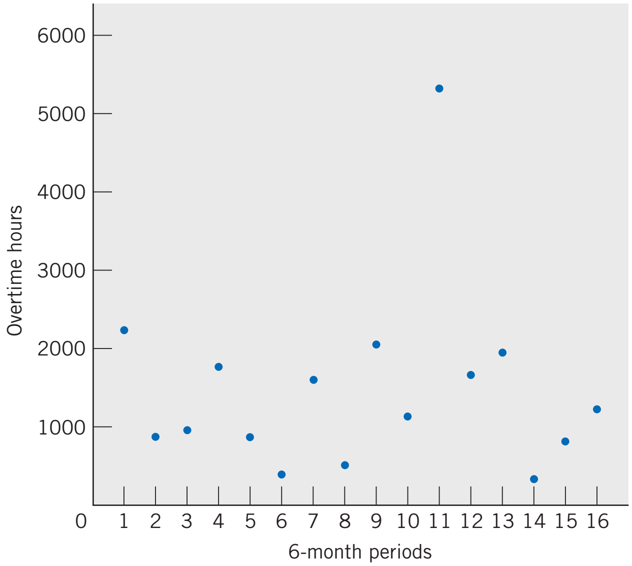


Figure 14 Time plot of extraordinary event hours versus time order.

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Exchange rate	238.5	168.4	144.6	128.2	138.1	145.0	134.6	126.8	111.1	102.2	94.0	108.8	121.1	130.9	113.7	107.8	121.6	125.2	115.9	108.2	110.1	116.3	117.8	103.4

Is this exchange rate in statistical control? Make a time plot and comment.

SOLUTION The time plot is shown in Figure 15 on page 62. There is a rather strong downhill trend over most of the time period so the exchange rate is definitely not in statistical control. A dollar has purchased fewer and fewer yen over the years. It is the downward trend that is the primary feature and a cause for serious concern with regard to trade deficits. There is somewhat of a leveling off in the last half suggesting a change in trend.

Manufacturers need to monitor critical dimensions, temperatures, and other variables so that, although they vary, the variation is kept small enough so that the quality of the final product is maintained. A graphical approach, called a **control chart,** is recommended for this purpose because it allows for the visual inspection of outliers and trends. It adds a **centerline** and **control limits** to the time plot to help identify unusual observations.

To construct a control chart:

1. Plot the observations versus time order.
2. Add a solid centerline at the level of the sample mean $\overline{x}$.
3. Add dashed lines for the control limits at $\overline{x} - 2s$ and $\overline{x} + 2s$.

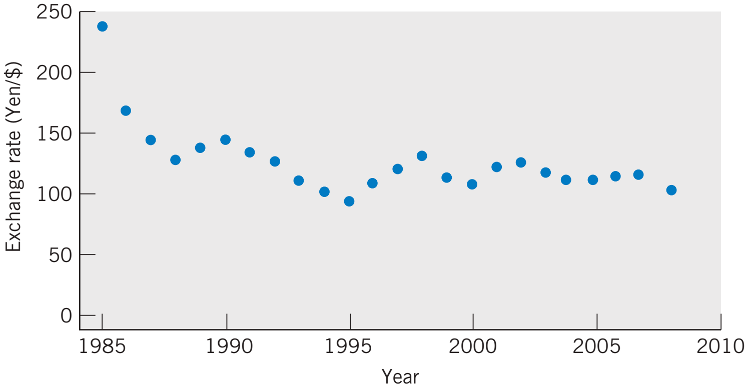


Figure 15 Time plot of the exchange rate.

According to the empirical rule, if the process is in statistical control so the observations are stable over time, only about 5% of the observations will fall outside of the control limits. In many applications, manufacturers use $\bar{x} - 3s$ and $\bar{x} + 3s$ as the control limits.

The upper and lower control limits on the control charts help to identify unusually low or unusually high observations. It allows us to distinguish between typical variation and variation that is especially large and could be due to special or assignable causes. Any time an observation falls outside of the control limits, an effort should be made to search for the reason.

Example 20 A Control Chart for Overtime Hours

Manufacturing processes are not the only ones that can benefit from control charting. Refer to the data in Example 18 on the number of overtime hours for police due to extraordinary events. Is the extraordinary event overtime hours process in control? Construct a control chart and comment.

SOLUTION A computer calculation gives $\bar{x} = 1478$ and $s = 1183$ so the centerline is drawn at the sample mean 1478 and the upper control limit is $\bar{x} + 2s = 1478 + 2 \times 1183 = 3844$. The lower control limit is negative; we replace it by 0. Figure 16 gives the resulting control chart for extraordinary event overtime hours.

There is no discernible trend, but one point does exceed the upper control limit. By checking more detailed records, it was learned that the point outside of the control limits occurred when protests took place on campus in response to the bombing of a foreign capital. These events required city police to serve 1773 extraordinary overtime hours in just one 2-week period and 683 in the next period. That is, there was really one exceptional event, or special cause, that could be identified with the unusual point.

The one large value, 5326 hours, not only affects the centerline by inflating the mean, but it also increases the variance and that raises the upper control limit. In Exercise 2.98, you are asked to remove this outlier and redo the control chart.

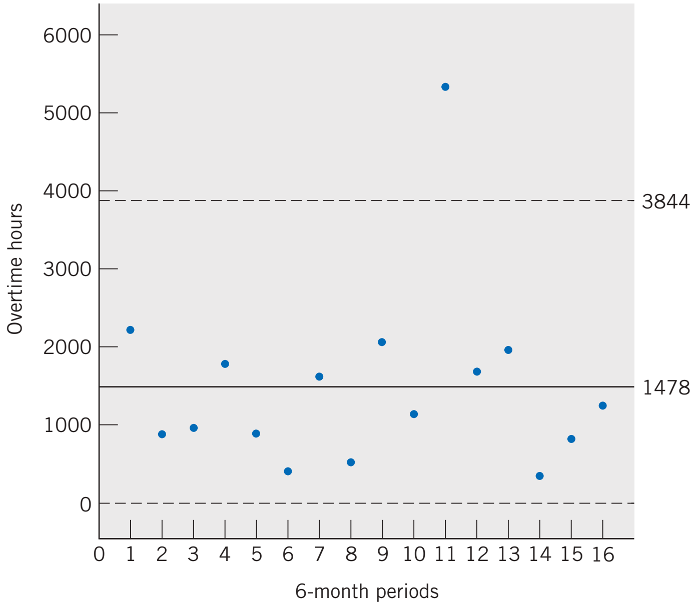


Figure 16 Control chart for extraordinary event overtime hours.

Exercises

2.92 Make a time plot of the phone call data in Exercise 2.48 and comment on the statistical control.

2.93 A city department has introduced a quality improvement program and has allowed employees to get credit for overtime hours when attending meetings of their quality groups. The total number of overtime meeting hours for each of the 26 pay periods in one year by row were

30	215	162	97	194	163	60	41	100
43	96	69	80	42	162	75	95	65
57	131	54	114	64	114	38	140	

Make a time plot of the overtime meeting hours data.

2.94 Make a control chart for the data referred to in Exercise 2.92 and comment.

2.95 Make a control chart for the data in Exercise 2.93 and comment.

2.96 The exchange rate between the United States and Canada can be stated as the number of Canadian dollars that can be purchased with $1. The official values for the year are

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Exchange rate	1.21	1.29	1.37	1.37	1.36	1.38	1.48	1.49	1.48	1.55	1.57	1.40	1.30	1.21	1.13	1.07	1.07

Is this exchange rate in statistical control? Make a time plot and comment.

2.97 Make a control chart of the data in Exercise 2.96 and comment.

2.98 Make a control chart for the extraordinary event overtime data in Example 18 after removing the outlier identified in that example. You need to recalculate the mean and standard deviation.

7. MORE ON GRAPHICS

The importance of graphing your data cannot be overemphasized. If a feature you expect to see is not present in the plots, statistical analyses will be of no avail. Moreover, creative graphics can often highlight features in the data and even give new insights.

The devastation of Napoleon's Grand Army during his ill-fated attempt to capture Russia was vividly depicted by Charles Minard. The 422,000 troops that entered Russia near Kaunas are shown as a wide (shaded) river flowing toward Moscow and the retreating army as a small (black) stream. The width of the band indicates the size of the army at each location on the map. Even the simplified version of the original graphic, appearing in Figure 17, dramatically conveys the losses that reduced the army of 422,000 men to 10,000 returning members. The temperature scale at the bottom, pertaining to the retreat, helps to explain the loss of life, including the incident where thousands died trying to cross the Berezina River in subzero temperatures. (For a copy of Minard's more detailed map and additional discussion, see E. R. Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press, 1983.)

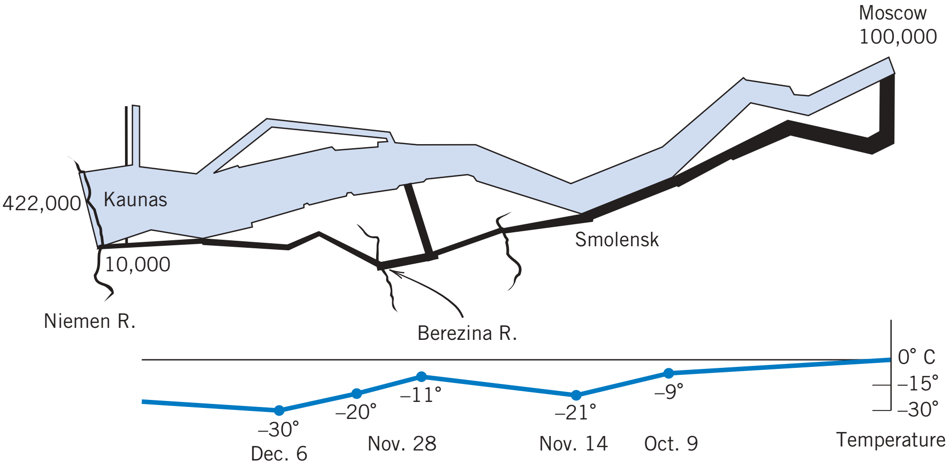


Figure 17 The demise of Napoleon's army in Russia, 1812–1813, based on Charles Minard.

Another informative graphic, made possible with modern software, is the display of ozone by city in Figure 18. This figure illustrates improved air quality relative to the previous year and the ten year average.

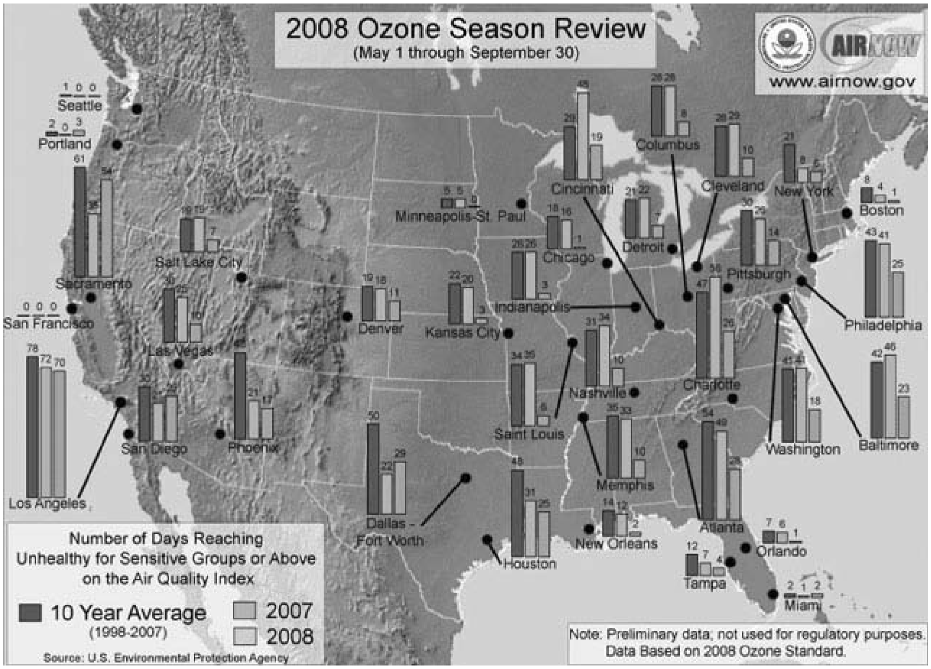


Figure 18 Ozone Season for 2008 compared to 2007 and ten year average.

Graphs can give a vivid overall picture.
Dave Carpenter/Cartoon Stock

8. STATISTICS IN CONTEXT

The importance of visually inspecting data cannot be overemphasized. We present a mini-case study[2] that shows the importance of first appropriately plotting and then monitoring manufacturing data. This statistical application concerns a ceramic part used in a popular brand of coffee makers. To make this ceramic part, a mixture of clay, water, and oil is poured into the cavity between two dies of a pressing machine. After pressing but before the part is dried to a hardened state, critical dimensions are measured. The depth of a slot is of interest here.

Sources of variation in the slot depth abound: the natural uncontrolled variation in the clay–water–oil mixture, the condition of the press, differences in operators, and so on. Some variation in the depth of the slot is inevitable. Even so, for the part to fit when assembled, the slot depth needs to be controlled within certain limits.

Every half hour during the first shift, slot depth is measured on three ceramic parts selected from production. Table 12 gives the data obtained on a Friday. The sample mean for the first sample of 218, 217, and 219 (thousandths of inch) is $(218 + 217 + 219)/3 = 654/3 = 218$, and so on.

TABLE 12 Slot Depth (thousandths of an inch)

Time	7:00	7:30	8:00	8:30	9:00	9:30	10:00
1	218	218	216	217	218	218	219
2	217	218	218	220	219	217	219
3	219	217	219	221	216	217	218
SUM	654	653	653	658	653	652	656
$\bar{x}$	218.0	217.7	217.7	219.3	217.7	217.3	218.7

Time	10:30	11:00	11:30	12:30	1:00	1:30	2:00	2:30
1	216	216	218	219	217	219	217	215
2	219	218	219	220	220	219	220	215
3	218	217	220	221	216	220	218	214
SUM	653	651	657	660	653	658	655	644
$\bar{x}$	217.7	217.0	219.0	220.0	217.7	219.3	218.3	214.7

An x-bar chart will indicate when changes have occurred and there is a need for corrective actions. Because there are 3 slot measurements at each time, it is the 15 sample means that are plotted versus time order. We will take the centerline to be the mean of the 15 sample means, or

$$\text{Centerline: } \bar{\bar{x}} = \frac{218.0 + \cdots + 214.7}{15} = 218.0$$

[2] Courtesy of Don Ermer.

When each plotted mean is based on several observations, the variance can be estimated by combining the variances from each sample. The first sample has variance $s_1^2 = [(218 - 218)^2 + (217 - 218)^2 + (219 - 218)^2]/(3 - 1) = 1.000$ and so on. The details are not important, but a computer calculation of the variance used to set control limits first determines the average of the 15 individual sample variances,

$$\frac{1.000 + 0.333 + \cdots + 0.333}{15} = 1.58$$

and, for reasons given in Chapter 7, divides by 3 to give the variance of a single sample mean. That is, $1.58/3 = .527$ is the appropriate s^2. The control limits are set at three times the estimated standard deviation $s = \sqrt{.527} = .726$, or $3 \times .726 = 2.2$ units from the centerline.

$$\text{Lower control limit: LCL} = 218.0 - 2.2 = 215.8$$
$$\text{Upper control limit: UCL} = 218.0 + 2.2 = 220.2$$

The x-bar chart is shown in Figure 19. What does the chart tell us?

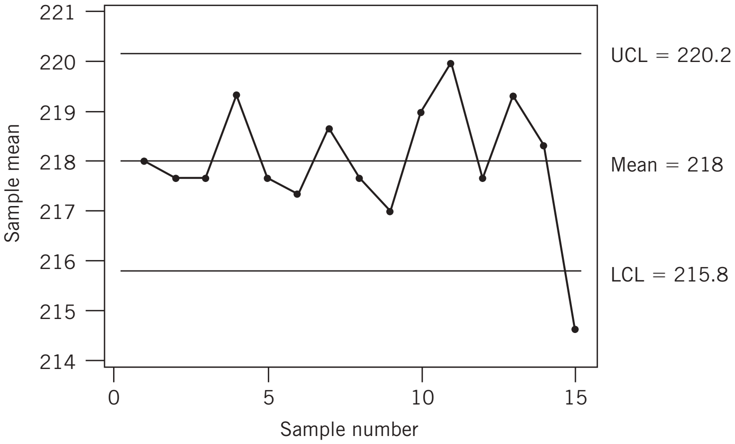


Figure 19 X-bar chart for depth.

The x-bar chart shows that the process was stable throughout the day and no points were out of control except the last sample. It was then that an unfortunate oversight occurred. Because it was near the end of her shift and the start of the weekend, the operator did not report the out-of-control value to either the setup person or the foreman. She knew the setup person was already cleaning up for the end of the shift and that the foreman was likely thinking about going across the street to the Legion Bar for some refreshments as soon as the shift ended. The operator did not want to ruin anyone's weekend plans so she kept quiet.

When the pressing machine was started up on Monday morning, one of the dies broke. The cost of the die was over a thousand dollars. But, when a customer was called and told there would be a delay in delivering the ceramic parts, he canceled the order. Certainly the loss of a customer is an even more expensive item.

Later, it was concluded that the clay had likely dried and stuck to the die leading to the break. A problem was predicted by the chart on Friday. Although the chart correctly indicated a problem at that time, someone had to act for the monitoring procedure to work.

USING STATISTICS WISELY

1. As a first step, always graph the data as a dot diagram or histogram to assess the overall pattern of data.
2. When comparing histograms based on different class intervals, be sure to create histograms whose height is relative frequency divided by width of interval.
3. Calculate summary statistics to describe the data set. Always determine the sample mean and standard deviation. The five-number summary

   minimum first quartile median third quartile maximum

   provides an additional summary when the sample sizes are moderately large. It helps describe cases where the dot diagram or histogram has a single long tail.
4. Use the median to describe the center when a small sample contains an extreme observation. The sample median is not influenced by a few very large or very small observations that may even be incorrectly recorded.
5. Do not routinely calculate summary statistics without identifying unusual observations (outliers) which may have undue influence on the value of a summary statistic.

KEY IDEAS AND FORMULAS

Qualitative data refer to frequency counts in categories. These are summarized by calculating the

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Total number of observations}}$$

for the individual categories.

The term **numerical-valued variable** or just **variable** refers to a characteristic that varies over units and is measured on a numerical scale. **Discrete variables** are usually counts and all discrete variables have gaps in their scale of measurement. **Continuous variables,** like height or weight, can conceptually take any value in an interval. Data resulting from measurements of a variable are either **discrete** or **continuous** data.

For a **discrete data** set, the **frequency** is the count of the number of observations having a distinct value. The **relative frequency** is the proportion of sample units having this property.

$$\text{Relative frequency} \;=\; \frac{\text{Frequency}}{\text{Total number of observations}}$$

The discrete data set is summarized by a **frequency distribution** that lists the distinct data points and their corresponding relative frequencies. Either a **line diagram** or a **histogram** can be used for a graphical display.

Continuous measurement data should be graphed as a **dot diagram** when the data set is small, say, fewer than 20 or 25 observations. Larger data sets are first summarized in a **frequency table.** This is constructed by grouping the observations in **class intervals,** preferably of equal lengths. The class intervals are non-overlapping and cover the range of the data set from smallest to largest. We recommend specifying an **endpoint convention** that tells which of the **class boundaries,** or endpoints of the class intervals, to include and which to exclude from each class interval. A list of the class intervals along with the corresponding relative frequencies provides a **frequency distribution** which can graphically be displayed as a **histogram.** The histogram is constructed to have total area 1, equal to total relative frequency. That is, for each class interval, we draw a rectangle whose **area represents the relative frequency** of the class interval.

A **stem-and-leaf display** is another effective means of display when the data set is not too large. It is more informative than a histogram because it retains the individual observations in each class interval instead of lumping them into a frequency count. Two variants are the **double-stem display** and **five-stem display.**

Pareto diagrams display events according to their frequency in order to highlight the most important few that occur most of the time.

A summary of measurement data (discrete or continuous) should also include numerical measures of center and spread.

Two important measures of center are

$$\textbf{Sample mean} \qquad \bar{x} = \frac{\sum x}{n}$$

$$\textbf{Sample median} \;=\; \text{middle most value of the ordered data set}$$

The **quartiles** and, more generally, **percentiles** are other useful locators of the distribution of a data set. The second quartile is the same as the median. The **sample quartiles** divide the **ordered data** into nearly four equal parts. The **100*p*-th percentile** has least proportion p at or below and proportion $1 - p$ at or above.

The amount of **variation,** or **spread,** of a data set is measured by the **sample standard deviation** s. The **sample variance** s^2 is given by

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$\text{Also, } s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right] \quad \text{(convenient for hand calculation)}$$

$$\text{Sample standard deviation } s = +\sqrt{s^2}$$

The standard deviation indicates the amount of spread of the data points around the mean $\bar{x}$. If the histogram appears **symmetric and bell-shaped,** then the interval

$\bar{x} \pm s$ includes approximately 68% of the data

$\bar{x} \pm 2s$ includes approximately 95% of the data

$\bar{x} \pm 3s$ includes approximately 99.7% of the data

Two other measures of variation are

Sample range = Largest observation − Smallest observation

and

Sample interquartile range = Third quartile − First quartile

The **five-number summary**, namely, the median, the first and third quartiles, the smallest observation, and the largest observation, together serve as useful indicators of the distribution of a data set. These are displayed in a **boxplot.**

TECHNOLOGY

Creating graphs and computing statistical summaries have become considerably easier because of recent developments in software. In all our professional applications of statistics, we begin by entering the data in a worksheet. We then read from the worksheet while another person checks against the original. This procedure has eliminated many errors in data entry and allowed us to proceed knowing that the computer software is using the correct data.

In the technology sections of this text, we give the essential details for using MINITAB, EXCEL, and the TI-84/83 Plus graphing calculator. The first two use the worksheet format.

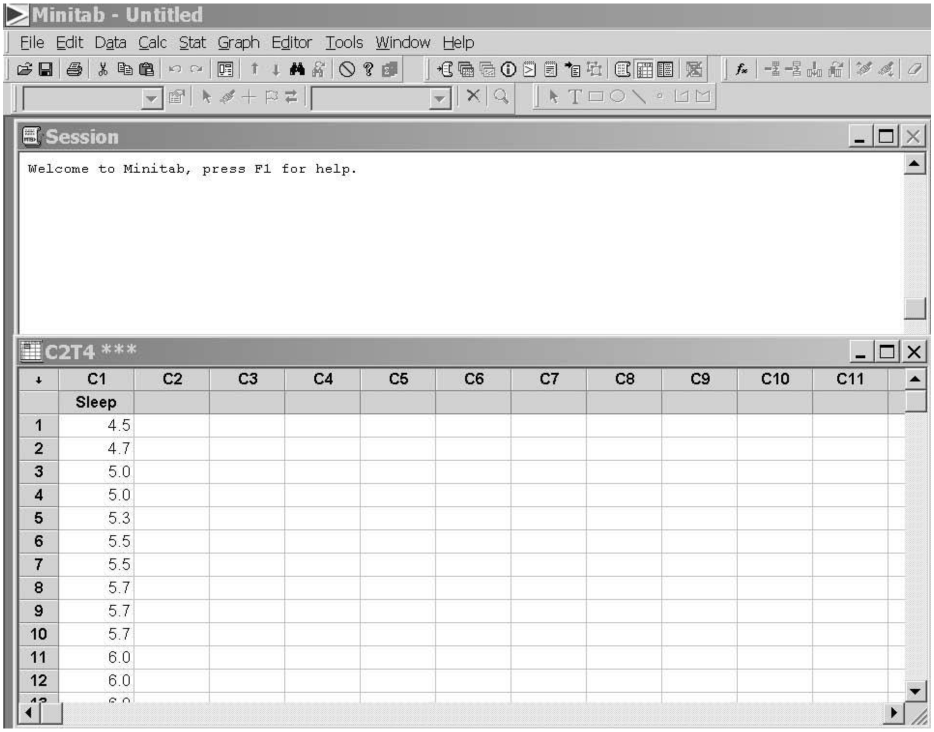
MINITAB

The MINITAB screen is split with the bottom part being the worksheet. In the example here, we have typed sleep in the top entry and then in row 1 of the first column we have typed the first number of hours of sleep in Table 4. The rest of the hours of sleep are typed in the other cells in column 1.

Alternatively, the data sets in the book are stored as MINITAB worksheets on the book's Web site. Go to www.wiley.com/college/johnson and click on this book. For instance, the hours of sleep are in C2T4.mtw, indicating Table 4 of Chapter 2. To open this worksheet:

Under **File** choose **Open Worksheet** and go to the drive that contains the downloaded files.
Click on the MINITAB folder and then click on the file name C2T4. Click **OK.**

This will activate the worksheet in MINITAB and you do not have to manually enter the numbers.



```
Welcome to Minitab, press F1 for help.
```

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
	Sleep										
1	4.5										
2	4.7										
3	5.0										
4	5.0										
5	5.3										
6	5.5										
7	5.5										
8	5.7										
9	5.7										
10	5.7										
11	6.0										
12	6.0										
13	6.0										

The summary statistics can be obtained by pulling down the menu in the top bar under **Stat,** then choosing **Basic Statistics** and then **Graphic Summary.** More specifically,

Data in C1

Dialog box:
Stat > Basic Statistics > Graphical Summary.
Type *C1* in **Variables.** Click **OK.**

MINITAB uses a slightly different definition of the quartiles and their values may slightly differ from those calculated by the method in this book.

MINITAB will also create a histogram. With the data in *C1*,

Dialog box:

Graph > Histogram.
Select *Simple* in **Variables.** Click **OK.**
Type *C1* in **X.** Click **OK.**

MINITAB will also create boxplots, dot plots, and stem-and-leaf displays. With the data in *C1*,

Dialog box:

Graph > Dotplot.
Select *Simple.* Click **OK.**
Type *C1* in **Graph variables.** Click **OK.**

produces a dot plot. To obtain a boxplot, replace the first step by

Graph > Boxplot.

and to obtain a stem-and-leaf display, replace the first two steps by the single step

Graph > Stem-and-Leaf.

Clicking on **Labels** before the last **OK** will allow you to put titles on your graph.

EXCEL

Begin with the data in column A. For the hours of sleep from Table 4, the spread sheet is given on page 73.

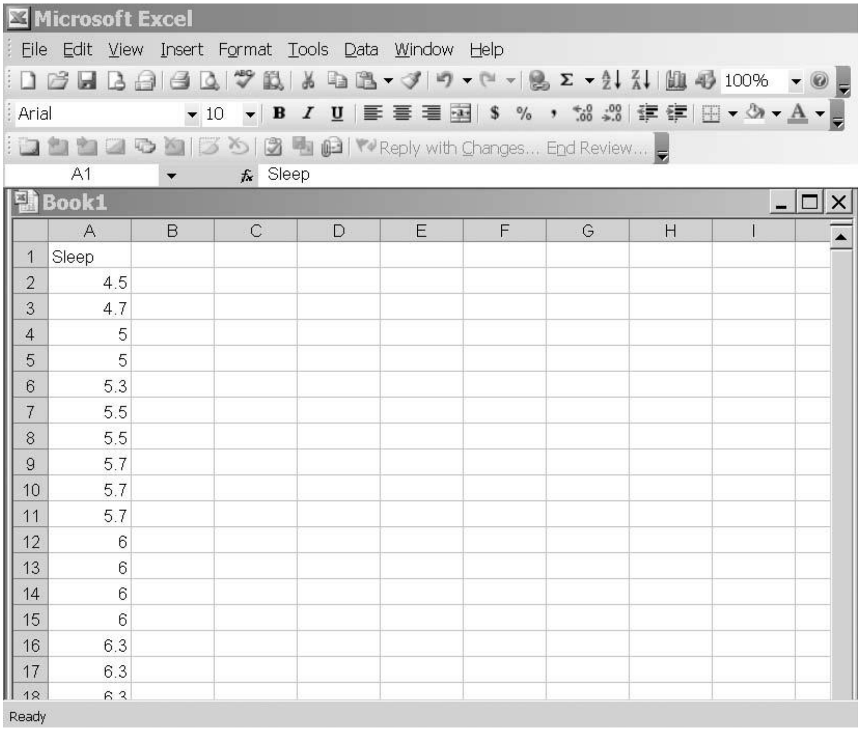
Alternatively, the data sets in the book are stored in EXCEL workbooks on the book's Web site. Go to www.wiley.com/college/johnson and click on this book. For instance, the hours of sleep are in C2T4.xls, indicating Table 4 of Chapter 2. Go to the drive containing the downloaded files and click on the EXCEL folder and then on the file name C2T4. The workbook having the data from Table 4 will then open.

Most of the statistical procedures we will use start with

Select **Tools** and then **Data Analysis.**
If **Data Analysis** is not listed, then it must be added once. To do so, select **Tools** then **Add-Ins.** Check **Analysis Toolpak** and click **OK.**

To obtain the summary statistics,

Select **Tools,** then **Data Analysis,** and then **Descriptive Statistics.**
Click **OK.** Place cursor in the **Input Range** window and use the mouse to highlight the data in column A. Check **Summary Statistics** and click **OK.**


Microsoft Excel
File Edit View Insert Format Tools Data Window Help
100%
Arial
10
Reply with Changes... End Review...
Book1
Ready

A1 Sleep

	A
1	Sleep
2	4.5
3	4.7
4	5
5	5
6	5.3
7	5.5
8	5.5
9	5.7
10	5.7
11	5.7
12	6
13	6
14	6
15	6
16	6.3
17	6.3
18	6.3

TI-84/83 PLUS

Press **STAT,** select **1:Edit,** and then enter the data in **L1**.
Press **STAT,** highlight **Calc,** and select **1:1–Var Stats.**
With **1:1–Var Stats** in the Home screen, press **2nd 1:1–Var Stats 1** to insert **L1** on the Home screen.
Then press **ENTER.**

9. REVIEW EXERCISES

2.99 Recorded here are the numbers of civilians employed in the United States by major occupation groups for the years 2000 and 2007. (*Source:* Statistical Abstract of the United States, 2009.)

	Number of Workers in Millions	
	2007	2000
Goods producing	22.2	24.6
Service (private)	115.4	107.1
Government	22.2	20.8
Total	159.8	152.5

(a) For each year, calculate the relative frequencies of the occupation groups.

(b) Comment on changes in the occupation pattern between 2000 and 2007.

2.100 Table 13 gives data collected from the students attending an elementary statistics course at the University of Wisconsin. These data include sex, height, number of years in college, and the general area of intended major [Humanities (H); Social Science (S); Biological Science (B); Physical Science (P)].

TABLE 13 Class Data

Student No.	Sex	Height in Inches	Year in College	Intended Major
1	F	67	3	S
2	M	72	3	P
3	M	70	4	S
4	M	70	1	B
5	F	61	4	P
6	F	66	3	B
7	M	71	3	H
8	M	67	4	B
9	M	65	3	S
10	F	67	3	B
11	M	74	3	H
12	M	68	3	S
13	M	74	2	P
14	F	64	4	P
15	M	69	3	S
16	M	64	3	B
17	M	72	4	P
18	M	71	3	B
19	F	67	2	S
20	M	70	4	S
21	M	66	4	S
22	F	67	2	B
23	M	68	4	S
24	M	71	3	H
25	M	75	1	S
26	M	67	1	B
27	M	68	3	P
28	M	72	4	B
29	F	68	3	P
30	F	66	2	B
31	F	65	2	B
32	M	64	4	B
33	M	72	1	H
34	M	67	4	B
35	M	73	3	S
36	F	71	4	B
37	M	71	3	B
38	M	69	2	S
39	F	69	4	P
40	M	74	4	S
41	M	73	3	B
42	M	68	3	B
43	F	66	2	S
44	M	73	2	P
45	M	73	2	S
46	M	67	4	S
47	F	62	3	S
48	M	68	2	B
49	M	71	3	S

(a) Summarize the data of "intended major" in a frequency table.

(b) Summarize the data of "year in college" in a frequency table and draw either a line diagram or a histogram.

2.101 Referring to Exercise 2.100, plot the dot diagrams of heights separately for the male and female students and compare.

2.102 Refer to the data on power outages in Table D.1 in the Data Bank. Make a Pareto chart for the cause of the outage.

2.103 The dollar amounts claimed by businessmen for their lunches are to be grouped into the following classes: 0–5, 5–10, 10–15, 15–20, 20 or more. The left endpoint is included. Is it possible to determine from this frequency distribution the exact number of lunches for which the amount claimed was:

(a) Less than 15?

(b) 10 or more?

(c) 30 or more?

2.104 Mung bean sprouts are more widely used in Asian cooking than the beans themselves. To study their growth, an experimenter presoaked some beans until they sprouted about 1 millimeter. Five were randomly selected and placed in individual petri dishes. After 96 hours, their lengths (mm)

143 131 101 143 111

were obtained. Find the mean and standard deviation.

2.105 The weights of twenty adult grizzly bears captured and released are summarized in the computer output

```
Descriptive Statistics: bearwt
Variable  N   Mean   Median   StDev
Bearwt   20  227.4    232.5    82.7
```

(a) Locate two measures of center tendency, or location, and interpret the values.

(b) Locate the standard deviation.

(c) Calculate the *z* score for a grizzly bear that weighs 320 pounds. See Exercise 2.82.

2.106 The stem-and-leaf display given here shows the final examination scores of students in a sociology course. (Leaf unit = 1.0)

Stem-and-Leaf Display of Scores	
2	57
3	244
4	1179
5	03368
6	012447
7	223556899
8	00457
9	0036

(a) Find the median score.

(b) Find the quartiles Q_1 and Q_3.

(c) What proportion of the students scored below 70? 80 and over?

2.107 The following are the numbers of passengers on the minibus tour of Hollywood.

9 12 10 11 11 7 12 6 11 4 10 10 11 9 10
7 10 8 8 9 8 9 11 9 8 6 10 6 8 11

(a) Find the sample median.

(b) Find the sample mean.

(c) Find the sample variance.

2.108 The following table shows the age at inauguration of each U.S. president.

(a) Make a stem-and-leaf display with a double stem.

(b) Find the median, Q_1 and Q_3.

Name	Age at Inauguration
1. Washington	57
2. J. Adams	61
3. Jefferson	57
4. Madison	57
5. Monroe	58
6. J. Q. Adams	57
7. Jackson	61
8. Van Buren	54
9. W. H. Harrison	68
10. Tyler	51
11. Polk	49
12. Taylor	64
13. Fillmore	50
14. Pierce	48
15. Buchanan	65
16. Lincoln	52
17. A. Johnson	56
18. Grant	46
19. Hayes	54
20. Garfield	49
21. Arthur	50
22. Cleveland	47
23. B. Harrison	55
24. Cleveland	55
25. McKinley	54
26. T. Roosevelt	42
27. Taft	51
28. Wilson	56
29. Harding	55
30. Coolidge	51
31. Hoover	54
32. F. D. Roosevelt	51
33. Truman	60
34. Eisenhower	62
35. Kennedy	43
36. L. Johnson	55
37. Nixon	56
38. Ford	61
39. Carter	52
40. Reagan	69
41. G. Bush	64
42. Clinton	46
43. G. W. Bush	54
44. Obama	47

2.109 (a) Calculate $\bar{x}$ and s for the data 6, 8, 4, 9, 8.

(b) Consider the data set 106, 108, 104, 109, 108, which is obtained by adding 100 to each number given in part (a). Use your results of part (a) and the properties stated in Exercises 2.52 and 2.74 to obtain the $\bar{x}$ and s for this modified data set. Verify your results by direct calculations with this new data set.

(c) Consider the data set −18, −24, −12, −27, −24, which is obtained by multiplying each number of part (a) by −3. Repeat the problem given in part (b) for this new data set.

2.110 Refer to the class data in Exercise 2.100. Calculate the following.

(a) $\bar{x}$ and s for the heights of males.

(b) $\bar{x}$ and s for the heights of females.

(c) Median and the quartiles for the heights of males.

(d) Median and the quartiles for the heights of females.

2.111 In a genetic study, a regular food was placed in each of 20 vials and the number of flies of a particular genotype feeding on each vial recorded. The counts of flies were also recorded for another set of 20 vials that contained grape juice. The following data sets were obtained (courtesy of C. Denniston and J. Mitchell).

No. of Flies (Regular Food)

15	20	31	16	22	22	23	33	38	28
25	20	21	23	29	26	40	20	19	31

No. of Flies (Grape Juice)

6	19	0	2	11	12	13	12	5	16
2	7	13	20	18	19	19	9	9	9

(a) Plot separate dot diagrams for the two data sets.

(b) Make a visual comparison of the two distributions with respect to their centers and spreads.

(c) Calculate $\bar{x}$ and s for each data set.

2.112 The data below were obtained from a detailed record of purchases of toothpaste over several years (courtesy of A. Banerjee). The usage times (in weeks) per ounce of toothpaste for a household taken from a consumer panel were

.74	.45	.80	.95	.84	.82	.78	.82	.89	.75	.76	.81
.85	.75	.89	.76	.89	.99	.71	.77	.55	.85	.77	.87

(a) Plot a dot diagram of the data.

(b) Find the relative frequency of the usage times that do not exceed .80.

(c) Calculate the mean and the standard deviation.

(d) Calculate the median and the quartiles.

2.113 To study how first-grade students utilize their time when assigned to a math task, a researcher observes 24 students and records their times off-task out of 20 minutes. The dotplot appears on page 3. (courtesy of T. Romberg).

Times Off-Task (minutes)					
4	0	2	2	4	1
4	6	9	7	2	7
5	4	13	7	7	10
10	0	5	3	9	8

For this data set, find:

(a) Mean and standard deviation.

(b) Median.

(c) Range.

2.114 The following summary statistics were obtained from a data set.

$\bar{x} = 80.5$ Median $= 84.0$

$s = 10.5$ $Q_1 = 75.5$

$Q_3 = 96.0$

Approximately what proportion of the observations are:

(a) Below 96.0?

(b) Above 84.0?

(c) In the interval 59.5 to 101.5?

(d) In the interval 75.5 to 96.0?

(e) In the interval 49.0 to 112.0?

State which of your answers are based on the assumption of a bell-shaped distribution.

2.115 The 50 measurements of acid rain in Wisconsin, whose histogram is given on the cover page of the chapter, are

3.58	3.80	4.01	4.01	4.05	4.05
4.12	4.18	4.20	4.21	4.27	4.28
4.30	4.32	4.33	4.35	4.35	4.41
4.42	4.45	4.45	4.50	4.50	4.50
4.50	4.51	4.52	4.52	4.52	4.57
4.58	4.60	4.61	4.61	4.62	4.62
4.65	4.70	4.70	4.70	4.70	4.72
4.78	4.78	4.80	5.07	5.20	5.26
5.41	5.48				

(a) Calculate the median and quartiles.

(b) Find the 90th percentile.

(c) Determine the mean and standard deviation.

(d) Display the data in the form of a boxplot.

2.116 Refer to Exercise 2.115.

(a) Determine the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$.

(b) What proportions of the measurements lie in those intervals?

(c) Compare your findings with the empirical guidelines for bell-shaped distributions.

2.117 Refer to the earthquake size data in Exercise 2.20.

(a) Calculate the median and quartiles.

(b) Calculate the mean and standard deviation.

(c) Display the data in the form of a boxplot.

2.118 The Dow Jones average provides an indication of overall market level. The changes in this average, from year end, from 1969 to 1970 through 2007 to 2008 are summarized in the following frequency table, where the left-hand endpoint is excluded.

Yearly Changes in the Dow Jones Average

Change in DJ Average	Frequency
(−4500, −1600]	1
(−1600, −850]	1
(−850, −250]	2
(−250, 0]	8
(0, 250]	13
(250, 850]	6
(850, 1650]	5
(1650, 2450]	3
Total	39

(a) Calculate the relative frequency for the intervals.

(b) Plot the relative frequency histogram. (*Hint:* Since the intervals have unequal widths, make the height of each rectangle equal to the relative frequency divided by the width of the interval.)

(c) What proportion of the changes were negative?

(d) Comment on the location and shape of the distribution.

2.119 The winning times of the men's 400-meter freestyle swimming in the Olympics (1964 to 2008) appear in the following table.

Winning Times in Minutes and Seconds

Year	Time
1964	4:12.2
1968	4:09.0
1972	4:00.27
1976	3:51.93
1980	3:51.31
1984	3:51.23
1988	3:46.95
1992	3:45.00
1996	3:47.97
2000	3:40.59
2004	3:43.10
2008	3:41.86

(a) Draw a dot diagram and label the points according to time order.

(b) Explain why it is not reasonable to group the data into a frequency distribution.

2.120 The **mode** of a collection of observations is defined as the observed value with largest relative frequency. The mode is sometimes used as a center value. There can be more than one mode in a data set. Find the mode for the data given in Exercise 2.13.

2.121 Lightning causes many deaths each year in the United States. The yearly number of deaths for 50 years, 1959 through 2008 are, by rows,

183	129	149	153	165	129	149	110	88	129	131	122	122
94	124	102	91	74	98	88	63	74	66	77	77	67
74	68	88	68	67	74	73	41	43	69	85	53	42
44	46	51	44	51	43	32	38	47	45	27		

Obtain the mean and standard deviation.

2.122 With reference to the lightning data in Exercise 2.121,

(a) Make a time plot of the data.

(b) Comment on the appropriateness of presenting the mean and standard deviation as summaries.

The Following Exercises Require a Computer.

Calculations of the descriptive statistics such as $\overline{x}$ and s are increasingly tedious with larger data sets. Current computer software programs alleviate the drudgery of hand calculations. Use MINITAB or some other package program.

2.123 Find $\overline{x}$ and s for:

(a) The lizard data in Exercise 2.19.

(b) The acid rain data in Exercise 2.115.

2.124 Lumber intended for building houses and other structures must be monitored for strength. The measurement of strength (pounds per square inch) for 61 specimens of Southern Pine (*Source:* U.S. Forest Products Laboratory) yielded

4001	3927	3048	4298	4000	3445
4949	3530	3075	4012	3797	3550
4027	3571	3738	5157	3598	4749
4263	3894	4262	4232	3852	4256
3271	4315	3078	3607	3889	3147
3421	3531	3987	4120	4349	4071
3686	3332	3285	3739	3544	
4103	3401	3601	3717	4846	
5005	3991	2866	3561	4003	
4387	3510	2884	3819	3173	
3470	3340	3214	3670	3694	

Using MINITAB, the sequence of choices produces a rather complete summary of the data. It includes the output on page 79.

Data(in 2.126.txt):
Strength
Dialog box:
Stat > Basic Statistics > Graphical Summary.
Type *Strength* in **Variables.** *Click* **OK.**

(a) Use this output to identify a departure from a bell-shaped pattern.

(b) MINITAB uses a slightly different scheme to determine the first and third quartiles, but the difference is not of practical importance with large samples. Calculate the first quartile using the definition in this book and compare with the value in the output.

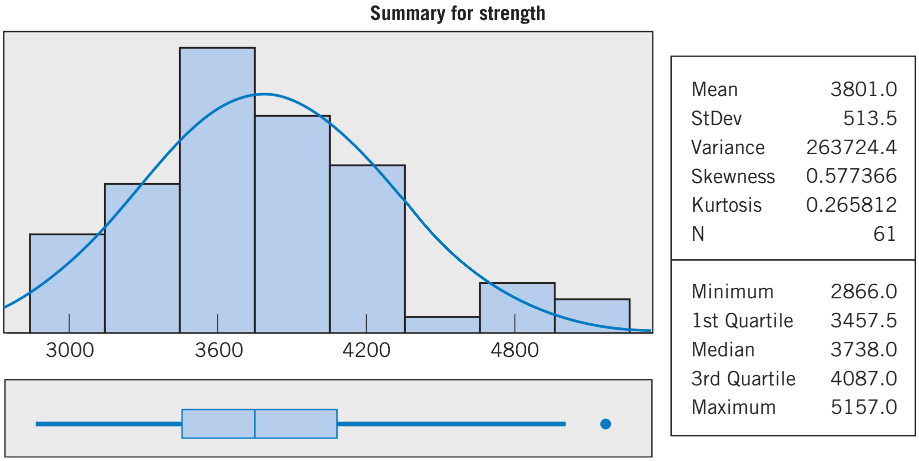
2.125 Find $\overline{x}$ and s for the data set in Table 4.

2.126 Find $\overline{x}$ and s for the final times to run 1.5 miles in Table D.5 in the Data Bank.

2.127 The SAS computer software package produced the output on page 79. Compare the mean and standard deviation with that of the MINITAB output in Exercise 2.124. Which output gives more digits?

2.128 The salmon fisheries support a primary industry in Alaska and their management is of high priority. Salmon are born in freshwater rivers and streams but then swim out into the ocean for a few years before returning to spawn and die. In order to identify the origins of mature fish, researchers studied growth rings on their scales. The growth the first year in freshwater is measured by the width of the growth rings for that period of life. The growth ring for the first year in the ocean environment will give an indication of growth for that period. A set of these measurements are given in Table D.7 in the Data Bank.

(a) Describe the freshwater growth for males by making a histogram and calculating the mean, standard deviation, and quartiles.



```
Mean            3801.0
StDev            513.5
Variance      263724.4
Skewness      0.577366
Kurtosis      0.265812
N                   61

Minimum         2866.0
1st Quartile    3457.5
Median          3738.0
3rd Quartile    4087.0
Maximum         5157.0
```

Figure 20 MINITAB output for Exercise 2.124.

```
                        UNIVARIATE PROCEDURE
VARIABLE = X1
                              MOMENTS
          N           61          VARIANCE    263724.4
          MEAN        3800.951
          STD DEV     513.5411
                        QUANTILES (DEF = 5)
                100% MAX    5157     99%    5157
                 75% Q3     4071     95%    4846
                 50% MED    3738     90%    4349
                 25% Q1     3470     10%    3173
                  0% MIN    2866      5%    3075
                                      1%    2866
                 RANGE      2291
                Q3-Q1        601
                              EXTREMES
              LOWEST     OBS        HIGHEST     OBS
                2866(       31)        4749(       28)
                2884(       42)        4846(       22)
                3048(        3)        4949(       12)
                3075(       14)        5005(       29)
                3078(       47)        5157(       26)
```

Figure 21 SAS output for Exercise 2.125.

(b) Describe the freshwater growth for females by making a histogram and calculating the mean, standard deviation, and quartiles.

(c) Construct boxplots to compare the growth of males and females.

2.129 Refer to the alligator data in Table D.11 of the Data Bank. Using the data on x_5 for thirty-seven alligators:

(a) Make a histogram.

(b) Obtain the sample mean and standard deviation.

2.130 Refer to Exercise 2.129.

(a) Obtain the quartiles.

(b) Obtain the 90th percentile. How many of the alligators above the 90th percentile are female?

2.131 Refer to the data on malt extract in Table D.8 of the Data Bank.

(a) Obtain sample mean and standard deviation.

(b) Obtain quartiles.

(c) Check conformity with the empirical rule.