MILLER AND FREUND'S

PROBABILITY AND STATISTICS FOR ENGINEERS

Richard Johnson

Department of Statistics

University of Wisconsin—Madison

Contents

Chapter 1

Introduction

Everything dealing with the collection, processing, analysis, and interpretation of numerical data belongs to the domain of statistics. In engineering, this includes such diversified tasks as calculating the average length of the downtimes of a computer, collecting and presenting data on the numbers of persons attending seminars on solar energy, evaluating the effectiveness of commercial products, predicting the reliability of a rocket, or studying the vibrations of airplane wings.

In Sections 1.2, 1.3, 1.4 and 1.5 we discuss the recent growth of statistics and, in particular, its applications to problems of engineering. Statistics plays a major role in the improvement of quality of any product or service. An engineer using the techniques described in this book can become much more effective in all phases of work relating to research, development, or production.

We begin our introduction to statistical concepts in Section 1.6 by emphasizing the distinction between a population and a sample.

1.1 Why Study Statistics?

Answers provided by statistical approaches can provide the basis for making decisions or choosing actions. For example, city officials might want to know whether the level of lead in the water supply is within safety standards. Because not all of the water can be checked, answers must be based on the partial information from samples of water that are collected for this purpose. As another example, a civil engineer must determine the strength of supports for generators at a power plant. A number of those available must be loaded to failure and their strengths will provide the basis for assessing the strength of other supports. The proportion of all supports available with strengths that lie below a design limit needs to be determined.

When information is sought, statistical ideas suggest a typical collection process with four crucial steps.

(a) **Set clearly defined goals for the investigations**.

(b) **Make a plan of what data to collect and how to collect it.**

(c) **Apply appropriate statistical methods to extract information from the data.**

(d) **Interpret the information and draw conclusions.**

These indispensable steps will provide a frame of reference throughout as we develop the key ideas of statistics. Statistical reasoning and methods can help you become efficient at obtaining information and making useful conclusions.

1.2 Modern Statistics

The origin of statistics can be traced to two areas of interest that, on the surface, have little in common: games of chance and what is now called political science. Mid-eighteenth century studies in probability, motivated largely by interest in games of chance, led to the mathematical treatment of errors of measurement and the theory that now forms the foundation of statistics. In the same century, interest in the numerical description of political units (cities, provinces, countries, etc.) led to what is now called **descriptive statistics**. At first, descriptive statistics consisted merely of the presentation of data in tables and charts; nowadays, it includes also the summarization of data by means of numerical descriptions and graphs.

In recent decades, the growth of statistics has made itself felt in almost every major phase of activity, and the most important feature of its growth has been the shift in emphasis from descriptive statistics to **statistical inference**. Statistical inference concerns generalization based on sample data; it applies to such problems as estimating an engine's average emission of pollutants from trial runs, testing a manufacturer's claim on the basis of measurements performed on samples of his product, and predicting the fidelity of an audio system on the basis of sample data pertaining to the performance of its components.

When one makes a statistical inference, namely, an inference that goes beyond the information contained in a set of data, one must always proceed with caution. One must decide carefully how far one can go in generalizing from a given set of data, whether such generalizations are at all reasonable or justifiable, whether it might be wise to wait until there are more data, and so forth. Indeed, some of the most important problems of statistical inference concern the appraisal of the risks and the consequences to which one might be exposed by making generalizations from sample data. This includes an appraisal of the probabilities of making wrong decisions, the chances of making incorrect predictions, and the possibility of obtaining estimates that do not lie within permissible limits.

We shall approach the subject of statistics as a science, developing each statistical idea insofar as possible from its probabilistic foundation, and applying each idea to problems of physical or engineering science as soon as it has been developed. The great majority of the methods we shall use in stating and solving these problems belong to the **classical approach**, because they do not formally take into account the various subjective factors mentioned above. However, we shall endeavor continually to make the reader aware that the subjective factors do exist, and to indicate whenever possible what role they might play in making the final decision. This "bread-and-butter" approach to statistics presents the subject in the form in which it has so successfully contributed to engineering science, as well as to the natural and social sciences, in the last half of the twentieth century and beyond.

1.3 Statistics and Engineering

There are few areas where the impact of the recent growth of statistics has been felt more strongly than in engineering and industrial management. Indeed, it would be difficult to overestimate the contributions statistics has made to solving production problems, to the effective use of materials and labor, to basic research, and to the development of new products. As in other sciences, statistics has become a vital tool to engineers. It enables them to understand phenomena subject to variation and to effectively predict or control them.

In this text, our attention will be directed largely toward engineering applications, but we shall not hesitate to refer also to other areas to impress upon the reader the great generality of most statistical techniques. Thus, the reader will find that the statistical method which is used to estimate the average coefficient of thermal expansion of a metal serves also to estimate the average time it takes a secretary to perform a given task, the average thickness of a pelican eggshell, or the average IQ of first year college students. Similarly, the statistical method that is used to compare the strength of two alloys serves also to compare the effectiveness of two teaching methods, the merits of two insect sprays, or the performance of men and women in a current-events test.

1.4 The Role of the Scientist and Engineer in Quality Improvement

Since the 1960's, the United States has found itself in an increasingly competitive world market. At present, we are in the midst of an international revolution in quality improvement. The teaching and ideas of W. Edwards Deming (1900-1993) were instrumental in the rejuvenation of Japan's industry. He stressed that American industry, in order to survive, must mobilize with a continuing commitment to quality improvement. From design to production, processes need to be continually improved. The engineer and scientist, with their technical knowledge and armed with basic statistical skills in data collection and graphical display, can be main participants in attaining this goal.

The **quality improvement** movement is based on the philosophy of "make it right the first time". Furthermore, one should not be content with any process or product but should continue to look for ways of improving it. We will emphasize the key statistical components of any modern quality improvement program. In Chapter 14, we outline the basic issues of quality improvement and present some of the specialized statistical techniques for studying production processes. The experimental designs discussed in Chapter 13 are also basic to the process of quality improvement.

Closely related to quality improvement techniques are the statistical techniques that have been developed to meet the **reliability** needs of the highly complex products of space-age technology. Chapter 15 provides an introduction to this area.

1.5 A Case Study : Visually Inspecting Data to Improve Product Quality

This study [1] dramatically illustrates the important advantages gained by appropriately plotting and then monitoring manufacturing data. It concerns a ceramic part used in popular coffee makers. This ceramic part is made by filling the cavity between two dies of a pressing machine with a mixture of clay, water and oil. After pressing, but before the part is dried to a hardened state, critical dimensions are measured. The depth of the slot is of interest here.

Because of natural uncontrolled variation in the clay-water-oil mixture, the condition of the press, differences in operators and so on, we cannot expect all of the slot measurements to be exactly the same. Some variation in the depth of slots is inevitable but the depth needs to be controlled within certain limits for the part to fit when assembled.

Slot depth was measured on three ceramic parts selected from production every half hour during the first shift from 6 A.M. to 3.P.M. The data in Table 1.1 were obtained on a Friday. The sample mean, or average, for the first sample of 214, 211 and 218 (thousandths of an inch) is

$$\frac{214+211+218}{3} = \frac{643}{3} = 214.3.$$

The graphical procedure, called an ***X*-bar** chart, consists of plotting the sample averages versus time order. This plot will indicate when changes have occurred and actions need to be taken to correct the process.

From a prior statistical study, it was known that the process was stable about a value of 217.5 thousandths of an inch. This value will be taken as the central line of the chart.

$$\text{central line : } \quad \overline{\overline{x}} = 217.5$$

[1] Courtesy of Don Ermer

It was further established that the process was capable of making mostly good ceramic parts if the average slot dimension for a sample remained between the

$$\text{Lower control limit: LCL} = 215.0$$

$$\text{Upper control limit: UCL} = 220.0$$

TABLE 1.1 Slot depth (thousandths of an inch)

Time	6.30	7.00	7.30	8.00	8.30	9.00	9.30	10.00
1	214	218	218	216	217	218	218	219
2	211	217	218	218	220	219	217	219
3	218	219	217	219	221	216	217	218
SUM	643	654	653	653	658	653	652	656
$\overline{x}$	214.3	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
$\overline{x}$	217.7	217.0	219.0	220.0	217.7	219.3	218.3	214.7

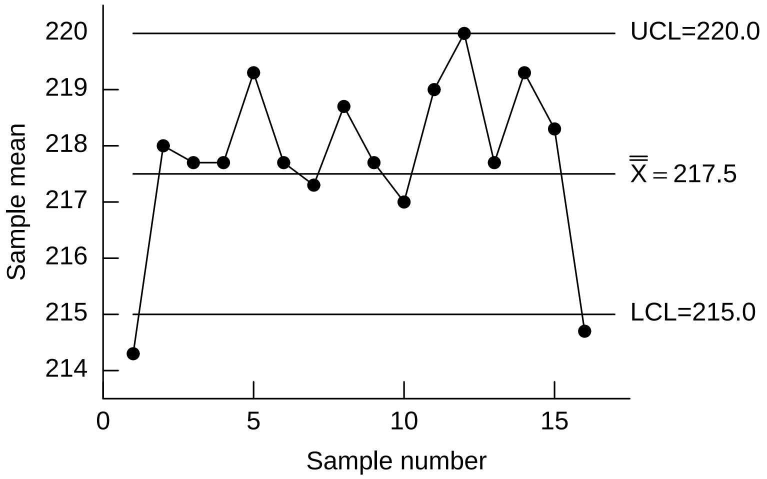


FIGURE 1.1 X - bar Chart for Depth

What does the chart tell us? The mean of 214.3 for the first sample, taken at approximately 6.30 A.M., is outside the lower control limit. Further, a measure of the variation in this sample

$$\text{range} = \text{largest} - \text{smallest} = 218 - 211 = 7$$

is large compared to the others. This evidence suggests that the pressing machine had not yet reached a steady state. The control chart suggests that it is necessary to warm up the pressing machine before the first shift begins at 6 A.M. Management and engineering implemented an early start-up and thereby improved the process. The operator and foreman did not have the authority to make this change. Deming claimed that 85% or more of our quality problems are in the system and that the operator and others responsible for the day-to-day operation are responsible for 15% or less of our quality problems.

The X-bar chart further shows that, throughout the day, the process was stable but a little on the high side although no points were out of control until the last sample of the day. Here an unfortunate oversight occurred. The operator did not report the out-of-control value to either the set-up person or the foreman because it was near the end of her shift and the start of her weekend. She also knew the set-up 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. She did not want to ruin anyone's plans so she kept quiet.

On Monday morning when the operator started up the pressing machine, one of the dies broke. The cost of the die was over a thousand dollars. But this was not the biggest cost. 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 expensive item. Deming referred to this type of cost as the unknown and unknowable, but at the same time it is probably the most important cost of poor quality.

On Friday the chart had predicted a problem. Afterward it was determined that the most likely difficulty was that the clay had dried and stuck to the die, leading to the break. The chart indicated the problem but someone had to act; for a statistical charting procedure to be truly effective action must be taken.

1.6 Two Basic Concepts - Population and Sample

The examples above, where the evaluation of actual information is essential for acquiring new knowledge, motivate the development of statistical reasoning and tools taught in this text. Most experiments and investigations conducted by engineers in the course of investigating, be it a physical phenomenon, production process, or manufactured unit, share some common characteristics.

A first step in any study is to develop a clear well defined **statement of purpose**. For example, a mechanical engineer wants to determine whether a new additive will increase the tensile strength of plastic parts produced on an injection molding machine. Not only must the additive increase the tensile strength, it needs to increase it by enough to be of engineering importance. He therefore created the following statement.

Purpose : Determine whether a particular amount of an additive can be found that will increase the tensile strength of the plastic parts by at least 10 pounds per square inch.

In any statement of purpose, try to avoid words like soft, hard, large enough, and so on which are difficult to quantify. The statement of purpose can help us to decide on what data to collect. For example, the mechanical engineer tried two different amounts of additive and produced 25 specimens of the plastic part with each mixture. The tensile strength was obtained for each of 50 specimens.

Relevant data must be collected. But it is often physically impossible or infeasible from a practical standpoint to obtain a complete set of data. When data are obtained from laboratory experiments, no matter how much experimentation has been performed, more could always be done. 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! In most situations, we must work with only partial information. The distinction between the data actually acquired and the vast collection of all potential observations is a key to understanding statistics. The source of each measurement is called a **unit**. It is usually an object or a person. To emphasize the term population, for the entire collection of units, we call the entire collection the **population of units**.

Units and population of units

unit: A single entity, usually an object or person, whose characteristics are of interest.

population of units: The complete collection of units about which information is sought.

Guided by the statement of purpose, we have a **characteristic of interest** for each unit in the population. The characteristic, which could be a qualitative trait, is called a **variable** if it can be expressed as a number.

There can be several characteristics of interest for a given population of units. Some examples are given in Table 1.2.

TABLE 1.2 Examples of populations, units, and variables.

Population	Unit	Variables/Characteristics
All students currently enrolled in school	student	GPA number of credits hours of work per week major right/left-handed
All printed circuit boards manufactured during a month	board	type of defects number of defects location of defects
All campus fast food restaurants	restaurant	number of employees seating capacity hiring / not hiring
All books in library	book	replacement cost frequency of check-out repairs needed

For any population there is the value, for each unit, of a characteristic or variable of interest. 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**. This collection of values is the population we will address in all later chapters. Here 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.

Statistical population

A statistical **population** is the set of all measurements (or record of some quality trait) corresponding to each unit in the entire population of units about which information is sought.

Generally, any statistical approach to learning about the population begins by taking a sample.

Samples from a population

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

The sample needs both to be representative of the population and to be large enough to contain sufficient information to answer the questions about the population that are crucial to the investigation.

EXAMPLE Self-selected samples – A bad practice

A magazine which features the latest computer hardware and software for home office use enclosed a short questionnaire on a postcard. Readers were asked to indicate whether or not they owned specific new software packages or hardware products. In past issues, this magazine used similar information from cards that were returned to make such statements as "40% of readers have purchased software package P." Is this sample representative of the population of magazine readers?

Solution It is clearly impossible to contact all magazine readers since not all are subscribers. One must necessarily settle for taking a sample. Unfortunately, the method used by the magazine editors is not representative and is badly biased. Readers who always update and try most of the new software will be more likely to respond indicating their purchases. In contrast, those who did not purchase any of the software or hardware mentioned in the survey will be very likely not to return the postcard. That is, the proportion of purchasers of software package P in the sample of returned postcards will likely be much higher than it is for the whole population consisting of the *purchase / not purchase* record for each reader.

■

To avoid bias due to self-selected samples, we must take an active role in the selection process. Random numbers can determine which. specific units to include in the sample of units.

Using a Random Number Table to Select Samples

The selection of a sample from a finite population must be done impartially and objectively. But, writing the unit names on slips of paper, putting the slips in a box, and drawing them out may not only be cumbersome, but proper mixing may not be possible. However, the selection is easy to carry out using a chance mechanism called a **random number table**. Suppose ten balls numbered $0, 1, \cdots, 9$ are placed in an urn and shuffled. Then one is drawn and the digit recorded. It is then replaced, the balls shuffled, another one drawn and the digit recorded. The digits in Table 7 at the end of the book were actually generated by a computer that closely simulates this procedure. A portion of this table is shown as Table 1.3.

The chance mechanism that generated the random number table ensures that each of the single digits has the same chance of occurrence, that all pairs $00, 01, \cdots, 99$ have the same chance of occurrence, and so on. Further, any collection of digits is unrelated to any other digit in the table. Because of these properties, the digits are called random.

TABLE 1.3 Random digits–a portion of Table 7 random digits.

1306	1189	5731	3968	5606	5084	8947	3897	1636	7810
0422	2431	0649	8085	5053	4722	6598	5044	9040	5121
6597	2022	6168	5060	8656	6733	6364	7649	1871	4328
7965	6541	5645	6243	7658	6903	9911	5740	7824	8520
7695	6937	0406	8894	0441	8135	9797	7285	5905	9539
5160	7851	8464	6789	3938	4197	6511	0407	9239	2232
2961	0551	0539	8288	7478	7565	5581	5771	5442	8761
1428	4183	4312	5445	4854	9157	9158	5218	1464	3634
3666	5642	4539	1561	7849	7520	2547	0756	1206	2033
6543	6799	7454	9052	6689	1946	2574	9386	0304	7945
9975	6080	7423	3175	9377	6951	6519	8287	8994	5532
4866	0956	7545	7723	8085	4948	2228	9583	4415	7065
8239	7068	6694	5168	3117	1568	0237	6160	9585	1133
8722	9191	3386	3443	0434	4586	4150	1224	6204	0937
1330	9120	8785	8382	2929	7089	3109	6742	2468	7025

EXAMPLE Using the table of random digits

Eighty specialty pumps were manufactured last week. Use Table 1.3 to select a sample of size $n = 5$ to carefully test and recheck for possible defects before they are sent to the purchaser. Select the sample without replacement so that the same pump does not appear twice in the sample.

Solution The first step is to number the pumps from 1 to 80, or to arrange them in some order so they can be identified. The digits must be selected two at a time because the population size $N = 80$ is a two-digit number. We begin by arbitrarily selecting a row and column. We select row 6 and column 21. Reading the digits in columns 21 and 22, and proceeding downward, we obtain

$$41 \quad 75 \quad 91 \quad 75 \quad 19 \quad 69 \quad 49.$$

We ignore the number 91 because it is greater than the population size 80. We also ignore any number when it appears a second time, as 75 does here. That is, we continue reading until five different numbers in the appropriate range are selected. Here the five pumps numbered

$$41 \quad 75 \quad 19 \quad 69 \quad 49$$

will be carefully tested and rechecked for defects.

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

■

EXAMPLE Selecting a sample by random digit dialing

Suppose there is a single three-digit exchange for the area in which you wish to conduct a survey. Use the random digit Table 7 to select five phone numbers.

Solution We arbitrarily decide to start on the second page of Table 7 at row 21 and column 13. Reading the digits in columns 13 through 16, and proceeding downward, we obtain

5619 0812 9167 3802 4449.

These five numbers, together with the designated exchange, become the phone numbers to be 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 number drawn from the table. If there are two exchanges in the area, separate selections could be done for each exchange.

■

Do's and Don'ts

Do's

1. Create a clear statement of purpose before deciding 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.

Don'ts

1. Don't unquestioningly accept self-selected samples.

REVIEW EXERCISES

1.1 A consumer magazine article asks "How Safe Is the Air in Airplanes?" and goes on to say that the air quality was measured on 158 different flights for U.S. based airlines. Let the variable of interest be a numerical measure of staleness. Identify the population and the sample.

1.2 A radio show host announced that she wanted to know which singer was the favorite among college students in your school. Listeners were asked to call and name their favorite singer. Identify the population, in terms of preferences, and the sample. Is the sample likely to be representative? Comment. Also describe how to obtain a sample that is likely to be more representative.

1.3 Consider the population of all laptop computers owned by students at your university. You want to know the size of the hard disk.

(a) Specify the population unit.

(b) Specify the variable of interest.

(c) Specify the statistical population.

1.4 Identify the statistical population, sample and the variable of interest in each of the following situations.

(a) To learn about starting salaries for engineers graduating from a Midwest university, twenty graduating seniors are asked to report their starting salary.

(b) Fifty computer memory chips were selected from the six thousand manufactured that day. The fifty computer memory chips were tested and 5 were found to be defective.

(c) Tensile strength was measured on 20 specimens made of a new plastic material. The intent is to learn about the tensile strengths for all specimens that could conceivably be manufactured with the new plastic material.

1.5 A campus engineering club has 40 active members listed on its membership roll. Use Table 7 of random digits to select 5 persons to be interviewed regarding the time they devote to club activities each week.

1.6 A city runs 50 buses daily. Use Table 7 of random digits to select 4 buses to inspect for cleanliness. (We started on the first page of Table 7 at row 31 columns 25 and 26 and read down).

1.7 Refer to the slot depth data in Table 1.1. After the machine was repaired, a sample of three new ceramic parts had slot depths 215, 216 and 213 (thousandths of an inch).

(a) Redraw the X-bar chart and include the additional mean $\overline{x}$.

(b) Does the new $\overline{x}$ fall within the control limits?

1.8 A Canadian manufacturer identified a critical diameter on a crank bore that needed to be maintained within a close tolerance for the product to be successful. Samples of size 4 were taken every hour. The values of the differences (measurement - specification), in ten-thousandths of an inch, are given in Table 1.4.

(a) Calculate the central line for an X-bar chart for the 24 hourly sample means. The centerline is $\overline{\overline{x}} = (4.25 - 3.00 - \cdots - 1.50 + 3.25)/24$.

(b) Is the average of all the numbers in the table, 4 for each hour, the same as the average of the 24 hourly averages? Should it be?

(c) A computer calculation gives the control limits

LCL = − 4.48
UCL = 7.88

Construct the X-bar chart. Identify hours where the process was out of control.

TABLE 1.4 The differences (measurement − specification), in ten-thousandths of an inch.

Hour	1	2	3	4	5	6	7	8	9	10	11	12
	10	-6	−1	−8	-14	−6	−1	8	−1	5	2	5
	3	1	−3	−3	−5	−2	−6	−3	7	6	1	3
	6	−4	0	−7	−6	−1	−1	9	1	3	1	10
	−2	−3	−7	−2	2	−6	7	11	7	2	4	4
$\overline{x}$	4.25	−3.00	−2.75	−5.00	−5.75	−3.75	−0.25	6.25	3.50	4.00	2.00	5.50

Hour	13	14	15	16	17	18	19	20	21	22	23	24
	5	6	−5	−8	2	7	8	5	8	−5	−2	−1
	9	6	4	−5	8	7	13	4	1	7	−4	5
	9	8	−5	1	−4	5	6	7	0	1	−7	9
	7	10	−2	0	1	3	6	10	−6	2	7	0
$\overline{x}$	7.50	7.50	−2.00	−3.00	1.75	5.50	8.25	6.50	0.75	1.25	−1.50	3.25

KEY TERMS : (*with page references*)

Classical approach to statistics **??**
Descriptive statistics **??**
Population **??**
Population of units **??**
Quality improvement **??**
Random number table **??**
Reliability **??**
Sample **??**
Statement of purpose **??**
Statistical inference **??**
X-bar chart **??**
Unit **??**

Chapter 2

Treatment of Data

Statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are virtually useless unless they are condensed, or reduced, into a more suitable form. We begin with the use of simple graphics. Next, Sections 2.2 and 2.3 deal with problems relating to the grouping of data and the presentation of such groupings in graphical form; in Section 2.4 we discuss a relatively new way of presenting data.

Sometimes it may be satisfactory to present data just as they are and let them speak for themselves; on other occasions it may be necessary only to group the data and present the result in tabular or graphical form. However, most of the time data have to be summarized further, and in Sections 2.5 through 2.7 we introduce some of the most widely used kinds of statistical descriptions.

2.1 Pareto Diagrams and Dot Diagrams

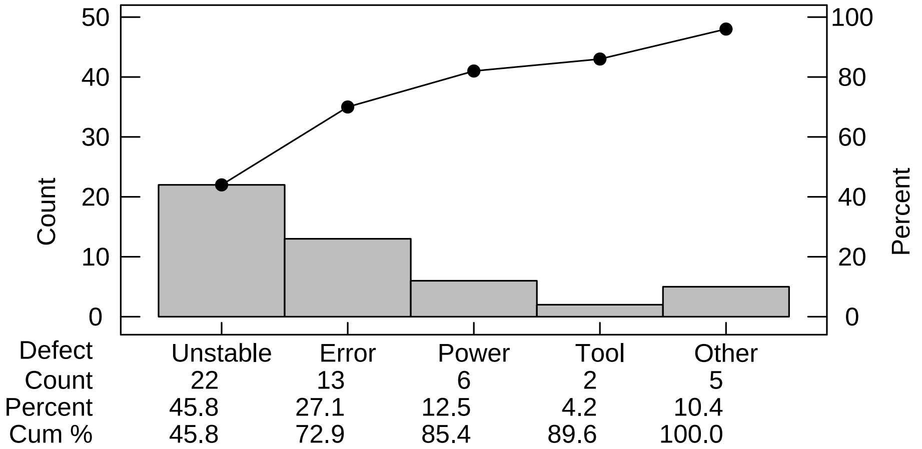
Data need to be collected to provide the vital information necessary to solve engineering problems. Once gathered, these data must be described and analyzed to produce summary information. Graphical presentations can often be the most effective way to communicate this information. To illustrate the power of graphical techniques, we first describe a **Pareto diagram**. This display, which orders each type of failure or defect according to its frequency, can help engineers identify important defects and their causes.

When a company identifies a process as a candidate for improvement, the first step is to collect data on the frequency of each type of failure. For example, for a computer-controlled lathe whose performance was below par, workers recorded the following causes and their frequencies:

power fluctuations	6
controller not stable	22
operator error	13
worn tool not replaced	2
other	5

These data are presented as a special case of a **bar chart** called a **Pareto diagram** in Figure 2.1. This diagram graphically depicts Pareto's empirical law that any assortment of events consists of a few major and many minor elements. Typically, two or three elements will account for more than half of the total frequency.

Concerning the lathe, 22 or $100\,(22/48) = 46\%$ of the cases are due to an unstable controller and $22{+}13 = 35$ or $100\,(35/48) = 73\%$ are due to either unstable controller or operator error. These cumulative percentages are shown in Figure 2.1 as a line graph whose scale is on the right-hand side of the Pareto diagram as in Figure 14.2.



Defect	Unstable	Error	Power	Tool	Other
Count	22	13	6	2	5
Percent	45.8	27.1	12.5	4.2	10.4
Cum %	45.8	72.9	85.4	89.6	100.0

FIGURE 2.1 A Pareto diagram of failures

In the context of quality improvement, to make the most impact we want to select the few vital major opportunities for improvement. This graph visually emphasizes the importance of reducing the frequency of controller misbehavior. An initial goal may be to cut it in half.

As a second step toward improvement of the process, data were collected on the deviations of cutting speed from the target value set by the controller. The seven observed values of (cutting speed) − (target),

$$3 \quad 6 \quad -2 \quad 4 \quad 7 \quad 4 \quad 3$$

are plotted as a **dot diagram** in Figure 2.2. The dot diagram visually summarizes the information that the lathe is, generally, running fast. In Chapters 13 and 14 we will develop efficient experimental designs and methods for identifying primary causal factors that contribute to the variability in a response such as cutting speed.

When the number of observations is small, it is often difficult to identify any pattern of variation. Still, it is a good idea to plot the data and look for unusual features.

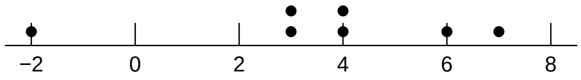


FIGURE 2.2 Dot diagram of cutting speed deviations

EXAMPLE Dot diagrams expose outliers

In 1987, for the first time, physicists observed neutrinos from a supernova that occurred outside of our solar system. At a site in Kamiokande, Japan, the following times (second) between neutrinos were recorded:

0.107 0.196 0.021 0.283 0.179 0.854 0.58 0.19 7.3 1.18 2.0

Draw a dot diagram.

Solution We plot to the nearest 0.1 second to avoid crowding. (See Figure 2.3). Note the extremely long gap between 2.0 and 7.3 seconds. Statisticians call such an unusual observation an **outlier**. Usually, outliers merit further attention. Was there a recording error, were neutrinos missed in that long time interval, or were there two separate explosions in the supernova? Important questions in physics may hinge on the correct interpretation of this outlier.

■

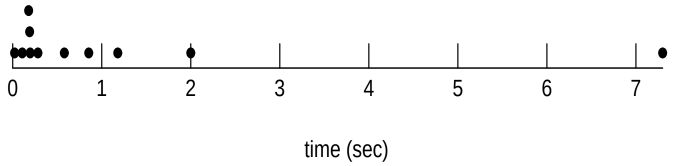


FIGURE 2.3 Dot diagram of time between neutrinos

EXAMPLE A dot diagram for multiple samples reveals differences

The vessels that contain the reactions at some nuclear power plants consist of two hemispherical components that are welded together. Copper in the welds could cause them to become brittle after years of service. Samples of welding material from one production run or "heat" that were used in one plant had the copper contents 0.27, 0.35, 0.37. Samples from the next heat had values 0.23, 0.15, 0.25, 0.24, 0.30, 0.33, 0.26. Draw a dot diagram that highlights possible differences in the two production runs (heats) of welding material. If the copper contents for the two runs are different, they should not be combined to form a single estimate.

Solution We plot the first group as solid circles and the second as open circles.(See Figure 2.4.) It seems unlikely that the two production runs are alike because the top two values are from the first run. (In Chapter 10 we confirm this fact). The two runs should be treated separately.

The copper content of the welding material used at the power plant is directly related to the determination of safe operating life. Combining the sample would lead to an unrealistically low estimate of copper content and too long an estimate of safe life.

■

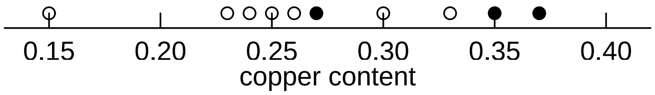


FIGURE 2.4 Dot diagram of copper content

When a set of data consists of a large number of observations, we take the approach in the next section. The observations are first summarized in the form of a table.

2.2 Frequency Distributions

A **frequency distribution** is a table that divides a set of data into a suitable number of classes (categories), showing also the number of items belonging to each class. Such a table sacrifices some of the information contained in the data; instead of knowing the exact value of each item, we only know that it belongs to a certain class. On the other hand, this kind of grouping often brings out important features of the data, and the gain in "legibility" usually more than compensates for the loss of information. In what follows, we shall consider mainly **numerical distributions**, that is, frequency distributions where the data are grouped according to size; if the data are grouped according to some quality, or attribute, we refer to such a distribution as a **categorical distribution**.

The first step in constructing a frequency distribution consists of deciding how many classes to use and the choosing the **class limits** for each class. That is, deciding from where to where each class is to go. Generally speaking, the number of classes we use depends on the number of observations, but it is seldom profitable to use fewer than 5 or more than 15. The exception to the upper limit is when data the size of the data set is several hundred or even a few thousand. It also depends on the range of a the data, namely, the difference between the largest observation and the smallest. Then, we tally the observations and thus determine the **class frequencies**, namely, the number of observations in each class.

To illustrate the construction of a frequency distribution, let us consider the following 80 determinations of the daily emission (in tons) of sulfur oxides from an industrial plant:

15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0	13.2
22.7	9.8	6.2	14.7	17.5	26.1	12.8	28.6	17.6	23.7
26.8	22.7	18.0	20.5	11.0	20.9	15.5	19.4	16.7	10.7
19.1	15.2	22.9	26.6	20.4	21.4	19.2	21.6	16.9	19.0
18.5	23.0	24.6	20.1	16.2	18.0	7.7	13.5	23.5	14.5
14.4	29.6	19.4	17.0	20.8	24.3	22.5	24.6	18.4	18.1
8.3	21.9	12.3	22.3	13.3	11.8	19.3	20.0	25.7	31.8
25.9	10.5	15.9	27.5	18.1	17.9	9.4	24.1	20.1	28.5

Since the largest observation is 31.8, the smallest is 6.2, and the range is 25.6, we might choose the six classes having the limits 5.0−9.9, 10.0−14.9, $\cdots$, 30.0−34.9, we might choose the seven classes 5.0−8.9, 9.0−12.9, $\cdots$ 29.0−32.9, or we might choose the nine classes 5.0−7.9, 8.0−10.9, $\cdots$, 29.0−31.9. Note that in each case **the classes do not overlap, they accommodate all the data, and they are all of the same width**.

Initially deciding on the second of these classifications, we now tally the 80 observations and obtain the results shown in the following table:

Class limits	Frequency
5.0 – 8.9	3
9.0 – 12.9	10
13.0 – 16.9	14
17.0 – 20.9	25
21.0 – 24.9	17
25.0 – 28.9	9
29.0 – 32.9	2
Total	80

Note that the class limits are given to as many decimal places as the original data. Had the original data been given to two decimal places, we would have used the class limits 5.00−8.99, 9.00−12.99, $\cdots$, 29.00−32.99, and if they had been rounded to the nearest ton, we would have used the class limits 5−8, 9−12, $\cdots$, 29−32.

In the preceding example, the data may be thought of as values of a continuous variable which could, conceivably, be any value in an interval. But, if we use classes such as 5.0−9.0, 9.0−13.0, $\cdots$, 29.0−33.0, there exists the possibility of ambiguities: 9.0 could go into the first class or into the second, 13.0 could go into the second class or into the third and so on. To avoid this difficulty, we take an alternative approach that is particularly applicable when graphing frequency distributions.

We make an **endpoint convention**. For the emission data we could take [5, 9) as the first class, [9, 11) as the second, and so on through [29, 33). That is, for this data set, we adopt the convention that the left-hand endpoint is included but the right-hand endpoint is not. For other data sets, we may prefer to reverse the endpoint convention so the right-hand endpoint is included but the left-hand endpoint is not. Whichever endpoint convention is adopted, it should appear in the description of the frequency distribution.

Using the convention that the left-endpoint is included, the frequency table for the sulfur emissions data is

Class limits	Frequency
[5.0, 9.0)	3
[9.0, 13.0)	10
[13.0, 17.0)	14
[17.0, 21.0)	25
[21.0, 25.0)	17
[25.0, 29.0)	9
[29.0, 33.0)	2
Total	80

The **class boundaries** are the endpoints of the intervals that specify each class. As we pointed out earlier, once data have been grouped, each observation has lost its identity in the sense that its exact value is no longer known. This may lead to difficulties when we want to give further descriptions of the data, but we can avoid them by representing each observation in a class by its midpoint, called the **class mark**. In general, the class marks of a frequency distribution are obtained by averaging successive class limits or successive class boundaries. If the classes of a distribution are all of equal length, as in our example, we refer to the common interval between any successive class marks as the **class interval** of the distribution. Note that the class interval may also be obtained from the difference between any successive class boundaries.

EXAMPLE Class marks and class interval for grouped data

With reference to the distribution of the sulfur oxide emission data, find (a) the class marks and (b) the class interval.

Solution

(a) The class marks are $\frac{5.0+9.0}{2} = 7.0, \frac{9.0+13.0}{2} = 11.0, 15.0, 19.0, 23.0, 27.0,$ and 31.0.

(b) The class interval is $11.0 - 7.0 = 4$.

There are several alternative forms of distributions into which data are sometimes grouped. Foremost among these are the "less than", "or less," "more than," and "or more" **cumulative distributions**. A cumulative "less than" distribution shows the total number of observations that are less than given values. These values must be class boundaries or appropriate class limits, but they may not be class marks.

■

EXAMPLE Cumulative distribution for sulfur emission data

Convert the distribution of the sulfur oxides emission data into a distribution showing how many of the observations are less than 5.0 less than 9.0 less than 13.0, $\cdots$, and less than 33.0.

Solution Since none of the values is less than 5.0, 3 are less than 9.0, 3+10 = 13 are less than 13.0, 3+10+14 = 27 are less than 17.0, $\cdots$, and all 80 are less than 33.0, we have

Tons of sulfur oxides	Cumulative Frequency
less than 5.0	0
less than 9.0	3
less than 13.0	13
less than 17.0	27
less than 21.0	52
less than 25.0	69
less than 29.0	78
less than 33.0	80

Cumulative "more than" and "or more" distributions are constructed similarly by adding the frequencies, one by one, starting at the other end of the frequency distribution. In practice, "less than" cumulative distributions are used most widely, and it is not uncommon to refer to “less than” cumulative distributions simply as cumulative distributions.

If it is desirable to compare frequency distributions, it may be necessary (or at least advantageous) to convert them into **percentage distributions**. We simply divide each class frequency by the total frequency (the total number of observations in the distribution) and multiply by 100; in this way we indicate what percentage of the data falls into each class of the distribution. The same can also be done with cumulative distributions, thus converting them to **cumulative percentage distributions**.

▬

2.3 Graphs of Frequency Distributions

Properties of frequency distributions relating to their shape are best exhibited through the use of graphs, and in this section we shall introduce some of the most widely used forms of graphical presentations of frequency distributions and cumulative distributions.

The most common form of graphical presentation of a frequency distribution is the **histogram**. The histogram of a frequency distribution is constructed of adjacent rectangles; the heights of the rectangles represent the class frequencies and the bases of the rectangles extend between successive class boundaries. A histogram of the sulfur oxides emission data is shown in Figure 2.5.

Using our endpoint convention, the interval (5.9] that defines the first class has frequency 3 so the rectangle has height 3. The second rectangle, over the interval (9, 13], has height 10 and so on. The tallest rectangle is over the interval (17, 21] and it has height 25. The histogram has a single peak and it is reasonably symmetric. About half of the area, representing half of the observations, is over the interval from 15 to 23 tons of sulfur oxides.

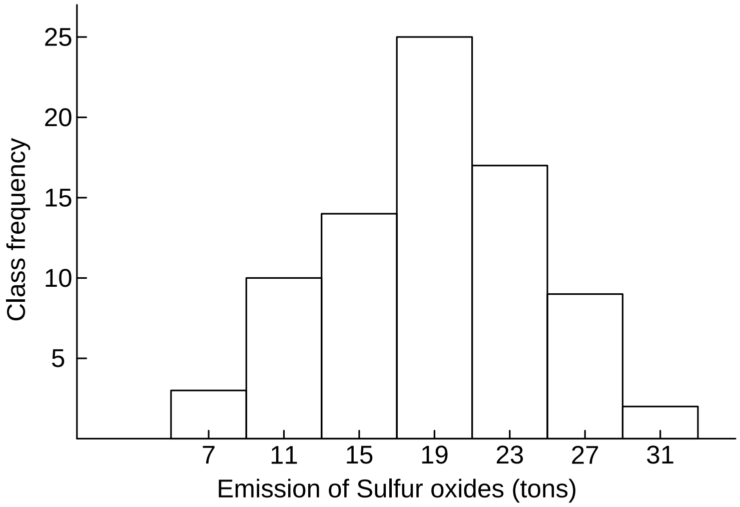


FIGURE 2.5 Histogram

Inspection of the graph of a frequency distribution as a histogram often brings out features that are not immediately apparent from the data themselves. Aside from the fact that such a graph presents a good overall picture of the data, it can also emphasize irregularities and unusual features. For instance, outlying observations which somehow do not fit the overall picture, that is, the overall pattern of variation in the data, may be due to errors of measurement, equipment failure and similar causes. Also, the fact that a histogram exhibits two or more *peaks* (maxima) can provide pertinent information. The appearance of two peaks may imply, for example, a shift in the process that is being measured, or it may imply that the data come from two or more sources. With some experience one learns to spot such irregularities or anomalies, and an experienced engineer would find it just as surprising if the histogram of a distribution of integrated-circuit failure times were symmetrical as if a distribution of American men's hat sizes were bimodal.

Sometimes it can be enough to draw a histogram in order to solve an engineering problem.

EXAMPLE A histogram reveals the solution to a grinding operation problem

A metallurgical engineer was experiencing trouble with a grinding operation. The grinding action was produced by pellets. After some thought he collected a sample of pellets used for grinding, took them home, spread them out on his kitchen table, and measured their diameters with a ruler. His histogram is displayed in Figure 2.6. What does the histogram reveal?

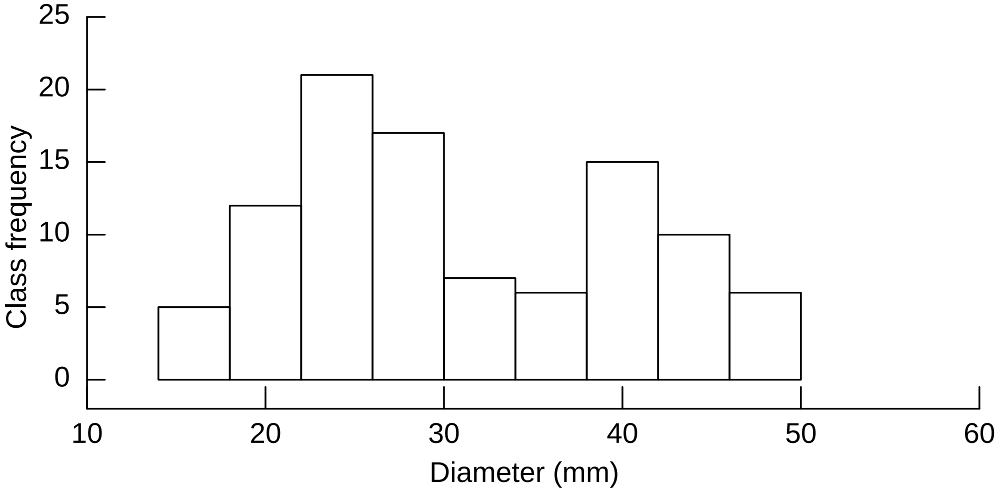


FIGURE 2.6 Histogram of pellet diameter

Solution The histogram exhibits two distinct peaks, one for a group of pellets whose diameters are centered near 25 and the other centered near 40.

By getting his supplier to do a better sort, so all the pellets would be essentially from the first group, the engineer completely solved his problem. Taking the action to obtain the data was the big step. The analysis was simple. ■

As illustrated by the next example concerning a system of supercomputers, not all histograms are symmetric.

EXAMPLE A histogram reveals the pattern of a supercomputer systems data

A computer scientist, trying to optimize system performance, collected data on the time, in microseconds, between requests for a particular process service.

2,808	4,201	3,848	9,112	2,082	5,913	1,620	6,719	21,657
3,072	2,949	11,768	4,731	14,211	1,583	9,853	78,811	6,655
1,803	7,012	1,892	4,227	6,583	15,147	4,740	8,528	10,563
43,003	16,723	2,613	26,463	34,867	4,191	4,030	2,472	28,840
24,487	14,001	15,241	1,643	5,732	5,419	28,608	2,487	995
3,116	29,508	11,440	28,336	3,440				

Draw a histogram using the equal length classes [0, 10,000), [10,000, 20,000), $\cdots$, [70,000, 80,000) where the left-hand endpoint is included but not the right-hand endpoint is not.

■

Solution The histogram of this interrequest time data, shown in Figure 2.7 has a long right hand tail. Notice that two classes are empty with this choice of equal length intervals. To emphasize that it is still possible to observe interrequest times in these intervals, it is preferable to regroup the data in the right-hand tail into classes of unequal lengths (see Exercise 2.62).

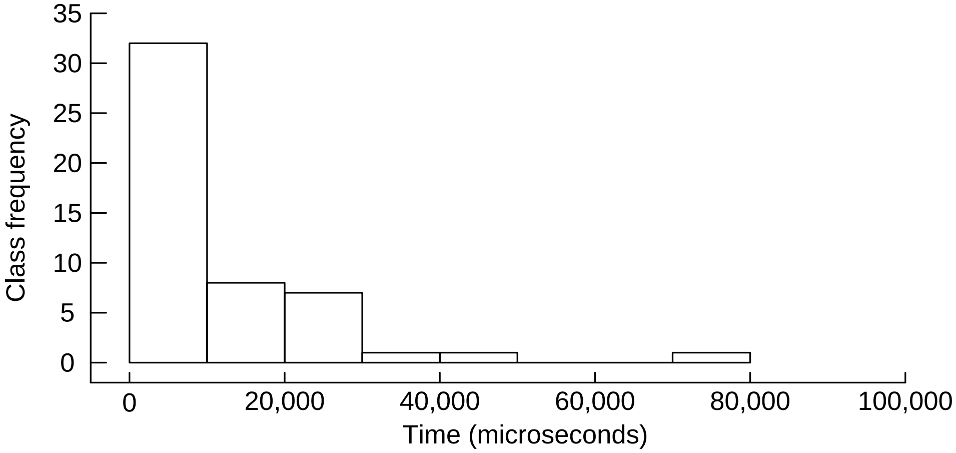


FIGURE 2.7 Histogram of interrequest time

When a histogram is constructed from a frequency table having classes of unequal lengths, the height of each rectangle must be changed to height $= \dfrac{\text{relative frequency}}{\text{width}}$. The area of the rectangle then represents the relative frequency for the class and the total area of the histogram is 1. We call this a **density histogram**.

EXAMPLE A density histogram has total area 1

Compressive strength was measured on 58 specimens of a new aluminum alloy undergoing development as a material for the next generation of aircraft.

66.4	67.7	68.0	68.0	68.3	68.4	68.6	68.8	68.9	69.0	69.1
69.2	69.3	69.3	69.5	69.5	69.6	69.7	69.8	69.8	69.9	70.0
70.0	70.1	70.2	70.3	70.3	70.4	70.5	70.6	70.6	70.8	70.9
71.0	71.1	71.2	71.3	71.3	71.5	71.6	71.6	71.7	71.8	71.8
71.9	72.1	72.2	72.3	72.4	72.6	72.7	72.9	73.1	73.3	73.5
74.2	74.5	75.3								

Draw a density histogram, that is, a histogram scaled to have a total area of 1 unit. For reasons to become apparent in Chapter 6, we call the vertical scale density.

Solution We make the height of each rectangle equal to *relative frequency /width*, so that its area equals the relative frequency. The resulting histogram, constructed by computer, has a nearly symmetric shape. (See Figure 2.8). We have also graphed a continuous curve that approximates the overall shape. In Chapter 6, we will be introduced to this bell-shaped family of curves.

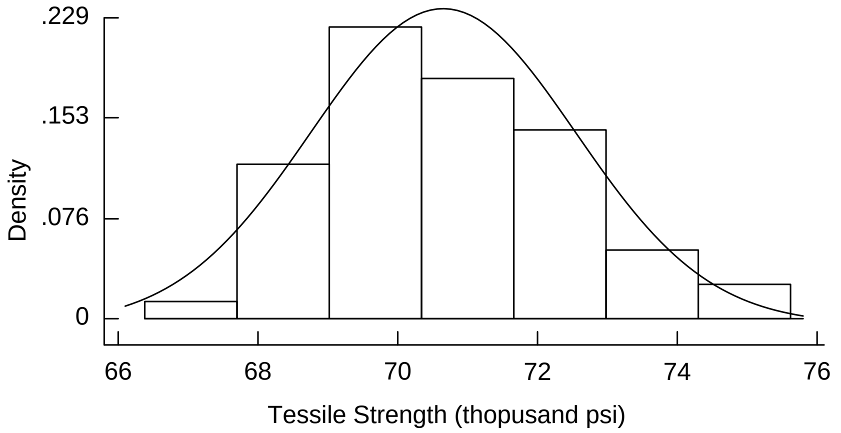


FIGURE 2.8 Histogram of aluminum alloy tensile strength

This example suggests that histograms, for observations that come from a continuous scale, can be approximated by smooth curves.

Cumulative distributions are usually presented graphically in the form of **ogives**, where we plot the cumulative frequencies at the class boundaries. The resulting points are connected by means of straight lines, as shown in Figure 2.9, which represents the cumulative "less than" distribution of the sulfur oxides emission data on page 15. The curve is steepest over the class with highest frequency.

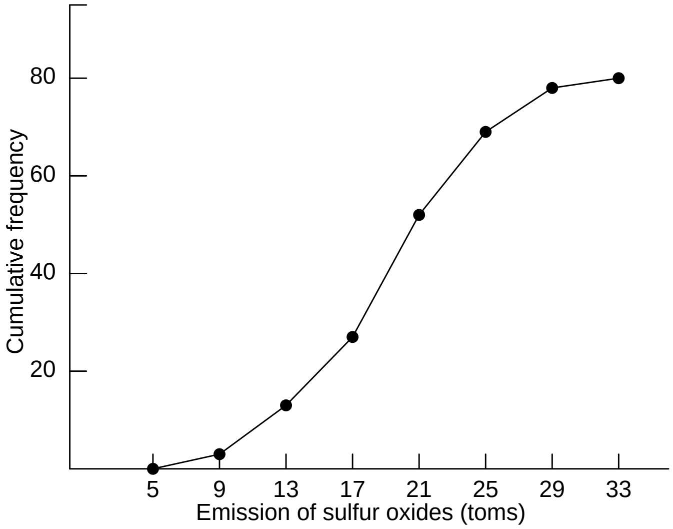


FIGURE 2.9 Ogive

2.4 Stem-and-leaf Displays

In the two preceding sections we directed our attention to the grouping of relatively large sets of data with the objective of putting such data into a manageable form. As we saw, this entailed some loss of information. Similar techniques have been proposed for the preliminary explorations of small sets of data, which yield a good overall picture of the data without any loss of information.

To illustrate, consider the following humidity readings rounded to the nearest percent:

29	44	12	53	21	34	39	25	48	23
17	24	27	32	34	15	42	21	28	37

Proceeding as in Section 2.2, we might group these data into the following distribution:

Humidity readings	Frequency
10− 19	3
20− 29	8
30− 39	5
40− 49	3
50− 59	1

If we wanted to avoid the loss of information inherent in the preceding table, we could keep track of the last digits of the readings within each class, getting

10−19	2 7 5
20−29	9 1 5 3 4 7 1 8
30−39	4 9 2 4 7
40−49	4 8 2
50−59	3

This can also be written as

1	2 7 5		1	2 5 7
2	9 1 5 3 4 7 1 8		2	1 1 3 4 5 7 8 9
3	4 9 2 4 7	or	3	2 4 4 7 9
4	4 8 2		4	2 4 8
5	3		5	3

where the left-hand column gives the ten digits 10, 20, 30, 40, and 50. In the last step, the leaves, are written in ascending order. The three numbers in the first row are 12, 15 and 17. This table is called a **stem-and-leaf display** (or simply a **stem-leaf display**) − each row has a **stem** and each digit on a stem to the right of the vertical line is a **leaf**. To the left of the vertical line are the **stem labels**, which, in our example, are 1, 2, . . . , 5. There should not be any gaps in the stem even if there are no leaves for that particular value.

Essentially, a stem-and-leaf display presents the same picture as the corresponding tally, yet it retains all the original information. For instance, if a stem-and-leaf display has the two-digit stem

$$1.2 \mid 0\ 2\ 3\ 5\ 8$$

the corresponding data are 1.20, 1.22, 1.23, 1.25 and 1.28, and if a stem-and-leaf display has the stem

$$0.3 \mid 03\ 17\ 55\ 89$$

with two-digit leaves, the corresponding data are 0.303, 0.317, 0.355 and 0.389.

There are various ways in which stem-and-leaf displays can be modified to meet particular needs (see Exercises 2.25 and 2.26), but we shall not go into this here in any detail as it has been our objective to present only one of the relatively new techniques, which come under the general heading of **exploratory data analysis**.

EXERCISES

2.1 Accidents at a potato chip plant are categorized according to the area injured.

fingers	17
eyes	5
arm	2
leg	1

Draw a Pareto chart.

2.2 Damages at a paper mill (thousands of dollars) due to breakage can be divided according to the product:

toilet paper	132
hand towels	85
napkins	43
12 other products	50

(a) Draw a Pareto chart.

What percent of the loss occurs in making

(b) toilet paper?

(c) toilet paper or hand towels?

2.3 The following are 15 measurements of the boiling point of a silicon compound (in degrees Celsius): 166, 141, 136, 153, 170, 162, 155, 146, 183, 157, 148, 132, 160, 175 and 150. Construct a dot diagram.

2.4 The following are 14 measurements on the strength (points) of paper to be used in cardboard tubes: 121, 128, 129, 132, 135, 133, 127, 115, 131, 125, 118, 114, 120, 116. Construct a dot diagram.

2.5 Civil engineers help municipal wastewater treatment plants operate more efficiently by collecting data on quality of the effluent. On seven occasions, the amounts of suspended solids (parts per million) at one plant were

14 12 21 28 30 65 26

Display the data in a dot diagram. Comment on your findings.

2.6 Jump River Electric serves part of Northern Wisconsin and because much of the area is forested, it is prone to outages. One August there were 11 power outages. Their durations(in hours) are

2.5 2.0 1.5 3.0 1.0 1.5 2.0 1.5 1.0 10.0 1.0

Display the data in a dot diagram.

2.7 The weights of certain mineral specimens given to the nearest tenth of an ounce are grouped to a table having the classes [10.5, 11.5), [11.5, 12.5), [12.5, 13.5) and [13.5, 14.5) ounces where the left-hand endpoint is included but the right-hand endpoint is not. Find

(a) the class marks.

(b) the class interval.

2.8 With reference to the preceding exercise, is it possible to determine from the grouped data how many of the mineral specimens weigh

(a) less than 11.5 ounces

(b) more than 11.5 ounces;

(c) at least 12.4 ounces;

(d) at most 12.4 ounces;

(e) from 11.5 to 13.5 ounces inclusive?

2.9 The following are measurements of the breaking strength (in ounces) of a sample of 60 linen threads:

32.5	15.2	35.4	21.3	28.4	26.9	34.6	29.3	24.5	31.0
21.2	28.3	27.1	25.0	32.7	29.5	30.2	23.9	23.0	26.4
27.3	33.7	29.4	21.9	29.3	17.3	29.0	36.8	29.2	23.5
20.6	29.5	21.8	37.5	33.5	29.6	26.8	28.7	34.8	18.6
25.4	34.1	27.5	29.6	22.2	22.7	31.3	33.2	37.0	28.3
36.9	24.6	28.9	24.8	28.1	25.4	34.5	23.6	38.4	24.0

Group these measurements into a distribution having the classes 15.0−19.9, 20.0−24.9, . . ., 35.0−39.9 and construct a histogram using [15, 20), [20, 25), $\cdots$, [35, 40) where the left-hand endpoint is included but the right-hand endpoint is not.

2.10 Convert the distribution obtained in the preceding exercise into a cumulative "less than" distribution and graph its ogive.

2.11 The class marks of a distribution of temperature readings (given to the nearest degree Celsius) are 16, 25, 34, 43, 52 and 61. Find

(a) the class boundaries;

(b) the class interval.

2.12 The following are the ignition times of certain upholstery materials exposed to a flame (given to the nearest hundredth of a second):

2.58	2.51	4.04	6.43	1.58	4.32	2.20	4.19
4.79	6.20	1.52	1.38	3.87	4.54	5.12	5.15
5.50	5.92	4.56	2.46	6.90	1.47	2.11	2.32
6.75	5.84	8.80	7.40	4.72	3.62	2.46	8.75
2.65	7.86	4.71	6.25	9.45	12.80	1.42	1.92
7.60	8.79	5.92	9.65	5.09	4.11	6.37	5.40
11.25	3.90	5.33	8.64	7.41	7.95	10.60	3.81
3.78	3.75	3.10	6.43	1.70	6.40	3.24	1.79
4.90	3.49	6.77	5.62	9.70	5.11	4.50	2.50
5.21	1.76	9.20	1.20	6.85	2.80	7.35	11.75

Group these figures into a table with a suitable number of equal classes and construct a histogram.

2.13 Convert the distribution obtained in Exercise 2.12 into a cumulative "less than" distribution and plot its ogive.

2.14 In a 2-week study of the productivity of workers, the following data were obtained on the total number of acceptable pieces which 100 workers produced:

65	36	49	84	79	56	28	43	67	36
43	78	37	40	68	72	55	62	22	82
88	50	60	56	57	46	39	57	73	65
59	48	76	74	70	51	40	75	56	45
35	62	52	63	32	80	64	53	74	34
76	60	48	55	51	54	45	44	35	51
21	35	61	45	33	61	77	60	85	68
45	53	34	67	42	69	52	68	52	47
63	65	55	61	73	50	53	59	41	54
41	74	82	58	26	35	47	50	38	70

Group these figures into a distribution having the classes 20-29, 30-39, 40-49, $\cdots$, and 80-89, and plot a histogram using [20, 30), $\cdots$, [80, 90) where the left-hand endpoint is included but the right-hand endpoint is not.

2.15 Convert the distribution obtained in Exercise 2.14 into a cumulative “less than” distribution and plots its ogive.

2.16 The following are the number of automobile accidents that occurred at 60 major intersections in a certain city during the Fourth of July weekend:

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

Group these data into a frequency distribution showing how often each of the values occurs and draw a bar chart.

2.17 Convert the distribution obtained in Exercise 2.16 into a cumulative “or more” distribution and draw its ogive.

2.18 Categorical distributions are often presented graphically by means of **pie charts**, in which a circle is divided into sectors proportional in size to the frequencies (or percentages) with which the data are distributed among the categories. Draw a pie chart to represent the following data, obtained in a study in which 40 drivers were asked to judge the maneuverability of a certain make of car.

Very good, good, good, fair, excellent, good, good, good, very good, poor, good, good, good, good, very good, good, fair, good, good, very poor, very good, fair good, good, excellent, very good, good good, good, fair, fair, very good, good, very good, excellent, very good, fair good, good and very good.

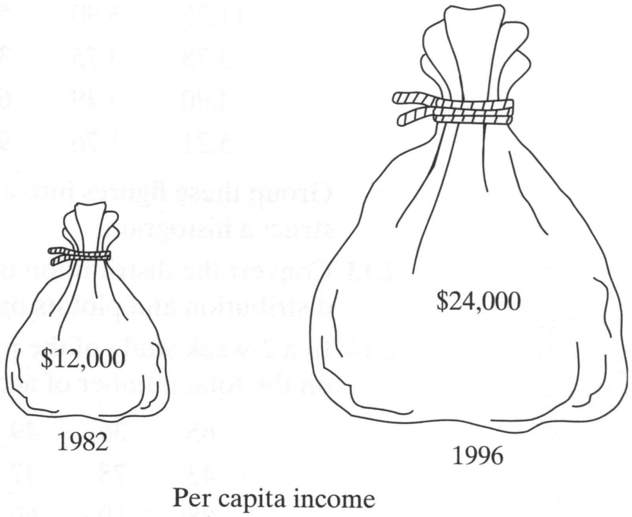


FIGURE 2.10 Pictogram for Exercise 2.19.

2.19 The pictogram of Figure 2.10 is intended to illustrate the fact that per capita income in the United States doubled from \$12,000 in 1982 to \$24,000 in 1996. Does this pictogram convey a “fair” impression of the actual change? If not, state how it might be modified.

2.20 Convert the distribution of the sulfur oxides emission data on page **??** into a distribution having the classes [5.0, 9.0), [9.0, 21.0), [21.0, 29.0), and [29.0, 33.0) where the left-endpoint is included. Draw two histograms of this distribution, one in which the class frequencies are given by the heights of the rectangles and one in which the class frequencies are given by the area of the rectangles. Explain why the first of these histograms gives a very misleading picture.

2.21 Given a set of observations $x_1, x_2, \ldots$, and x_n, we define their **empirical cumulative distribution** as the function whose values $F(x)$ equal the proportion of the observations less than or equal to x. Graph the empirical cumulative distribution for the 15 measurements of Exercise 2.3.

2.22 The following are figures on a well’s daily production of oil in barrels: 214, 203, 226, 198, 243, 225, 207, 203, 208, 200, 217, 202, 208, 212, 205 and 220. Construct a stem-and-leaf display with the stem labels 19, 20,. . . , and 24.

2.23 The following are determinations of a river's annual maximum flow in cubic meters per second: 405, 355, 419, 267, 370, 391, 612, 383, 434, 462, 288, 317, 540, 295, and 508. Construct a stem-and-leaf display with two-digit leaves.

2.24 List the data that correspond to the following stems of stem-and-leaf displays:

(a) 1 | 1 2 3 4 5 7 8. Leaf unit = 1.0.

(b) 23 | 0 0 1 4 6 . Leaf unit = 1.0.

(c) 2 | 03 18 35 57 First leaf unit = 10.0

(d) 3.2 | 1 3 4 4 7 . Leaf unit = 0.01

2.25 If we want to construct a stem-and-leaf display with more stems than there would be otherwise, we might repeat each stem. The leaves 0, 1, 2, 3 and 4 would be attached to the first stem and leaves 5, 6, 7, 8 and 9 to the second. For the humidity readings on page **??**, we would thus get the **double-stem display**.

1	2
1	5 7
2	1 1 3 4
2	5 7 8 9
3	2 4 4
3	7 9
4	2 4
4	8
5	3

where we doubled the number of stems by cutting the interval covered by each stem in half. Construct a double-stem display with one-digit leaves for the data in Exercise 2.14.

2.26 If the double stem display has too few stems, we might wish to have 5 stems where the first holds leaves 0 and 1, the second holds 2 and 3, and so on. The resulting stem-and-leaf display is called a **five-stem display**.

(a) The following are the IQs of 20 applicants to an undergraduate engineering program: 109, 111, 106, 106, 125, 108, 115, 109, 107, 109, 108, 110, 112, 104, 110, 112, 128, 106, 111 and 108. Construct a five stem display with one-digit leaves.

(b) The following is part of a five-stem display:

53	4 4 4 4 5 5	Leaf unit = 1.0
53	6 6 6 7	
53	8 9	
54	1	

List the corresponding measurements.

2.5 Descriptive Measures

Histograms, dot diagrams, and stem-and-leaf diagrams summarize a data set pictorially so we can visually discern the overall pattern of variation. We now develop numerical measures to describe a data set. To proceed, we introduce the notation

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

for a general sample consisting of n measurements. Here x_i is the i-th observation in the list so x_1 represents the value of the first measurement, x_2 represents the value of the second measurement, and so on.

Given a set of n measurements or observations, $x_1, x_2, \ldots, x_n$, there are many ways in which we can describe their center (middle, or central location). Most popular among these are the **arithmetic mean** and the **median**, although other kinds of “averages” are sometimes used for special purposes. The arithmetic mean – or, more succinctly, the **mean** – is defined by the formula
Sample mean

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

To emphasize that it is based on a set of observations, we often refer to $\overline{x}$ as the **sample mean**.

Sometimes it is preferable to use the **median** as a descriptive measure of the center, or location, of a set of data. This is particularly true if it is desired to minimize the calculations or if it is desired to eliminate the effect of extreme (very large or very small) values. The median of n observations $x_1, x_2, \ldots, x_n$ can be defined loosely as the “middlemost” value once the data are arranged according to size. More precisely, if the observations are arranged according to size and n is an odd number, the median is the value of the observation numbered $\frac{n+1}{2}$; if n is an even number, the median is defined as the mean (average) of the observations numbered $\frac{n}{2}$ and $\frac{n+2}{2}$.

Sample median

Order the n observations from smallest to largest

$$\begin{aligned}\text{sample median} \;&=\; \text{observation in position } \frac{n+1}{2}\;, \quad \text{if } n \text{ odd.}\\ &=\; \text{average of two observations in}\\ &\;\; \text{positions } \frac{n}{2} \text{ and } \frac{n+2}{2}\;, \qquad \text{if } n \text{ even.}\end{aligned}$$

EXAMPLE Calculation of the sample mean and median

In order to control costs, a company collects data on the weekly number of meals claimed on expense accounts. The numbers for five weeks are

15 14 2 7 and 13.

Find the mean and the median.

Solution The mean is

$$\overline{x} = \frac{15+14+2+27+13}{5} = 14.2 \quad \text{meals}$$

and, ordering the data from smallest to largest

$$2 \quad 13 \quad \underbrace{14} \quad 15 \quad 27$$

the median is the third largest value, namely, 14 meals.

Both the mean and median give essentially the same central value. ■

EXAMPLE Calculation of the sample median with even sample size

An engineering group receives email requests for technical information from sales and service persons. The daily numbers for six days are

11 9 17 19 4 and 15.

Find the mean and the median.

Solution The mean is

$$\overline{x} = \frac{11+9+17+19+4+15}{6} = 12.5 \;\; \text{requests}$$

and, ordering the data from the smallest to largest

$$4 \quad 9 \quad \underbrace{11 \quad 15} \quad 17 \quad 19$$

the median, the mean of the third and fourth largest values, is 13 requests.

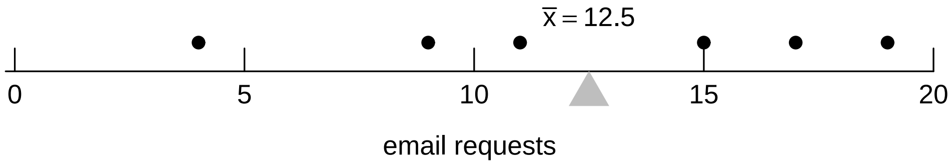


FIGURE 2.11 The interpretation of the sample mean as a balance point.

The sample mean has a physical interpretation as the balance point, or center of mass, of a data set. Figure 2.11 is the dot diagram for the data on the number of email requests given in the previous example. In the dot diagram, each observation is represented by a ball placed at the appropriate distance along the horizontal axis. If the balls are considered as masses having equal weights and the horizontal axis is weightless, then the mean corresponds to the center of inertia or balance point of the data. This interpretation of the sample mean, as the balance point of the observations, holds for any data set.

Although the mean and the median each provide a single number to represent an entire set of data, the mean is usually preferred in problems of estimation and other problems of statistical inference. An intuitive reason for preferring the mean is that the median does not utilize all the information contained in the observations.

The following is an example where the median actually gives a more useful description of a set of data than the mean.

EXAMPLE The median is unaffected by a few outliers

A small company employs four young engineers, who each earn $40,000, and the owner (also an engineer), who gets $130,000. Comment on the claim that on the average the company pays $58,000 to its engineers and, hence, is a good place to work.

■

Solution The mean of the five salaries is \$58,000, but it hardly describes the situation. The median, on the other hand, is \$40,000 and it is most representative of what a young engineer earns with the firm. Money wise, the company is not such a good place for young engineers.

This example illustrates that there is always an inherent danger when summarizing a set of data by means of a single number.

■

One of the most important characteristics of almost any set of data is that the values are not all alike; indeed, the extent to which they are unlike, or vary among themselves, is of basic importance in statistics. Measures such as the mean and median describe one important aspect of a set of data – their “middle” or their “average” – but they tell us nothing about this other basic characteristic. We observe that the dispersion of a set of data is small if the values are closely bunched about their mean, and that it is large if the values are scattered widely about their mean. It would seem reasonable, therefore, to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean.

If a set of numbers $x_1, x_2, \ldots, x_n$ has mean $\overline{x}$, the differences $x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_n - \overline{x}$ are called the **deviations from the mean**. It suggests itself that we might use their average as a measure of variation in the data set. Unfortunately, this will not do. For instance, refer to the observations 11 9 17 19 4 15 displayed above in Figure 2.11 where $\overline{x} = 12.5$ is the balance point. The six deviations are $-1.5 \quad -3.5 \quad 4.5 \quad 6.5 \quad -8.5 \quad 2.5$ and the sum of positive deviations $4.5 + 6.5 + 2.5 = 13.5$ exactly cancels the sum of the negative deviations $-1.5 \quad -3.5 \quad -8.5 = -13.5$ so the sum of all the deviations is 0.

As the reader will be asked to show in Exercise 2.50, the sum of the deviations is always zero. That is $\sum_{i=1}^{n} (\, x_i - \overline{x}\,) \;=\; 0$, so the mean of the deviations is always zero. Because the deviations sum to zero, we need to remove their signs. Absolute value and square are two natural choices. If we take their absolute value, so each negative deviation is treated as positive, we would obtain a measure of variation. However, to obtain the most common measure of variation, we square each deviation. The **sample variance**, s^2, is essentially the average of the squared deviations from the mean $\overline{x}$, and is defined by the formula

Sample variance

$$\boxed{\; s^2 = \frac{\sum\limits_{i=1}^{n} (\, x_i \,-\, \overline{x}\,)^2}{n-1} \;}$$

Our reason for dividing by $n-1$ instead of n is that there are only $n-1$ independent deviations $x_i - \overline{x}$. Because their sum is always zero, the value of any particular one is always equal to the negative of the sum of the other $n-1$ deviations. Also, using divisor $n - 1$ produces an estimate that will not, on average, lead to consistent overestimation or consistent underestimation.

If many of the deviations are large in magnitude, either positive or negative, their squares will be large and s^2 will be large. When all the deviations are small, s^2 will be small.

EXAMPLE Calculation of sample variance

The delay times (handling, setting, and positioning the tools) for cutting 6 parts on an engine lathe are 0.6 1.2 0.9 1.0 0.6 and 0.8 minutes. Calculate s^2.

Solution First we calculate the mean:

$$\overline{x} = \frac{0.6+1.2+0.9+1.0+0.6+0.8}{6} = 0.85$$

Then we set up the work required to find $\Sigma\,(\,x_i - \overline{x}\,)^2$ in the following table:

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
0.6	-0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	-0.25	0.0625
0.8	-0.05	0.0025
5.1	0.00	0.2750

We divide 0.2750 by 6−1 = 5 to obtain

$$s^2 = \frac{0.2750}{5} = 0.055 \ \ \text{(minute)}^2.$$

By calculating the sum of deviations in the second column, we obtain a check on our work. In other data sets, this sum should be 0 up to rounding error. ■

Notice that the units of s^2 are not those of the original observations. The data are delay times in minutes, but s^2 has the unit (minute)^2. Consequently, we define the **standard deviation** of n observations $x_1, x_2, \ldots, x_n$ as the square root of their variance, namely

Sample standard deviation

$$\boxed{s = \sqrt{\frac{\sum_{i=1}^{n} (\, x_i \, - \, \overline{x} \,)^2}{n-1}}}$$

The standard deviation is by far the most generally useful measure of variation. Its advantage over the variance is that it is expressed in the same units as the observations.

EXAMPLE Calculation of sample standard deviation

With reference to the previous example, calculate s.

Solution From the previous example, $s^2 = 0.055$. Take the square root and get

$$s \, = \, \sqrt{0.055} \, = \, 0.23 \ \text{ minutes}$$

■

The standard deviation s has a rough interpretation as the average distance from an observation to the sample mean.

The standard deviation and the variance are measures of **absolute variation**, that is, they measure the actual amount of variation in a set of data, and they depend on the scale of measurement. To compare the variation in several sets of data, it is generally desirable to use a measure of **relative variation**, for instance, the **coefficient of variation**, which gives the standard deviation as a percentage of the mean.

Coefficient of variation

$$\boxed{V = \frac{s}{\overline{x}} \cdot 100}$$

EXAMPLE The coefficient of variation for comparing relative preciseness

Measurements made with one micrometer of the diameter of a ball bearing have a mean of 3.92 mm and a standard deviation of 0.0152 mm, whereas measurements made with another micrometer of the unstretched length of a spring have a mean of 1.54 inches and a standard deviation of 0.0086 inch. Which of these two measuring instruments is relatively more precise?

Solution For the first micrometer the coefficient of variation is

$$V = \frac{0.0152}{3.92} \, \cdot \, 100 \, = \, 0.39\%$$

and for the second micrometer the coefficient of variation is

$$V = \frac{0.0086}{1.54} \, \cdot \, 100 \, = \, 0.56\%$$

Thus, the measurements made with the first micrometer are relatively more precise.

■

In this section, we have limited the discussion to the mean, the median, the variance, and the standard deviation, but there are many other ways of describing sets of data.

2.6 Quartiles and Percentiles

In addition to the median, which divides a set of data into halves, we can consider other division points. When an ordered data set is divided into quarters, the resulting division points are called sample **quartiles**. The **first quartile**, Q_1, is a value that has one-fourth, or 25%, of the observations below its value. The first quartile is also the sample 25th **percentile** $P_{0.25}$. More generally, we define the sample 100 p-th percentile as follows.

Sample percentiles

The sample 100 p-th percentile is a value such that 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.

As in the case of the median, which is the 50th percentile, this may not uniquely define a percentile. Our convention is to take an observed value for the sample percentile unless two adjacent values both satisfy the definition. In this latter case, take their mean. This coincides with the procedure for obtaining the median when the sample size is even. (Most computer packages linearly interpolate between the two adjacent values. For moderate or large sample sizes, the particular convention used to locate a sample percentile between the two observations are inconsequential.)

The following rule simplifies the calculation of sample percentiles.

Calculating the sample 100 p-th Percentile
1. Order the n observations from smallest to largest.
2. Determine the product 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 mean of the k-th and $(k + 1)$-st ordered observations.

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

Sample quartiles

first quartile $Q_1 = $ 25th percentile
second quartile $Q_2 = $ 50th percentile
third quartile $Q_3 = $ 75th percentile

EXAMPLE Calculation of percentiles from the sulfur emission data

Obtain the quartiles and the 97th percentile for the sulfur emission data on page **??**.
Solution The ordered data are:

6.2	7.7	8.3	9.0	9.4	9.8	10.5	10.7	11.0	11.2	11.8
12.3	12.8	13.2	13.3	13.5	13.9	14.4	14.5	14.7	15.2	15.5
15.8	15.9	16.2	16.7	16.9	17.0	17.3	17.5	17.6	17.9	18.0
18.0	18.1	18.1	18.4	18.5	18.7	19.0	19.1	19.2	19.3	19.4
19.4	20.0	20.1	20.1	20.4	20.5	20.8	20.9	21.4	21.6	21.9
22.3	22.5	22.7	22.7	22.9	23.0	23.5	23.7	23.9	24.1	24.3
24.6	24.6	24.8	25.7	25.9	26.1	26.4	26.6	26.8	27.5	28.5
28.6	29.6	31.8								

According to our calculation rule, $n\,p \;=\; 80\,(\,\frac{1}{4}\,) \;=\; 20$ is an integer, so we take the mean of the 20th and 21st ordered observations.

$$Q_1 \;=\; \frac{14.7 + 15.2}{2} \;=\; 14.95.$$

Since $n\,p \;=\; 80\,(\,\frac{1}{2}\,) \;=\; 40$, the second quartile, or median, is the mean of the 40th and 41st ordered observations

$$Q_2 \;=\; \frac{19.0 + 19.1}{2} \;=\; 19.05$$

while the third quartile is the mean of the 60th and 61st:

$$Q_3 \;=\; \frac{22.9 + 23.0}{2} \;=\; 22.95.$$

To obtain the 97th percentile $P_{0.97}$, we determine that $0.97 \times 80 \;=\; 77.6$ which we round up to 78. Counting in to the 78-th position, we obtain

$$P_{0.95} \;=\; 28.6.$$

The 97th percentile provides a useful description regarding days of high emission. On only 3% of the days are more than 28.6 tons of sulfur put into the air.

In the context of monitoring high values, we also record that the maximum emission was 31.8 tons.

■

In the context of monitoring high values, we also record that the maximum emission was 31.8 tons.

The **minimum** and **maximum** observations also convey information concerning the amount of variability present in a set of data. Together, they describe the interval containing all of the observed values and whose length is the

$$\textbf{range} \quad = \quad \text{maximum} \; - \; \text{minimum}$$

Care must be taken when interpreting the range since a single large or small observation can greatly inflate its value.

The amount of variation in the middle half of the data is described by the

$$\textbf{interquartile range} \quad = \quad \text{third quartile} \; - \; \text{first quartile} \quad = Q_3 - Q_1$$

EXAMPLE Calculation of their range and interquartile range

Obtain the range and interquartile range for the sulfur emission data on page **??**.

Solution The minimum = 6.2. From the previous example, the maximum = 31.8, $Q_1 =$ 14.95 and $Q_3 = 22.95$.

$$\begin{aligned} \text{range} &= \text{maximum} - \text{minimum} = 31.8 - 6.2 = 25.6 \text{ tons} \\ \text{interquartile range} &= Q_3 - Q_1 = 22.95 - 14.95 = 8.00 \text{ tons.} \end{aligned}$$

■

Boxplots

The summary information contained in the quartiles is highlighted in a graphic display called a **boxplot**. The center half of the data, extending from the first to the third quartile, is represented by a rectangle. The median is identified by a bar within this box. A line extends from the third quartile to the maximum and another line extends from the first quartile to the minimum. (For large data sets the lines may only extend to the 95th and 5th percentiles).

Figure 2.12 gives the boxplot for the sulfur emission data on page **??** . The symmetry seen in the histogram is also evident in this boxplot.

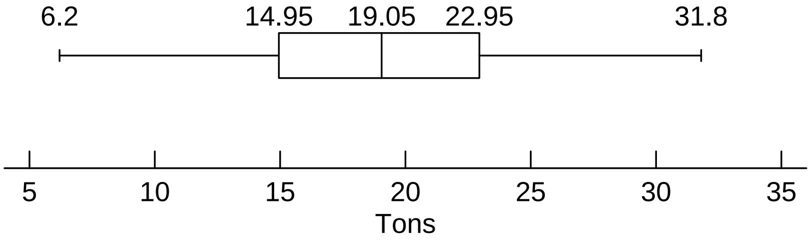


FIGURE 2.12 Boxplot of the sulfur emission data

A **modified boxplot** can both identify outliers and reduce their effect on the shape of the boxplot. The outer line extends to the largest observation only if it is not too far from the third quartile. That is, for the line to extend to the largest observation, it must be within 1.5 × (interquartile range) units of Q_3. The line from Q_1 extends to the smallest observation if it is within that same limit. Otherwise the line extends to the next most extreme observations that fall within this interval.

EXAMPLE A modified boxplot–possible outliers are detached

Construct a modified boxplot for the neutrino interarrival time data (see page **??**)

.021	.107	.179	.190	.196	.283	.580	.854	1.18	2.00	7.30

Construct a modified boxplot.

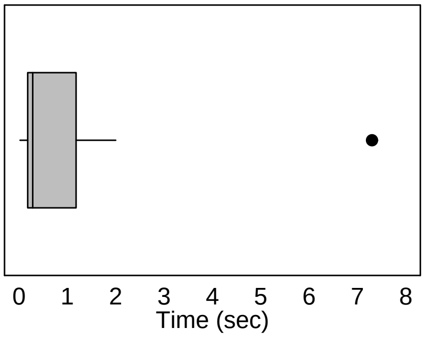


FIGURE 2.13 Modified boxplot for nutrino data

Solution Since $n/4 = 11/4 = 2.75$, the first quartile is the third ordered time .179 and $Q_3 = 1.18$ so the interquartile range is $1.18 - .179 = 1.001$. Further $1.5 \times 1.001 = 1.502$ and the smallest observation is closer than this to $Q_1 = .179$ but

$$\text{maximum} - Q_3 \;=\; 7.30 \;-\; 1.18 = 6.12 \text{ exceeds } 1.502 \;=\; 1.5 \times (\text{interquartile range}).$$

As shown in Figure 2.13, the line to the right extends to 2.00, the most extreme observation within 1.502 units, but not to the largest observation which is shown as detached from the line.

■

Boxplots are particularly effective for graphically portraying comparisons among sets of observations. They are easy to understand and have a high visual impact.

EXAMPLE Multiple boxplots can reveal differences and similarities

Sometimes, with rather complicated components like hard disk drives or RAM chips for computers, quality is quantified as an index with target value 100. Typically, a quality index will be based upon the deviations of several physical characteristics from their engineering specifications. Figure 2.14 shows the quality index at 4 manufacturing plants.

Comment on the relationships between quality at different plants.

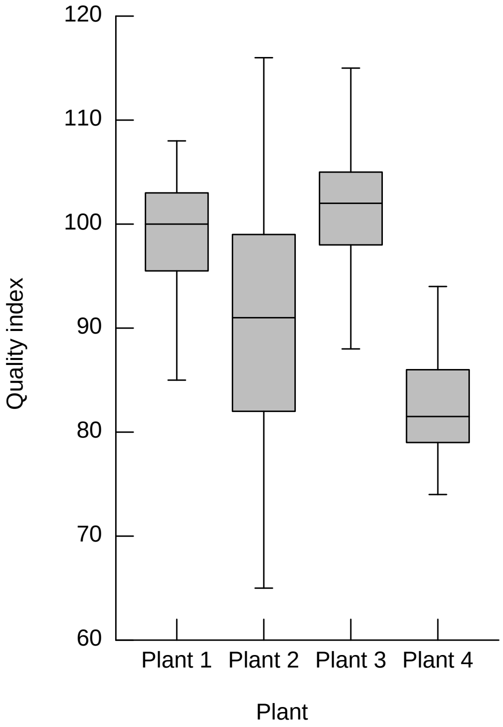


FIGURE 2.14 Boxplot of the quality index

Solution It is clear from the graphic that plant 2 needs to reduce its variability and that plants 2 and 4 need to improve their quality level.

■

We conclude this section with a warning. Sometimes it is the trend over time that it is the most important feature of the data. This feature would be lost entirely if the set of data were summarized in a dot diagram, stem-and-leaf display, or boxplot. Figure 2.15 illustrates this point by a time plot of the ozone in October, in Dobson units, over a region of the South Pole. The apparent downward trend, if real, is of major scientific interest and may be vital to life on our planet.

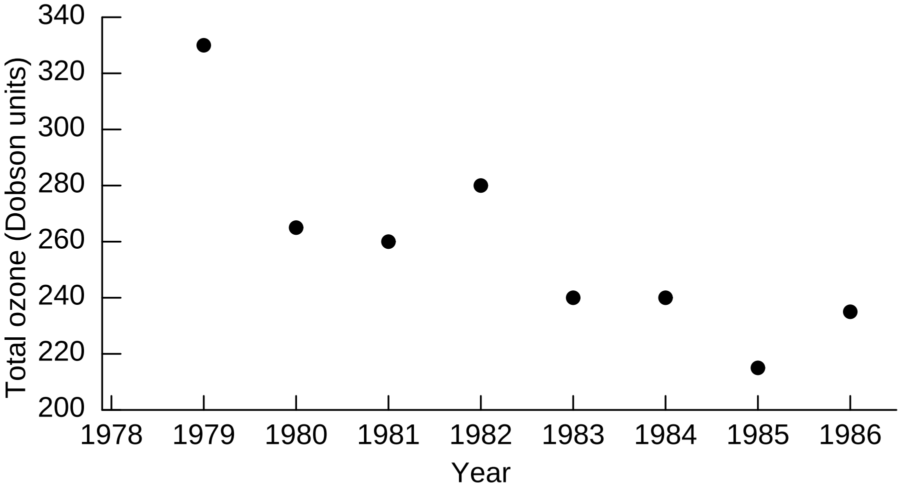


FIGURE 2.15 The monthly average total atmospheric ozone, for October, over the South Polar latitudes

2.7 The Calculation of $\overline{x}$ and s

In this section we discuss methods for calculating $\overline{x}$ and s for **raw data** (ungrouped) as well as grouped data. These methods are particularly well suited for small hand held calculators and they are rapid. They are also accurate, except in extreme cases where, say, the data differ only in the seventh or higher digits.

The calculation of $\overline{x}$ for ungrouped data does not pose any problems; we have only to add the values of the observations and divide by n. On the other hand, the calculation of s^2 is usually cumbersome if we directly use the formula defining s^2 on page **??**. Instead, we shall use the algebraically equivalent form.

Variance (hand-held calculator formula)

$$s^2 = \frac{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}{n(n-1)}$$

which requires less labor to evaluate with a calculator. (In Exercise 2.51 the reader will be asked to show that this formula is, in fact, equivalent to the one on page **??**). This expression for variance is without $\overline{x}$, which reduces roundoff error.

EXAMPLE Calculating variance using the hand-held calculator formula

Find the mean and the standard deviation of the following miles per gallon (mpg) obtained in 20 test runs performed on urban roads with an intermediate-size car:

19.7	21.5	22.5	22.2	22.6
21.9	20.5	19.3	19.9	21.7
22.8	23.2	21.4	20.8	19.4
22.0	23.0	21.1	20.9	21.3

Solution Using a calculator, we find that the sum of these figures is 427.7 and that the sum of their squares is 9,173.19. Consequently,

$$\overline{x} = \frac{427.7}{20} = 21.39 \text{ mpg}$$

and

$$s^2 = \frac{20(9,173.19) - (427.7)^2}{20 \times 19} = 1.412$$

and it follows that $s = 1.19$ mpg. In computing the necessary sums we usually retain all decimal places, but as in this example, at the end we usually round to one more decimal than we had in the original data.

■

See Exercise 2.58 for a computer calculation. This is the recommended procedure because it is easy to check the data entered for accuracy and the calculation is free of human error. Most importantly, the calculation of variance can be done using the square of the deviations $x_i - \overline{x}$ rather than the squares of the observations x_i and this is numerically more stable.

Not too many years ago, one of the main reasons for grouping data was to simplify the calculation of descriptions such as the mean and the standard deviation. With easy access to statistical calculators and computers, this is no longer the case, but we shall nevertheless discuss here the calculation of $\overline{x}$ and s from grouped data, since some data (for instance, from government publications) may be available only in grouped form.

To calculate $\overline{x}$ and s from grouped data, we shall have to make some assumption about the distribution of the values within each class. If we represent all values within a class by the corresponding class mark, the sum of the x's and the sum of their squares can now be written

$$\sum_{i=1}^{k} x_i\, f_i \quad \text{and} \quad \sum_{i=1}^{k} x_i^2\, f_i$$

where x_i is the class mark of the i-th class, f_i is the corresponding class frequency, and k is the number of classes in the distribution. Substituting these sums into the formula for $\overline{x}$ and the computing formula for s^2, we get

Mean and variance (grouped data)

$$\overline{x} = \frac{\sum\limits_{i=1}^{k} x_i\, f_i}{n}$$

$$s^2 = \frac{n \cdot \sum\limits_{i=1}^{k} x_i^2\, f_i \; - \; \left(\sum\limits_{i=1}^{k} x_i\, f_i \right)^2}{n(n-1)}$$

EXAMPLE Calculating a mean and variance from grouped data

Use the distribution obtained on page **??** to calculate the mean and the variance of the sulfur oxides emission data.

Solution Recording the class marks and the class frequencies in the first two columns, and the products $x_i\, f_i$ and $x_i^2\, f_i$ in the third and fourth columns, we obtain

x_i	f_i	$x_i\, f_i$	$x_i^2\, f_i$
7	3	21	147
11	10	110	1,210
15	14	210	3,150
19	25	475	9,025
23	17	391	8,993
27	9	243	6,561
31	2	62	1,922
	80	1,512	31,008

Then, substitution into the formula yields

$$\overline{x} = \frac{1,512}{80} = 18.90$$

and

$$s^2 = \frac{80(31,008) - (1,512)^2}{80 \times 79} = 30.77.$$

■

EXERCISES

2.27 In each of the following situations, should your value be near the average or an outlier? If an outlier, should it be too large or too small?

(a) Income on your starting job?

(b) Your score on the final exam in a physics class?

(c) Your weight in ten years.

2.28 In each of the following situations, should your value be near the average or an outlier? If outlier, should it be too large or too small?

(a) The time you take to complete a lab assignment next week.

(b) Your white blood cell count.

2.29 Is the influence of a single outlier greater on the mean or the median? Explain.

2.30 Is the influence of a single outlier greater on the sample range or the interquartile range? Explain.

2.31 Referring to Exercise 1.6, we see that the sample of 4 deviations (observation − specification) for a critical crank bore diameter is

$$-6 \quad 1 \quad -4 \quad -3$$

ten-thousandths of an inch during the second hour. For these 4 deviations

(a) calculate the sample mean $\overline{x}$;

(b) calculate the sample standard deviation s;

(c) on average, is the hole too large or too small?

2.32 A company was experiencing a chronic weld defect problem with a water outlet tube assembly. Each assembly manufactured is leak tested in a water tank. Data were collected on a gap between the flange and the pipe for 6 assemblies that leaked and 6 good assemblies that passed the leak test.

Leaker	.290	.104	.207	.145	.104	.124

(a) Calculate the sample mean $\overline{x}$.

(b) Calculate the sample standard deviation s.

2.33 Refer to Exercise 2.32. The measurements for 6 assemblies that did not leak were

Good .207 .124 .062 .301 .186 .124

(a) Calculate the sample mean $\overline{x}$.

(b) Calculate the sample standard deviation s.

(c) Does there appear to be a major difference in gap between assemblies that leaked and those which did not? The quality improvement group turned their focus to welding process variables.

2.34 A contract for the maintenance of a national railway's high horsepower locomotives was given to a large private company. After one year of experience with the maintenance program, those in charge of the program felt that major improvements could be made in the reliability of the locomotives. To document the current status, they collected data on the cost of materials for rebuilding traction motors. Use the data below to

(a) calculate the sample mean $\overline{x}$,

(b) calculate the sample standard deviation s.

Materials costs for rebuilding traction motors (1000s of dollars):

1.41	1.70	1.03	0.99	1.68	1.09	1.68	1.94
1.53	2.25	1.60	3.07	1.78	0.67	1.76	1.17
1.54	0.99	0.99	1.17	1.54	1.68	1.62	0.67
0.67	1.78	2.12	1.52	1.01			

2.35 If the mean annual salary paid to the chief executives of three engineering firms is \$125,000, can one of them receive \$400,000?

2.36 Records show that in Phoenix, Arizona, the normal daily maximum temperature for each month is 65, 69, 74, 84, 93, 102, 105, 102, 98, 88, 74 and 66 degrees Fahrenheit. Verify that the mean of these figures is 85 and comment on the claim that, in Phoenix, the average daily maximum temperature is a very comfortable 85 degrees.

2.37 The following are the numbers of twists that were required to break 12 forged alloy bars: 33, 24, 39, 48, 26, 35, 38, 54, 23, 34, 29 and 37. Find

(a) the mean;

(b) the median.

2.38 With reference to the preceding exercise, find s using

(a) the formula that defines s;

(b) the hand-held calculator formula for s.

2.39 The following are the numbers of minutes that a person had to wait for the bus to work on 15 working days

10 1 13 9 5 9 2 10 3 8 6 17 2 10 15

Find

(a) the mean;

(b) the median;

(c) Draw a boxplot.

2.40 With reference to the preceding exercise, find s^2 using

(a) the formula that defines s^2;

(b) the hand-held calculator formula for s^2.

2.41 Material manufactured continuously before being cut and wound into large rolls must be monitored for thickness (caliper). A sample of 10 measurements on paper, in mm, yielded

32.2 32.0 30.4 31.0 31.2 31.2 30.3 29.6 30.5 30.7

Find the mean and quartiles for this sample.

2.42 For the four observations 9 7 15 5

(a) calculate the deviations $(x_i - \overline{x})$ and check that they add to 0;

(b) calculate the variance and the standard deviation.

2.43 With reference to Exercise 2.14 on page **??**, draw a boxplot.

2.44 With reference to Exercise 2.3 on page **??**, calculate $\overline{x}$ and s.

2.45 Find the mean and the standard deviation of the 20 humidity readings on page **??** by using

(a) the raw (ungrouped) data;

(b) the distribution obtained in that example.

2.46 Use the distribution in Exercise 2.9 on page **??** to find the mean and the variance of the breaking strengths.

2.47 Use the distribution obtained in Exercise 2.12 on page **??** to find the mean and the standard deviation of the ignition times. Also determine the coefficient of variation.

2.48 Use the distribution obtained in Exercise 2.14 on page **??** to find the coefficient of variation of the productivity data.

2.49 In three recent years, the price of copper was 69.6, 66.8 and 66.3 cents per pound, and the price of bituminous coal was 19.43, 19.82 and 22.40 dollars per short ton. Which of the two sets of prices is relatively more variable?

2.50 Show that $\sum_{i=1}^{n} (\, x_i \; - \; \overline{x} \,) = 0$ for any set of observations $x_1, x_2, \ldots, x_n$.

2.51 Show that the computing formula for s^2 on page **??** is equivalent to the one used to define s^2 on page **??**.

2.52 If data are coded so that $x_i \; = \; c \cdot u_i \; + \; a$, show that $\overline{x} \; = \; c \cdot \overline{u} \; + \; a$ and $s_x \; = \; |c| \cdot s_u$.

2.53 To find the **median** of a distribution obtained for n observations, we first determine the class into which the median must fall. Then, if there are j values in this class and k values below it, the median is located $\frac{(n/2) - k}{j}$ of the way into this class, and to obtain the median we multiply this fraction by the class interval and add the result to the lower class boundary of the class into which the median must fall. This method is based on the assumption that the observations in each class are "spread uniformly" throughout the class interval, and this is why we count $\frac{n}{2}$ of the observations instead of $\frac{n+1}{2}$ as on page **??**. To illustrate, let us refer to the distribution of the sulfur oxides emission data on page **??**. Since $n = 80$, it can be seen that the median must fall into the class $17.0 - 20.9$, and since $j = 25$ and $k = 27$, it follows that the median is $16.95 + \frac{40 - 27}{25} \cdot 4 = 19.03$.

(a) Use the distribution obtained in Exercise 2.9 on page **??** to find the median of the grouped breaking strengths.

(b) Use the distribution obtained in Exercise 2.12 on page **??** to find the median of the grouped ignition times.

2.54 For each of the following distributions, decide whether it is possible to find the mean and whether it is possible to find the median. Explain your answers.

(a)

Grade	Frequency
40−49	5
50−59	18
60−69	27
70−79	15
80−89	6

(b)

IQ	Frequency
less than 90	3
90−99	14
100−109	22
110−119	19
more than 119	7

(c)

Weight	Frequency
110 or less	41
101−110	13
111−120	8
121−130	3
131−140	1

2.55 To find the first and third quartiles Q_1 and Q_3 for grouped data, we proceed as in Exercise 2.54, but count $\frac{n}{4}$ and $\frac{3n}{4}$ of the observations instead of $\frac{n}{2}$.

(a) With reference to the distribution of the sulfur oxides emission data on page **??**, find Q_1, Q_3 and the interquartile range.

(b) Find Q_1 and Q_3 for the distribution of the ignition time data given on page **??**.

2.56 If k sets of data consist, respectively, of $n_1, n_2, \ldots, n_k$ observations and have the means $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$, then the overall mean of all the data is given by the formula

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i \, \overline{x}_i}{\sum_{i=1}^{k} n_i}$$

(a) The average annual salaries paid to top-level management in three companies are \$94,000, \$102,000, and \$99,000. If the respective numbers of top-level executives in these companies are 4, 15, and 11, find the average salary paid to these 30 executives.

(b) In a nuclear engineering class there are 22 juniors, 18 seniors, and 10 graduate students. If the juniors averaged 71 in the midterm examination, the seniors averaged 78, and the graduate students averaged 89, what is the mean for the entire class?

2.57 The formula for preceding exercise is a special case of the following formula for the **weighted mean**

$$\overline{x}_w = \frac{\sum_{i=1}^{k} w_i \, x_i}{\sum_{i=1}^{k} w_i}$$

where w_i is a weight indicating the relative importance of the i-th observation.

(a) If an instructor counts the final examination in a course four times as much as each 1-hour examination, what is the weighted average grade of a student who received grades of 69, 75, 56, and 72 in four 1-hour examinations and a final examination grade of 78?

(b) From 1999 to 2004 the cost of food increased by 53% in a certain city, the cost of housing increased by 40% and the cost of transportation increased by 34%. If the average salaried worker spent 28% of his or her income on food, 35% on housing, and 14% on transportation, what is the combined percentage increase in the cost of these items?

2.58 Modern computer software packages have come a long way toward removing the tedium of calculating statistics. *MINITAB* is one common and easy-to-use package. We illustrate the use of the computer using *MINITAB* commands. Other easy-to-use packages have a quite similar command structure.

The lumber used in the construction of buildings must be monitored for strength. Data for strength, in pounds per square inch, for 2 × 4 pieces of lumber are in the file 2.58dat. We give the basic commands that calculate n, $\overline{x}$ and s as well as the quartiles.

The session commands require the data to be set in the first column, C1, of the *MINITAB* work sheet. The command for creating a box plot is also included.

Data in 2.58.dat
strength
Dialog box:
Stat> Basic Statistics > Descriptive Statistics
Type *strength* in **Variables**. Click **Graphs**. Click **Graphical Summary**
Click **OK**. Click **OK**.

Output : (partial)

```
Variable          N       Mean   Median   StDev
Strength         30     1908.8   1863.0   327.1

Variable   Minimum    Maximum       Q1      Q3
Strength    1325.0     2983.0   1711.5  2071.8
```

Use *MINITAB*, or some other statistical package, to find $\overline{x}$ and s for

(a) the decay times on page **??**;

(b) the interrequest times on page **??**.

2.59 (Further *MINITAB* calculation and graphs) With the observations on the strength (in pounds per square inch) of 2×4 pieces of lumber already set in C1, the sequence of choices and clicks

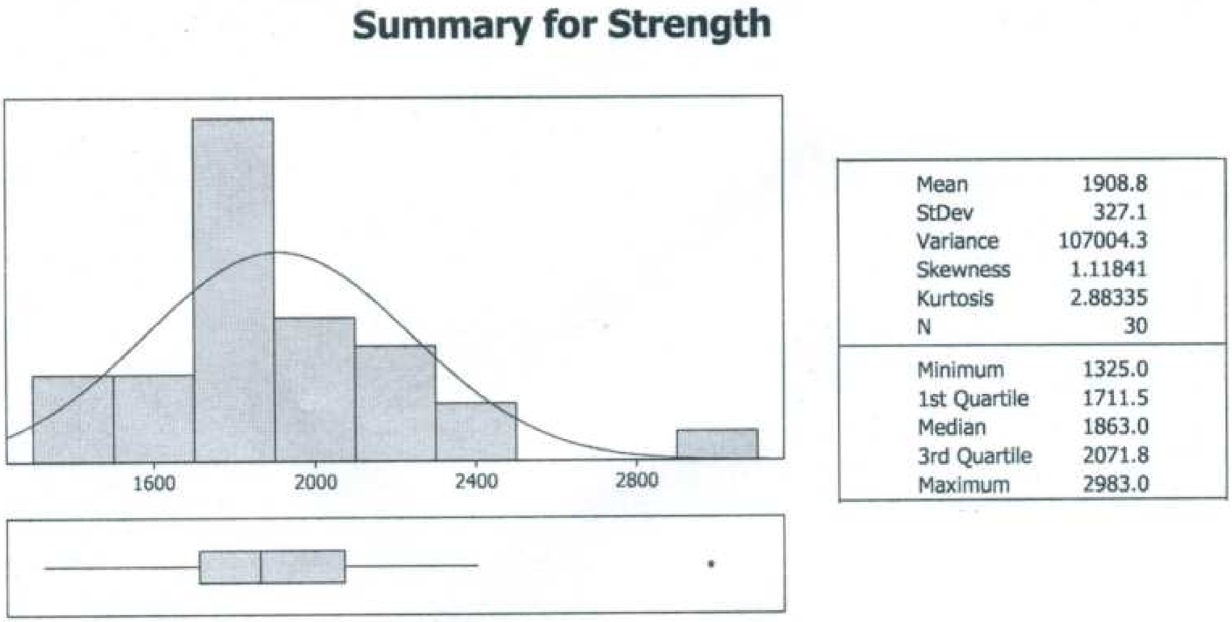
Stat> Basic Statistics > Graphical summary

Type *strength* in **Variables**. Click **OK**

produces an even more complete summary.

The ordered strength data are

1325 1419 1490 1633 1645 1655 1710 1712 1725 1727 1745
1828 1840 1856 1859 1867 1889 1899 1943 1954 1976 2046
2061 2104 2168 2199 2276 2326 2403 2983



Mean	1908.8
StDev	327.1
Variance	107004.3
Skewness	1.11841
Kurtosis	2.88335
N	30
Minimum	1325.0
1st Quartile	1711.5
Median	1863.0
3rd Quartile	2071.8
Maximum	2983.0

From the ordered data

(a) obtain the quartiles;

(b) construct a histogram and locate the mean, median, Q_1, and Q_3 on the horizontal axes;

(c) repeat parts (a) and (b) with the aluminum alloy data on page **??**.

2.8 A Case Study: Problems with Aggregating Data

As circuit boards and other components move through a company's surface mount technology assembly line, a significant amount of data is collected for each assembly.[1] The data are recorded at several stages of manufacture in a serial tracking database by means of computer terminals located throughout the factory. The data include the board serial number, the type of defect, number of defects and their location. The challenge here is to transform a large amount of data into manageable and useful

[1] Courtesy of Don Ermer

information. When there is a variety of products and lots of data are collected on each, record management and the extraction of appropriate data for product improvement must be done well.

Originally, an attempt was made to understand this large database by *aggregating*, or grouping together, data from all products and performing an analysis of the data as if it were one product! This was a poor practice that decreased the resolution of the information obtained from the database. The products on the assembly line ranged in complexity, maturity, method of processing, and lot size.

To see the difficulties caused by aggregation, consider a typical week's production where 100 printed circuit boards of Product A were produced; 40 boards of Product B; and 60 boards of Product C. Following a wave soldering process, a total of 400 solder defects was reported. This translates to an overall average of 400/200 = 2 defects per board. It was this company's practice to circulate the weekly aggregate average throughout the factory floor for review and comment. It was then the operator's responsibility to take action according to the *misleading* report. Over time, it became apparent that this process was ineffective for improving quality.

However, further analysis of this data on a product by product basis revealed that products A, B, and C actually contributed 151, 231, and 18 defects. Thus, the number of defects per board was 1.51, 5.78, and 0.30 for products A, B and C, respectively. Figure 2.17 correctly shows the average number of defects. Product C has a significantly lower and Product B has a significantly higher defect rate relative to the incorrect aggregated average. These latter are also the more complex boards.

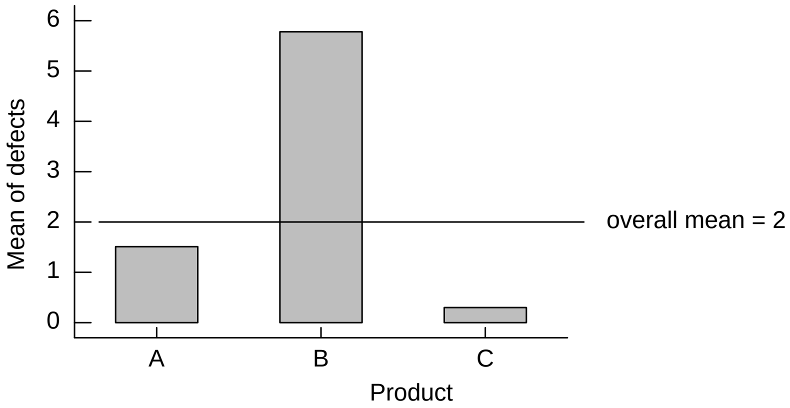


FIGURE 2.17 Average number of defects per product type

These data concern the number of defects that occurred when boards were wave soldered after an assembly stage. The next step was to implement control charts for the number of defects for each of the three products. The numbers of defects for Product B were:

10	8	8	4	6	8	8	10	6	7	4	2	4	5	5
5	2	11	6	6	5	7	3	4	3	2	6	5	1	7
3	1	1	5	4	5	12	13	11	8					

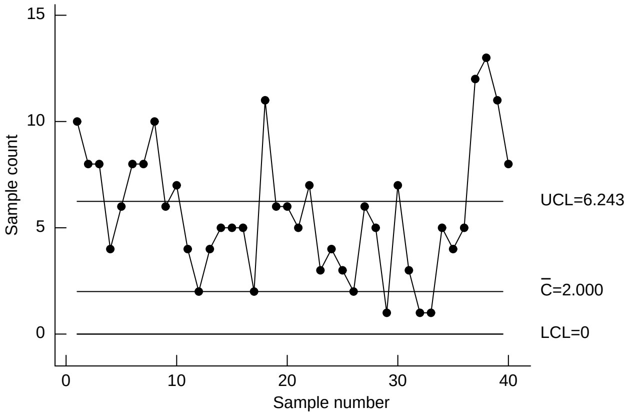
The appropriate control chart is a time plot where the serial numbers of the product or sample are on the horizontal axis and the corresponding number of defects on the vertical axis. In this C-chart, the central line labeled $\overline{C}$ is the average number of defects over all cases in the plot. The dashed lines are the control limits set at three standard deviations about the central line. (For reasons explained in Section 14.6, we use $\sqrt{\overline{C}}$ rather than s when the data are numbers of defects.)

$$\text{LCL} = \overline{C} - 3\sqrt{\overline{C}}$$

$$\text{UCL} = \overline{C} + 3\sqrt{\overline{C}}.$$

Figure 2.18(a) gives a $C-$chart constructed for Product B, but where the centerline is incorrectly calculated from the aggregated data is $\overline{C} = 2.0$. This is far too low and so is the upper control limit 6.24. The lower control limit is negative so we use 0. It looks like a great many of the observations are out of control because they exceed the upper control limit.

When the $C-$chart is correctly constructed on the basis of data from Product B alone, the centerline is $\overline{C} = 231/40 = 5.775$ and the upper control limit is 12.98. The lower control limit is again negative so we use 0. From Figure 2.18(b), the correct C-chart, the wave soldering process for Product B appears to be in control except for time 38 when 13 defects were observed.



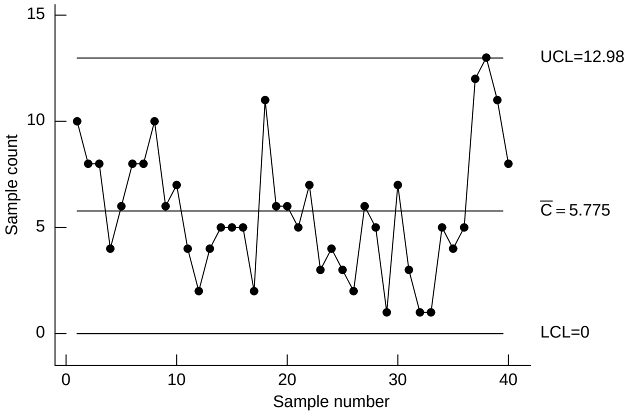


FIGURE 2.18a Incorrect C-chart for Defects
FIGURE 2.18b Correct C-chart for Defects

With the data segregated into products, separate charts were constructed for each of the three products. With this new outlook on data interpretation, a number of improvement opportunities surfaced that were previously disguised by aggregation. For example, by reducing the dimensions of an electrical pad a significant reduction was achieved in the number of solder bridges between pins. This same design change was added to all of the board specifications and improvements were obtained on all products.

In summary, the aggregation of data from different products, or more generally from different sources, can lead to incorrect conclusions and mask opportunities for quality improvement. Segregating data by product, although more time-consuming initially, can lead to significant reduction in waste and manufacturing costs.

Do's and Don'ts

Do's

1. Graph the data as a dot diagram or histogram to assess the overall pattern of data.
2. Calculate the summary statistics: sample mean, standard deviation and quartiles to describe the data set.

Don'ts

1. Don't routinely calculate summary statistics without identifying unusual observations which may have undue influence on the values of the statistics.

REVIEW EXERCISES

2.60 From 2000 computer chips inspected by the manufacturer, the following numbers of defects were recorded.

holes not open	182
holes too large	55
poor connections	31
incorrect size chip	5
other	7

Draw a Pareto chart.

2.61 Draw

(a) a frequency table of the aluminum alloy strength data on page **??** using the classes [66.0, 67.5), [67.5, 69.0), [69.0, 70.5), [70.5−72.0), [72.0, 73.5), [73.5, 75.0), [75.0, 76.5), where the right-hand endpoint is excluded.

(b) a histogram using the frequency table in part (a).

2.62 Draw

(a) A frequency table of the interrequest time data on page 2? using the intervals [0, 2,500), [2,500, 5,000), [5,000, 10,000), [10,000, 20,000), [20,000, 40,000), [40,000, 60,000), [60,000, 80,000) where the left-hand endpoint is included but the right-hand endpoint is not.

(b) a histogram using the frequency table in part (a) (note that the intervals are unequal, so make the height of the rectangle equal relative frequency/width).

2.63 Direct evidence of Newton's universal law of gravitation was provided from a renowned experiment by Henry Cavendish (1731-1810). In the experiment, masses of objects were determined by weighing and the measured force of attraction was used to calculate the density of the earth. The values of the earth's density, in time order by row, are

5.36	5.29	5.58	5.65	5.57	5.53	5.62	5.29
5.44	5.34	5.79	5.10	5.27	5.39	5.42	5.47
5.63	5.34	5.46	5.30	5.75	5.68	5.85	

(Source : *Philosophical Transactions* 17 (1998) : 469).

(a) Find the mean and standard deviation.

(b) Find the median, Q_1, and Q_3.

(c) Plot the observations versus time order. Is there any obvious trend?

2.64 J. J. Thomson (1856-1940) discovered the electron by isolating negatively charged particles for which he could measure the mass-charge ratio. This ratio appeared to be constant over a wide range of experimental conditions and, consequently, could be a characteristic of a new particle. His observations, from two different cathode ray tubes that used air as the gas, are

Tube 1	0.57	0.34	0.43	0.32	0.48	0.40	0.40
Tube 2	0.53	0.47	0.47	0.51	0.63	0.61	0.48

(Source : *Philosophical Magazine* 44; 5 (1897): 293.)

(a) Draw a dot diagram with solid dots for Tube 1 observations and circles for Tube 2 observations.

(b) Calculate the mean and standard deviation for the Tube 1 observations.

(c) Calculate the mean and standard deviation for the Tube 2 observations.

2.65 With reference to Exercise 2.64,

(a) calculate the median, maximum, minimum, and range for Tube 1 observations;

(b) calculate the median, maximum, minimum, and range for the Tube 2 observations.

2.66 A. A. Michelson (1852-1931) made many series of measurements of the speed of light. Using a revolving mirror technique, he obtained

12 30 30 27 30 39 18 27 48 24 18

for the differences (velocity of light in air) − (229, 700) km/s. (Source: *The Astrophysical Journal* 65 (1927):11.)

(a) Draw a dot diagram.

(b) Find the median and the mean. Locate both on the dot diagram.

(c) Find the variance and standard deviation.

2.67 With reference to Exercise 2.66

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) draw a boxplot.

2.68 A civil engineer monitors water quality by measuring the amount of suspended solids in a sample of river water. Over 11 weekdays, she observed

14 12 21 28 30 63 29 63 55 19 20

suspended solids (parts per million).

(a) Draw a dot diagram.

(b) Find the median and the mean. Locate both on the dot diagram.

(c) Find the variance and standard variation.

2.69 With reference to Exercise 2.68

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) construct a boxplot.

2.70 With reference to the aluminum alloy strength data in the example on page **??**

(a) find the quartiles;

(b) find the minimum, maximum, range, and interquartile range;

(c) find the 10th percentile and 20th percentile.

2.71 With reference to Exercise 2.70, draw a boxplot.

2.72 With reference to the aluminum alloy strength data in the example on page **??**, make a stem-and-leaf display.

2.73 In five tests, one student averaged 63.2 with a standard deviation of 3.3, whereas another student averaged 78.8 with a standard deviation of 5.3. Which student is relatively more consistent?

2.74 With reference to the lumber strength data in Exercise 2.59, the statistical software package *SAS* produced the output in Figure 2.19. Using this output

(a) identify the mean and standard deviation and compare these answers with the values given in Exercise 2.59.

(b) draw a boxplot.

```
                       UNIVARIATE PROCEDURE

VARIABLE = STRENGTH

                             Moments

               N                30  Sum Wgts          30
               Mean       1908.767  Sum            57263
               Std Dev     327.115  Variance    107004.3
               Skewness   1.118406  Kurtosis    2.883349
               USS         1.124E8  CSS          3103123
               CV         17.13751  Std Mean    59.72276

                         Quantiles (Def = 5)

                  100%  Max  2983  99%     2983
                   75%  Q3   2061  95%     2403
                   50%  Med  1863  90%     2301
                   25%  Q1   1712  10%   1561.5
                    0%  Min  1325   5%     1419
                                    1%     1325

                  Range      1658
                  Q3-Q1       349
```

Figure 2.19 Selected SAS output to describe the lumber strength data from Exercise 2.59

2.75 Civil engineers must monitor flow on rivers where power is generated. The following are the daily mean flow rates (MGD) on the Namekagon River during the month of May for 47 years.

602.0	517.5	572.5	392.4	505.8	547.5	389.1	497.2
794.8	657.6	904.7	595.5	611.9	482.9	698.6	606.7
986.4	567.7	400.1	634.9	448.4	479.1	1156.0	718.5
575.6	743.3	1146.0	461.6	644.0	480.8	429.1	626.9
833.9	889.0	752.6	516.5	817.2	895.8	572.2	563.7
679.3	738.0	618.9	390.8	550.9	425.9	760.6	

(a) obtain the quartiles;

(b) obtain the 90th percentile;

(c) construct a histogram.

2.76 The national Highway Traffic Safety Administration reported the relative speed (rounded to the nearest 5 mph) of automobiles involved in accidents one year. The percentages at different speeds were

20 mph or less	2.0%
25 or 30 mph	29.7%
35 or 40 mph	30.4%
45 or 50 mph	16.5%
55 mph	19.2%
60 or 65 mph	2.2%

(a) From these data can we conclude that it is quite safe to drive at high speeds? Why or why not?

(b) Why do most accidents occur in the 35 or 40 mph and in the 25 or 30 mph ranges?

(c) Construct a density histogram using the endpoints 0, 22.5, 32.5, 42.5, 52.5, 57.5, 67.5 for the intervals.

2.77 Given a five-number summary,

$$\text{minimum} \qquad Q_1 \qquad Q_2 \qquad Q_3 \qquad \text{maximum}$$

is it possible to determine whether or not an outlier is present? Explain.

2.78 Given a stem-and-leaf display, is it possible to determine whether or not an outlier is present? Explain.

2.79 Traversing the same section of interstate highway on eleven different days, a driver recorded the number of cars pulled over by the highway patrol.

$$0 \quad 1 \quad 3 \quad 0 \quad 2 \quad 0 \quad 1 \quad 0 \quad 2 \quad 1 \quad 0$$

(a) Create a dot plot.

(b) There is a long tail to the right. You might expect the sample mean to be larger than the median. Calculate the sample mean and median and compare the two measures of center. Comment.

KEY TERMS

(*with page references*)

Absolute variation **??**
Arithmetic mean **??**
Bar chart **??**
Boxplot **??**
Categorical distribution **??**
Class boundary **??**
Class frequency **??**
Class interval **??**
Class limit **??**
Class mark **??**
Coefficient of variation **??**
Cumulative distribution **??**
Cumulative percentage distribution **??**
Density histogram **??**
Dot diagram **??**
Double-stem display **??**
Empirical cumulative distribution **??**
Endpoint convention **??**
Exploratory data analysis **??**
Five-stem display **??**
Frequency distribution **??**
Histogram **??**
Interquartile range **??**
Leaf **??**
Mean **??**
Median **??**
Modified boxplot **??**
Numerical distribution **??**
Ogive **??**
Outlier **??**
Pareto diagram **??**
Percentage distribution **??**
Percentile **??**
Pie chart **??**
Quartile **??**
Range **??**
Raw data **??**
Relative variation **??**
Sample mean **??**
Sample variance **??**
Standard deviation **??**
Stem **??**
Stem-and-leaf display **??**
Stem label **??**
Variance **??**
Weighted mean **??**