Probability & Statistics

Chapter 1 Introduction to Statistics

Statistics

Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

Data Collection & Descriptive Statistics

Sometimes a statistical analysis begins with a given set of data: For instance, the government regularly collects and publicizes data concerning earthquake occurrences, the unemployment rate and the rate of inflation. Statistics can be used to describe, summarize, and analyze these data.

In some situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data. For instance, suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective. To study this question, the instructor might divide the students into two groups, and use a

different teaching method for each group. At the end of the class the students can be tested and the scores of the members of the different groups compared. If the data, consisting of the test scores of members of each group, are significantly higher in one of the groups, then it might seem reasonable to suppose that the teaching method used for that group is superior.

It is important to note, however, that in order to be able to draw a valid conclusion from the data, it is essential that the students were divided into groups in such a manner that neither group was more likely to have the students with greater natural aptitude for programming. For instance, the instructor should not have let the male class members be one group and the females the other. For if so, then even if the women scored significantly higher than the men, it would not be clear whether this was due to the them,

or to the fact that women may be inherently better than men at learning programming skills. The accepted way of avoiding this pitfall is to divide the class members into the two groups "at random." This term means that the division is done in such a manner that all possible choices of the members of a group are equally likely.

At the end of the experiment, the data should be described. For instance, the scores of the two groups should be presented. In addition, summary measures such as the average score of members of each of the groups should be presented. This part of statistics, concerned with the description and summarization of data, is called descriptive statistics.

Inferential Statistics & Probability Models

After the preceding experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about which teaching method is superior. This part of statistics, concerned with the drawing of conclusions, is called inferential statistics.

To be able to draw a conclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average score of members of the first group is quite a bit higher than that of the second. Can we conclude that this increase is due to the teaching method used? Or is it possible that the teaching method was not responsible for the increased scores but rather that the higher scores of the first group were just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips.

Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.)

To be able to draw logical conclusions from data, we usually make some assumptions about the chances (or probabilities) of obtaining the different data values. The totality of these assumptions is referred to as a probability model for the data.

Population & Samples

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the population. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a sample.

The sample is to be informative because it is representative of the population.

Chapter 2 | Descriptive Statistics

Describing Data Set

The numerical findings of a study should be presented clearly, concisely, and in such a manner that an observer can quickly obtain a feel for the essential characteristics of the data. Over the years it has been found that tables and graphs are particularly useful ways of presenting data, often revealing important features such as the range, the degree of concentration, and the symmetry of the data.

Frequency Tables & Graphs

A data set having a relatively small number of distinct values can be conveniently presented in a frequency table. For instance, Table 1 is a frequency table for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees in computer science. Table 1 tells us, among other things, that the lowest starting salary of $47,000 was received by four of the graduates, whereas the highest salary of $60,000 was received by a single student. The most common starting salary was $52,000, and was received by 10 of the students.

Table 1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Data from a frequency table can be graphically represented by a line graph that plots the distinct data values on the horizontal axis and indicates their frequencies by the heights of vertical lines. A line graph of the data presented in Table 1 is shown in Figure 1.

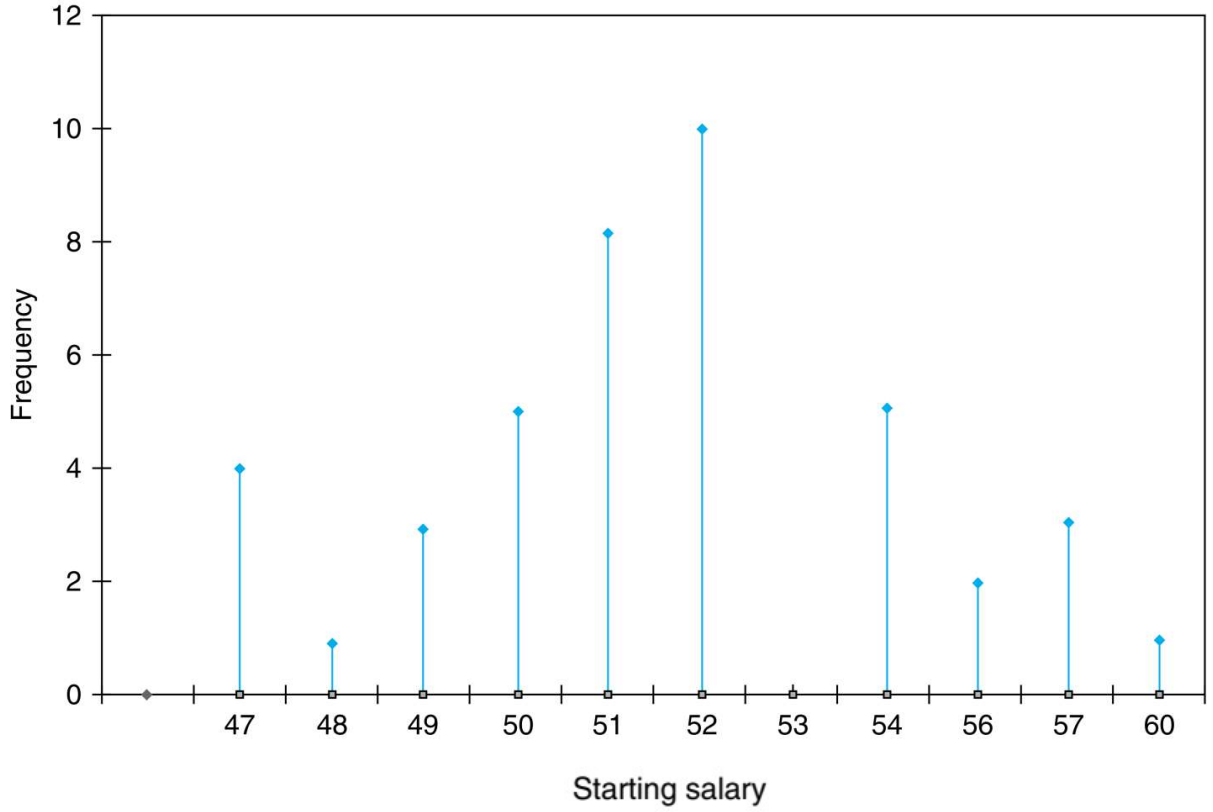


Figure 1

When the lines in a line graph are given added thickness, the graph is called a bar graph.

Figure 2 presents a bar graph.

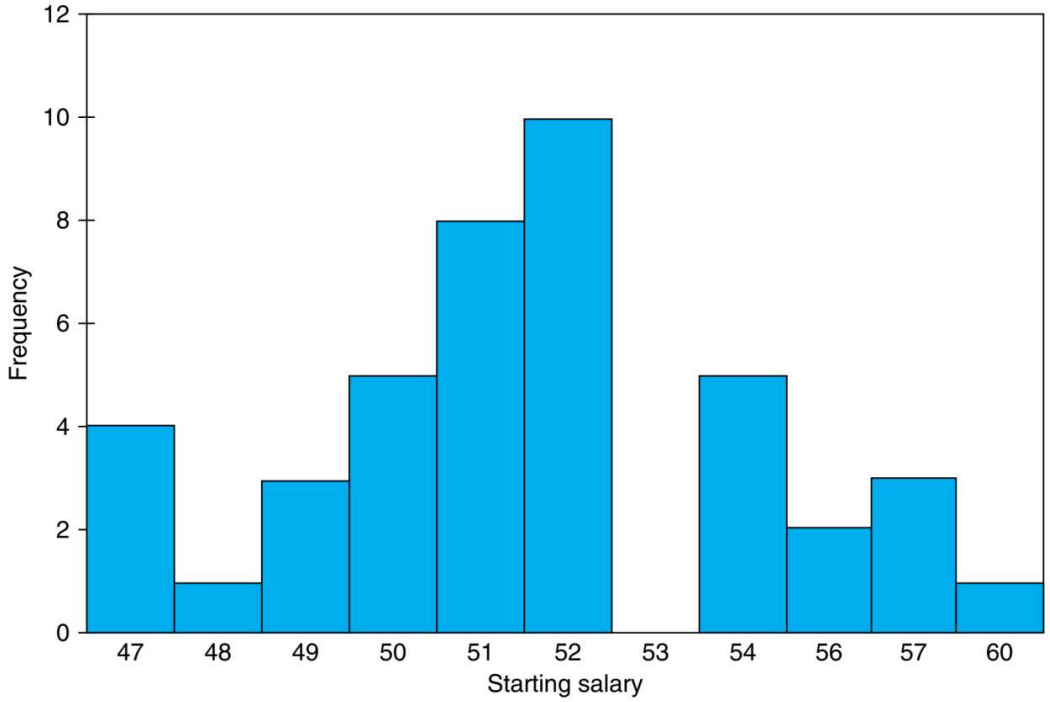


Figure 2 *Bar graph for starting salary data.*

Another type of graph used to represent a frequency table is the frequency polygon, which plots the frequencies of the different data values on the vertical axis, and then connects the plotted points with straight lines. Figure 3 presents a frequency polygon for the data of Table 1.

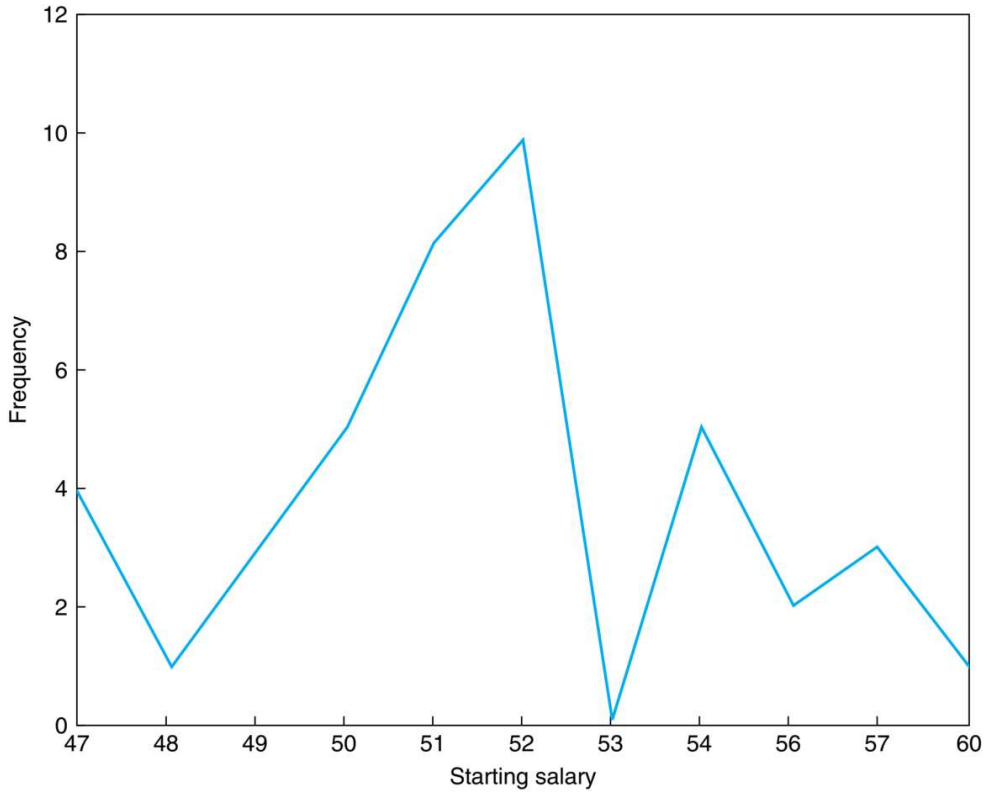


Figure 3

Frequency polygon for starting salary data.

Relative Frequency Tables & Graphs

Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f/n is called its relative frequency. That is, the relative frequency of a data value is the proportion of the data that have that value. The relative frequencies can be represented graphically by a relative frequency line or bar graph or by a relative frequency polygon. Indeed, these relative frequency graphs will look like the corresponding graphs of the absolute frequencies except that the labels on the vertical axis are now the old labels (that gave the frequencies) divided by the total number of data points.

Table 4 is a relative frequency table for the data of Table 1. The relative frequencies are obtained by dividing the corresponding frequencies of Table 1 by 42, the size of the data set.

Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42

Figure 4

Pie Chart

We can construct pie chart by dividing a circle into various sections or slices. It should be used when we want to compare individual categories with the whole. If you want to compare the values of categories with each other, a bar chart may be more useful.

Problem

The following table shows the yearly budget of a family

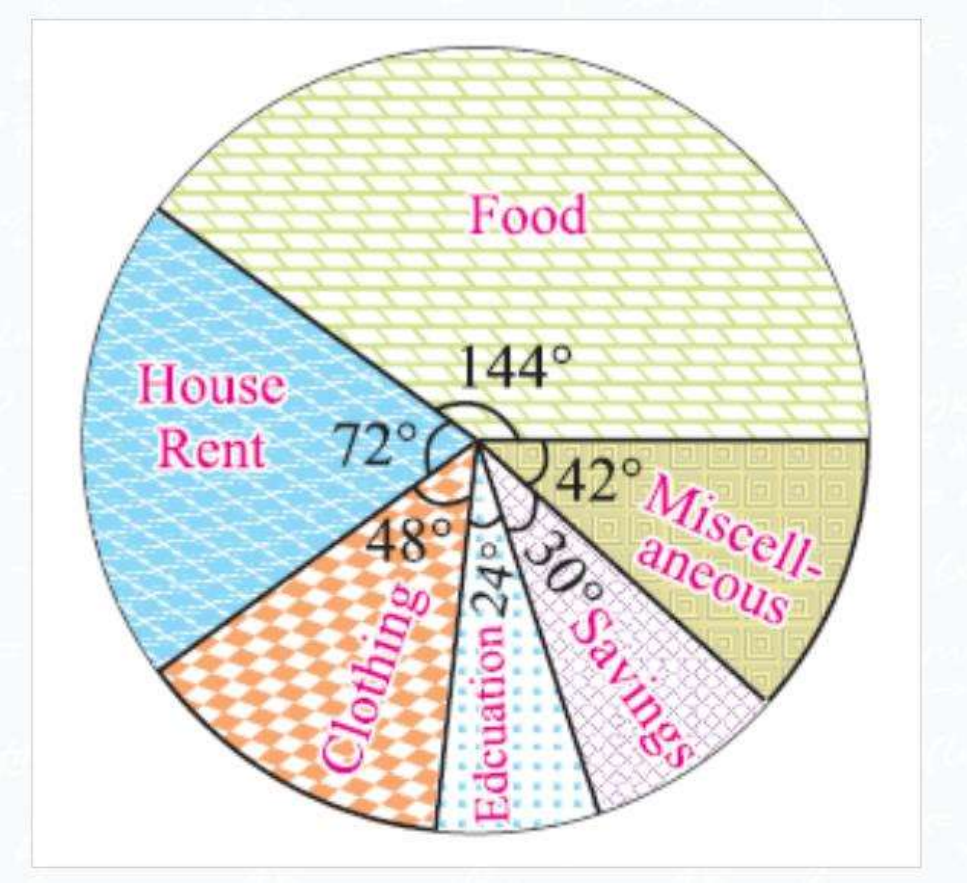
Particulars	Food	House Rent	Clothing	Education	Savings	Miscella-neous
Expenses (in $)	4800	2400	1600	800	1000	1400

Draw a pie chart to represent the above information.

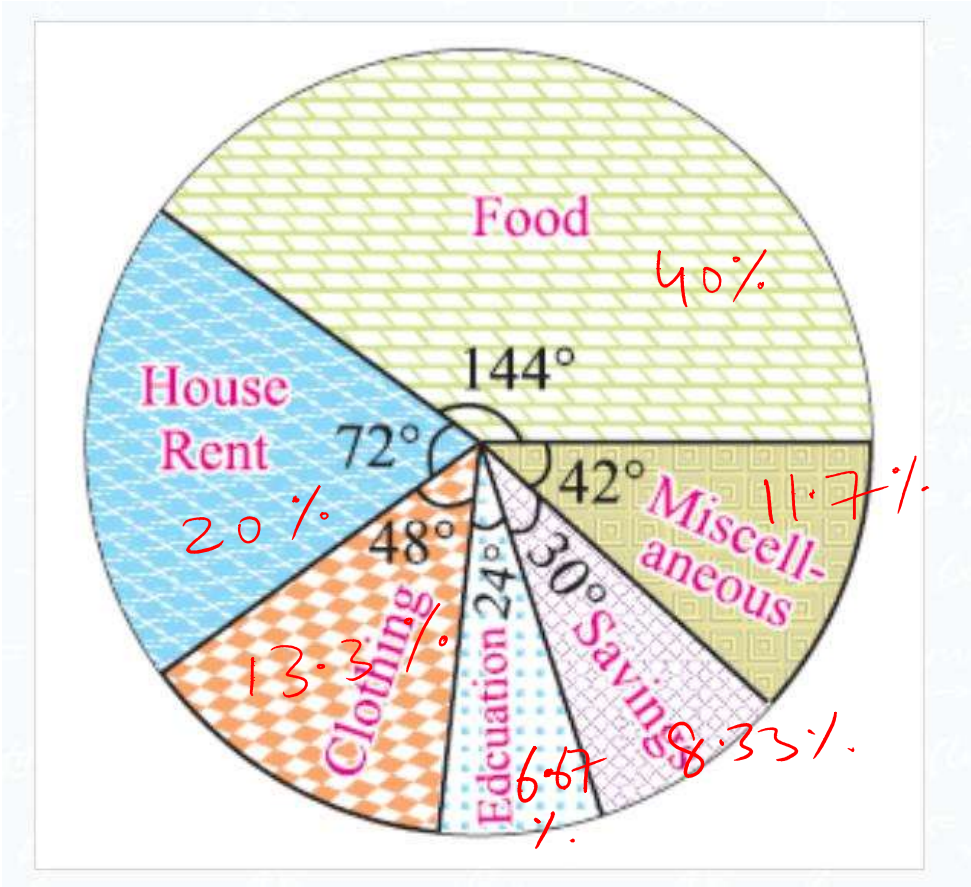
Solution

Particulars	Expenses ($)	Central angle
Food	4800	$\frac{4800}{12000} \times 360° = 144°$
House rent	2400	$\frac{2400}{12000} \times 360° = 72°$
Clothing	1600	$\frac{1600}{12000} \times 360° = 48°$
Education	800	$\frac{800}{12000} \times 360° = 24°$
Savings	1000	$\frac{1000}{12000} \times 360° = 30°$
Miscellaneous	1400	$\frac{1400}{12000} \times 360° = 42°$
Total	**12000**	**360°**

From the table, we obtain the required pie chart as shown below.



Particulars	Expenses ($)	%
Food	4800	$\frac{4800}{12000} \times 100 = 40$
House rent	2400	$\frac{2400}{12000} \times 100 = 20$
Clothing	1600	$\frac{1600}{12000} \times 100 = 13.3$
Education	800	6.67
Savings	1000	8.33
Miscellaneous	1400	11.7
Total	**12000**	100


Food
40%
144°
House
Rent
72°
20%
42°
Miscell-
aneous
11.7%
48°
13.3%
Clothing
24°
Edcuation
6.67
%
30°
Savings
8.33%

Grouped data, histograms, ogives, and stem and leaf plots

Using a line or a bar graph to plot the frequencies of data values is often an effective way of portraying a data set. However, for some data sets the number of distinct values is too large to utilize this approach. Instead, in such cases, it is useful to divide the values into groupings, or class intervals, and then plot the number of data values falling in each class interval. The number of class intervals chosen should be a trade-off between (1) choosing too few classes at a cost of losing too much information about the actual data values in a class and (2) choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible.

The endpoints of a class interval are called the class boundaries. We will adopt the left end inclusion convention, which stipulates that a class interval contains its left-end but not its right-end boundary point. Thus, for instance, the class interval 20–30 contains all values that are both greater than or equal to 20 and less than 30.

Table 2 (on next slide) presents the lifetimes of 200 incandescent lamps. A class frequency table for the data of Table 2 is presented in Table 3. The class intervals are of length 100, with the first one starting at 500.

Table 2 *Life in Hours of 200 Incandescent Lamps*

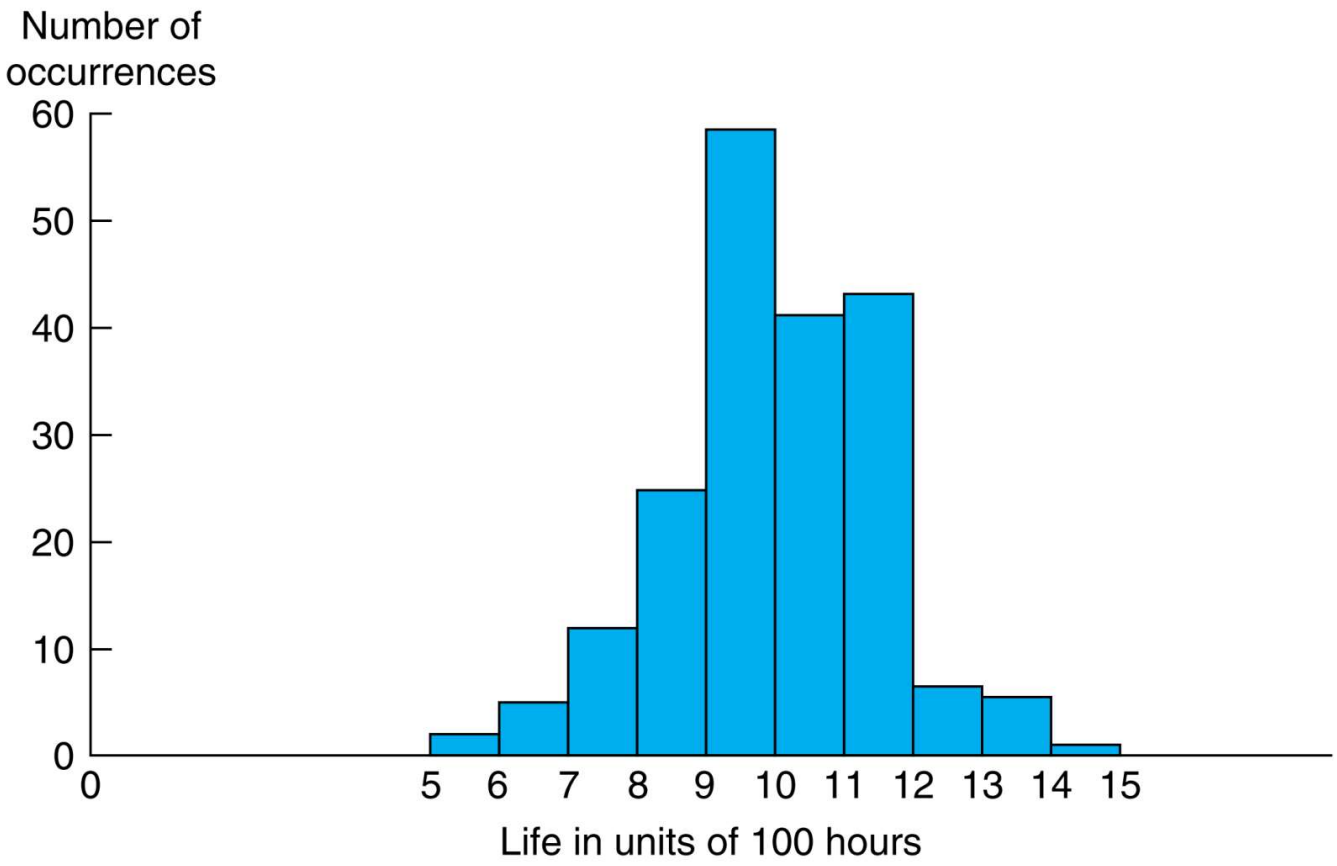
Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Table 3 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

A bar graph plot of class data, with the bars placed adjacent to each other, is called a histogram. The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a frequency histogram and in the latter a relative frequency histogram.

A frequency histogram



An efficient way of organizing a small- to moderate-sized data set is to utilize a stem and leaf plot. Such a plot is obtained by first dividing each data value into two parts —its stem and its leaf. For instance, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit. Thus, for instance, the value 62 is expressed as

Stem	Leaf
6	2

and the two data values 62 and 67 can be represented as

Stem	Leaf
6	2, 7

Example

Table 4 (on next slide) presents the per capita personal income for each of the 50 states and the District of Columbia. We can represent the data by a stem and leaf plot.

Table 4 Capita Personal Income (Dollars per Person), 2002

State name		State name		State name	
United States	30,941	Kentucky	25,579	Ohio	29,405
Alabama	25,128	Louisiana	25,446	Oklahoma	25,575
Alaska	32,151	Maine	27,744	Oregon	28,731
Arizona	26,183	Maryland	36,298	Pennsylvania	31,727
Arkansas	23,512	Massachusetts	39,244	Rhode Island	31,319
California	32,996	Michigan	30,296	South Carolina	25,400
Colorado	33,276	Minnesota	34,071	South Dakota	26,894
Connecticut	42,706	Mississippi	22,372	Tennessee	27,671
Delaware	32,779	Missouri	28,936	Texas	28,551
District of Columbia	42,120	Montana	25,020	Utah	24,306
Florida	29,596	Nebraska	29,771	Vermont	29,567
Georgia	28,821	Nevada	30,180	Virginia	32,922
Hawaii	30,001	New Hampshire	34,334	Washington	32,677
Idaho	25,057	New Jersey	39,453	West Virginia	23,688
Illinois	33,404	New Mexico	23,941	Wisconsin	29,923
Indiana	28,240	New York	36,043	Wyoming	30,578
Iowa	28,280	North Carolina	27,711		
Kansas	29,141	North Dakota	26,982		

The data presented in Table 4 are represented in the following stem-and-leaf plot. Note that the values of the leaves are put in the plot in increasing order.

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575, 57
26	183, 894, 982
27	671, 711, 744
28	240, 280, 551, 731, 821, 936
29	141, 405, 567, 596, 771, 923
30	001, 180, 296, 578
31	319, 727
32	151, 677, 779, 922, 996
33	276, 404
34	071, 334
36	043, 298
39	244, 453
42	120, 706

Example

The following stem-and-leaf plot represents the weights of 80 attendees at a sporting convention. The stem represents the tens digit and the hundred digits, and the leaves are the ones digit.

10	2, 3, 3, 4, 7	(5)
11	0, 1, 2, 2, 3, 6, 9	(7)
12	1, 2, 4, 4, 6, 6, 6, 7, 9	(9)
13	1, 2, 2, 5, 5, 6, 6, 8, 9	(9)
14	0, 4, 6, 7, 7, 9, 9	(7)
15	1, 1, 5, 6, 6, 6, 7	(7)
16	0, 1, 1, 1, 2, 4, 5, 6, 8, 8	(10)
17	1, 1, 3, 5, 6, 6, 6	(7)
18	1, 2, 2, 5, 5, 6, 6, 9	(8)
19	0, 0, 1, 2, 4, 5	(6)
20	9, 9	(2)
21	7	(1)
22	1	(1)
23		(0)
24	9	(1)

The numbers in parentheses on the right represent the number of values in each stem class. These summary numbers are often useful. They tell us, for instance, that there are 10 values having stem 16; that is, 10 individuals have weights between 160 and 169. Note that a stem without any leaves (such as stem value 23) indicates that there are no occurrences in that class.

Summarizing data sets

Modern-day experiments often deal with huge sets of data. To obtain a feel for such a large amount of data, it is useful to be able to summarize it by some suitably chosen measures.

Sample mean, sample median, and sample mode

Let's introduce some statistics that are useful for describing the center of a set of data value.

Definition

The *sample mean*, designated by $\bar{x}$, is defined by

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

The computation of the sample mean can often be simplified by noting that if for constants a and b

$$y_i = ax_i + b, \qquad i = 1, \ldots, n$$

then the sample mean of the data set $y_1, \ldots, y_n$ is

$$\bar{y} = \sum_{i=1}^{n} (ax_i + b)/n = \sum_{i=1}^{n} ax_i/n + \sum_{i=1}^{n} b/n = a\bar{x} + b$$

Note: More often we use the term statistic for numerical quantity computed from a data set.

Consider the data

$x_1 = 16.0$	$x_2 = 30.5$	$x_3 = 17.7$	$x_4 = 17.5$	$x_5 = 14.1$
$x_6 = 10.0$	$x_7 = 15.6$	$x_8 = 15.0$	$x_9 = 19.1$	$x_{10} = 17.9$
$x_{11} = 18.9$	$x_{12} = 18.5$	$x_{13} = 12.2$	$x_{14} = 6.0$	

Figure 1.14 shows a dotplot of the data; a water-absorption percentage in the mid-teens appears to be "typical." With $\Sigma x_i = 229.0$, the sample mean is

$$\bar{x} = \frac{229.0}{14} = 16.36$$

A physical interpretation of the sample mean demonstrates how it assesses the center of a sample. Think of each dot in the dotplot as representing a 1 kg weight. Then a fulcrum placed with its tip on the horizontal axis will balance precisely when it is located at x. So the sample mean can be regarded as the balance point of the distribution of observations.

Problem

The winning scores in the U.S. Masters golf tournament in the years from 1999 to 2008 were as follows:

280, 278, 272, 276, 281, 279, 276, 281, 289, 280

Find the sample mean of these scores.

SOLUTION Rather than directly adding these values, it is easier to first subtract 280 from each one to obtain the new values $y_i = x_i - 280$:

$$0, -2, -8, -4, 1, -1, -4, 1, 9, 0$$

Because the arithmetic average of the transformed data set is

$$\bar{y} = -8/10$$

it follows that

$$\bar{x} = \bar{y} + 280 = 279.2$$

Sometimes we want to determine the sample mean of a data set that is presented in a frequency table listing the k distinct values $v_1, \ldots, v_k$ having corresponding frequencies $f_1, \ldots, f_k$. Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, with the value v_i appearing f_i times, for each $i = 1, \ldots, k$, it follows that the sample mean of these n data values is

$$\bar{x} = \sum_{i=1}^{k} v_i f_i / n$$

In expanded form, we have

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \cdots + \frac{f_k}{n} v_k$$

Problem

The number of iPhone sold daily by a small company for the past 6 days has been arranged in the following frequency table:

Value	Frequency
3	2
4	1
5	3

What is the sample mean?

Solution Since the original data set consists of the 6 values

$$3, 3, 4, 5, 5, 5$$

it follows that the sample mean is

$$\bar{x} = \frac{3 + 3 + 4 + 5 + 5 + 5}{6}$$

$$= \frac{3 \times 2 + 4 \times 1 + 5 \times 3}{6} = \boxed{\frac{25}{6}}$$

Another statistic used to indicate the center of a data set is the sample median; loosely speaking, it is the middle value when the data set is arranged in increasing order.

Definition

Order the values of a data set of size n from smallest to largest. If n is odd, the *sample median* is the value in position $(n+1)/2$; if n is even, it is the average of the values in positions $n/2$ and $n/2+1$.

Thus the sample median of a set of three values is the second smallest; of a set of four values, it is the average of the second and third smallest.

Order the data values from smallest to largest. If the number of data values is odd, then the sample median is the middle value in the ordered list; if it is even, then the sample median is the average of the two middle values.

Problem

The following data represent the number of weeks it took seven individuals to obtain their driver's licenses. Find the sample median.

$$2, 110, 5, 7, 6, 7, 3$$

Solution

First arrange the data in increasing order.

$$2, 3, 5, 6, 7, 7, 110$$

Since the sample size is 7, it follows that the sample median is the fourth-smallest value. That is, the sample median number of weeks it took to obtain a driver's license is $m = 6$ weeks.

Problem

The following data represent the number of days it took 6 individuals to quit smoking after completing a course designed for this purpose.

$$1, 2, 3, 5, 8, 100$$

What is the sample median?

Solution

Since the sample size is 6, the sample median is the average of the two middle values; thus,

$$m = \frac{3 + 5}{2} = 4$$

That is, the sample median is 4 days.

Comparison of Mean with Median

The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean, being the arithmetic average, makes use of all the data values. The sample median, which makes use of only one or two middle values, is not affected by extreme values.

Mode

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

If no number in a set of numbers occurs more than once, that set has no mode.
A set of numbers with two modes is **bimodal**, a set of numbers with three modes is **trimodal**, and any set of numbers with more than one mode is **multimodal**.

Problem

The following frequency table gives the values obtained in 40 rolls of a die.

Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Find **(a)** the sample mean, **(b)** the sample median, and **(c)** the sample mode.

Solution

(a) The sample mean is

$$\bar{x} = (9 + 16 + 15 + 20 + 30 + 42)/40 = 3.05$$

(b) The sample median is the average of the 20th and 21st smallest values, and is thus equal to 3. **(c)** The sample mode is 1, the value that occurred most frequently.

Consider the following data sets

$$A: 1, 2, 5, 6, 6 \qquad B: -40, 0, 5, 20, 35$$

Although both data sets have the same sample mean and sample median there is a much greater variability or spread in the values of Set **B** than in the set **A**. For measuring variability in data we have statistics sample variance and standard deviation which we are discussing in coming slides.

Sample Variance

The *sample variance*, call it s^2, of the data set $x_1, \ldots, x_n$ is defined by

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

Just as $\bar{x}$ will be used to make inferences about the population mean μ, we should define the sample variance so that it can be used to make inferences about σ^2. Now note that σ^2 involves squared deviations about the population mean μ. If we actually knew the value of μ, then we could define the sample variance as the average squared deviation of the sample x_i's about μ. However, the value of μ is almost never known, so the sum of squared deviations about $\bar{x}$ must be used. But *the* x_i*'s tend to be closer to their average* $\bar{x}$ *than to the population average* μ. To compensate for this, the divisor $n - 1$ is used rather than the sample size n. In other words, if we used a divisor n in the sample variance, then the resulting quantity would tend to underestimate σ^2 (produce estimated values that are too small on the average), whereas dividing by the slightly smaller $n - 1$ corrects this underestimating.

Problem

Find the sample variance of data set A: 1, 2, 5, 6, 6

Solution

It is determined as follows:

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	−3	−2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

Hence, for data set A,

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{4} = 5.5$$

Problem

Find the sample variance of data set B: $-40, 0, 5, 20, 35$

Solution

The sample mean for data set B is also $\bar{x} = 4$. Therefore, for this set, we have

x_i	-40	0	5	20	35
$x_i - \bar{x}$	-44	-4	1	16	31
$(x_i - \bar{x})^2$	1936	16	1	256	961

Thus,

$$s^2 = \frac{3170}{4} = 792.5$$

The following algebraic identity is useful for computing the sample variance by hand:

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

The identity is proven as follows:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\
&= \sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\\
&= \sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\\
&= \sum_{i=1}^{n}x_i^2-n\bar{x}^2
\end{aligned}$$

The sample variance remains unchanged when a constant is added to each data value.

The above statement make sense because adding or subtracting c shifts the location of the data set but leaves distances between data values unchanged.

Problem

The following data give the yearly numbers of law enforcement officers killed in the United States over 10 years:

164, 165, 157, 164, 152, 147, 148, 131, 147, 155

Find the sample variance of the number killed in these years.

Solution

Rather than working directly with the given data, let us subtract the value 150 from each data item. (That is, we are adding $c = -150$ to each data value.) This results in the new data set

$$14, 15, 7, 14, 2, -3, -2, -19, -3, 5$$

Its sample mean is

$$\bar{y} = \frac{14 + 15 + 7 + 14 + 2 - 3 - 2 - 19 - 3 + 5}{10} = 3.0$$

The sum of the squares of the new data is

$$\sum_{i=1}^{10} y_i^2 = 14^2 + 15^2 + 7^2 + 14^2 + 2^2 + 3^2 + 2^2 + 19^2 + 3^2 + 5^2 = 1078$$

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 1078 - 10(9) = 988$$

Hence, the sample variance of the revised data, which is equal to the sample variance of the original data, is

$$s^2 = \frac{988}{9} \approx \boxed{109.78}$$

Sample Standard Deviation

The quantity s, defined by

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

is called the *sample standard deviation*.

The sample standard deviation is measured in the same units as the data.

It may appear to the reader that the use of both the sample variance and the sample standard deviation is redundant. Both measures reflect the same concept in measuring variability, but the sample standard deviation measures variability in linear units whereas the sample variance is measured in squared units. For instance, if the data are in feet, then the sample variance will be expressed in units of square feet and the sample standard deviation in units of feet.

Note

If we shift the measurements by adding or subtracting a constant, then the measure of center gets shifted by the same amount, but the measure of variation is unaffected by any shift in measurements.

Sample Percentiles and Box Plots

The *sample 100*p *percentile* is that data value such that $100p$ percent of the data are less than or equal to it and $100(1 - p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

Method for finding Sample Percentile

- First arrange the data in increasing order.
- If np is not an integer, then the data value whose position is the smallest integer exceeding np is the sample 100p percentile.
- On the other hand, if np is an integer, then the sample 100p percentile is the average of the values in positions np and np + 1.

Problem

Table 2.6 lists the populations of the 25 most populous U.S. cities for the year 1994.

For this data set, find (a) the sample 10 percentile and (b) the sample 80 percentile.

TABLE 2.6 *Population of 25 Largest U.S. Cities, July 2006*

Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951

9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

Solution

(a) Because the sample size is 25 and 25(.10) = 2.5, the sample 10 percentile is the third smallest value, equal to 590, 763.

(b) Because 25(.80) = 20, the sample 80 percentile is the average of the twentieth and the twenty-first smallest values. Hence, the sample 80 percentile is

$$\frac{1{,}512{,}986 + 1{,}448{,}394}{2} = 1{,}480{,}690$$

The sample 50 percentile is, of course, just the sample median. Along with the sample 25 and 75 percentiles, it makes up the sample quartiles.

Quartiles

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the sample median or the *second quartile*; the sample 75 percentile is called the *third quartile*.

The quartiles break up a data set into four parts, with roughly 25 percent of the data being less than the first quartile, 25 percent being between the first and second quartile, 25 percent being between the second and third quartile, and 25 percent being greater than the third quartile.

Problem

Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB.

The following data give noise levels measured at 36 different times directly outside of Grand Central Station in New York.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85

69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

Solution A stem and leaf plot of the data is as follows:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

The first quartile is 76, the average of the 9th and 10th smallest data values; the second quartile is 89.5, the average of the 18th and 19th smallest values; the third quartile is 104.5, the average of the 27th and 28th smallest values.

Box Plot

A box plot is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line.

Table 2.1 Starting Yearly Salaries.

Starting Salary	Frequency
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
66	2
67	3
70	1

For instance, the 42 data values presented in Table 2.1 go from a low value of 57 to a high value of 70. The value of the first quartile (equal to the value of the 11th smallest on the list) is 60; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 61.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 64. The box plot for this data set is shown in Figure 2.7.

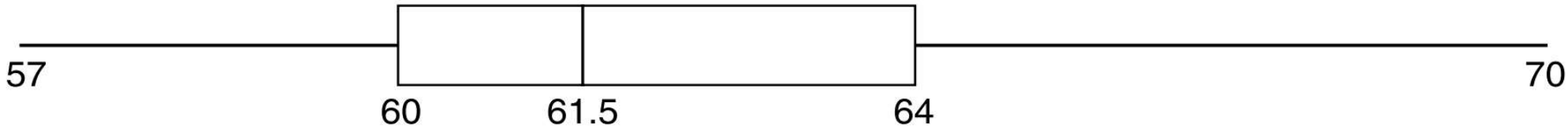


FIGURE 2.7

The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the **range of the data.** Also, the length of the box itself, equal to the third quartile minus the first quartile, is called the **interquartile range**.

Normal Data Set

Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be *normal* and their histograms are called *normal histograms*. Figure 2.8 is the histogram of a normal data set.

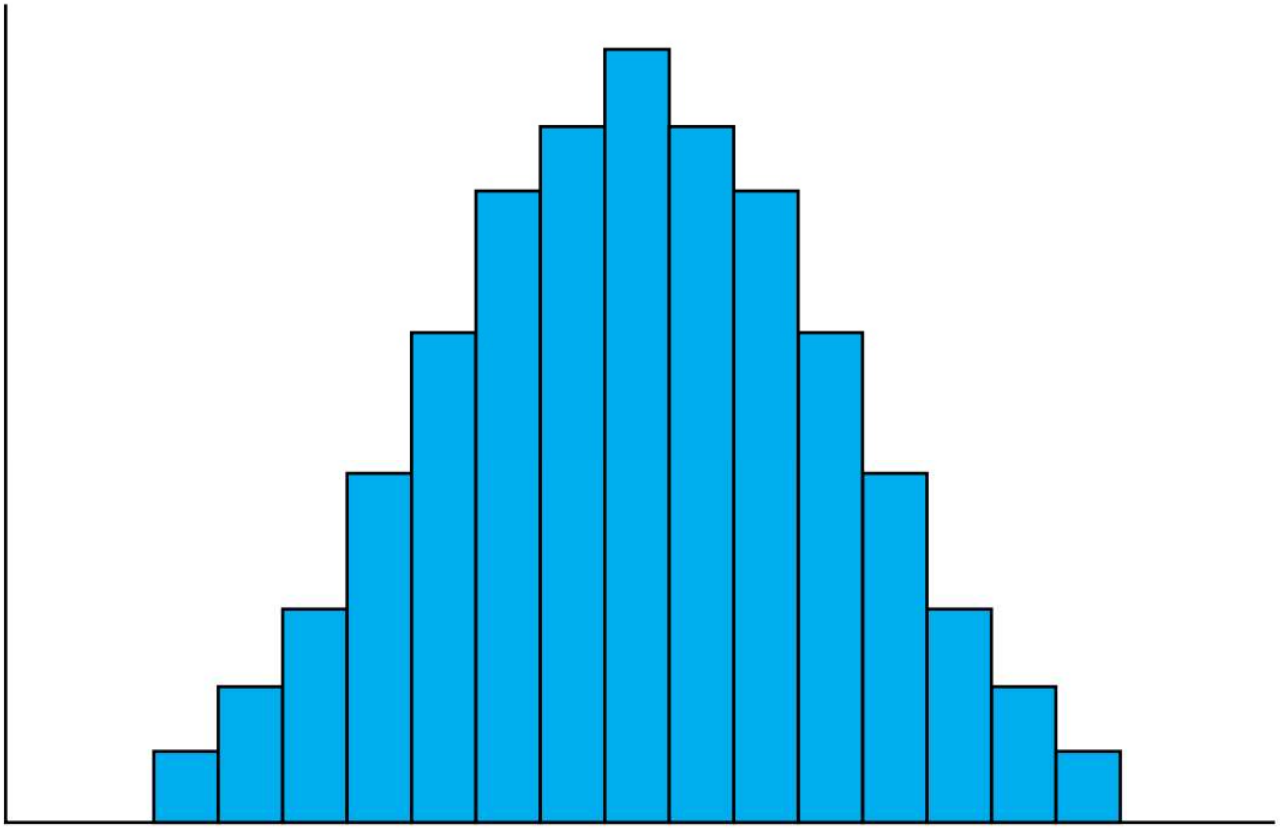


FIGURE 2.8 *Histogram of a normal data set.*

Definition *A data set is said to be* normal *if a histogram describing it has the following properties:*

1. *It is highest at the middle interval.*
2. *Moving from the middle interval in either direction, the height decreases in such a way that the entire histogram is bell-shaped.*
3. *The histogram is symmetric about its middle interval.*

If the histogram of a data set is close to being a normal histogram, then we say that the data set is approximately normal. For instance, we would say that the histogram given in Figure 2.9 is from an approximately normal data set, whereas the ones presented in Figures 2.10 and 2.11 are not (because each is too nonsymmetric).

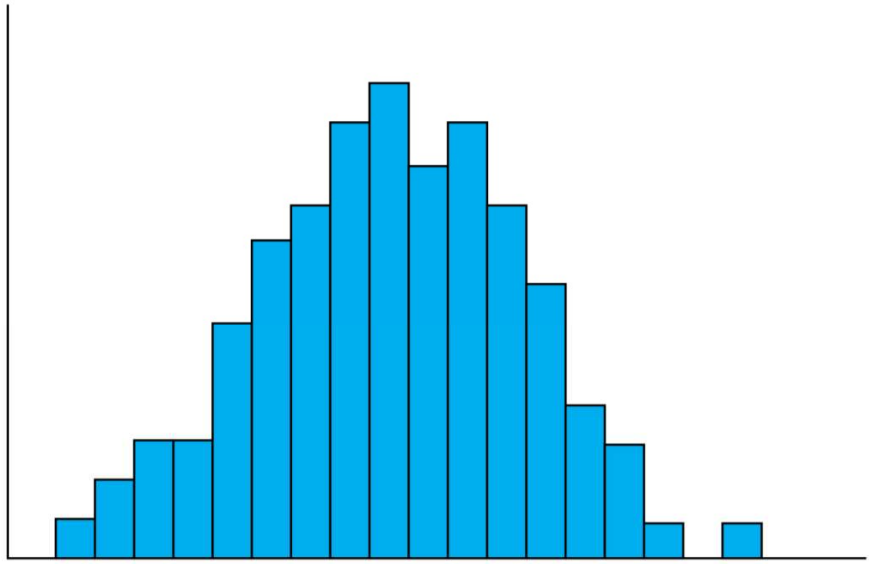


FIGURE 2.9 *Histogram of an approximately normal data set.*

Any data set that is not approximately symmetric about its sample median is said to be skewed. It is "skewed to the right" if it has a long tail to the right and "skewed to the left" if it has a long tail to the left. Thus the data set presented in Figure 2.10 is skewed to the left and the one of Figure 2.11 is skewed to the right.

It follows from the symmetry of the normal histogram that a data set that is approximately normal will have its sample mean and sample median approximately equal.

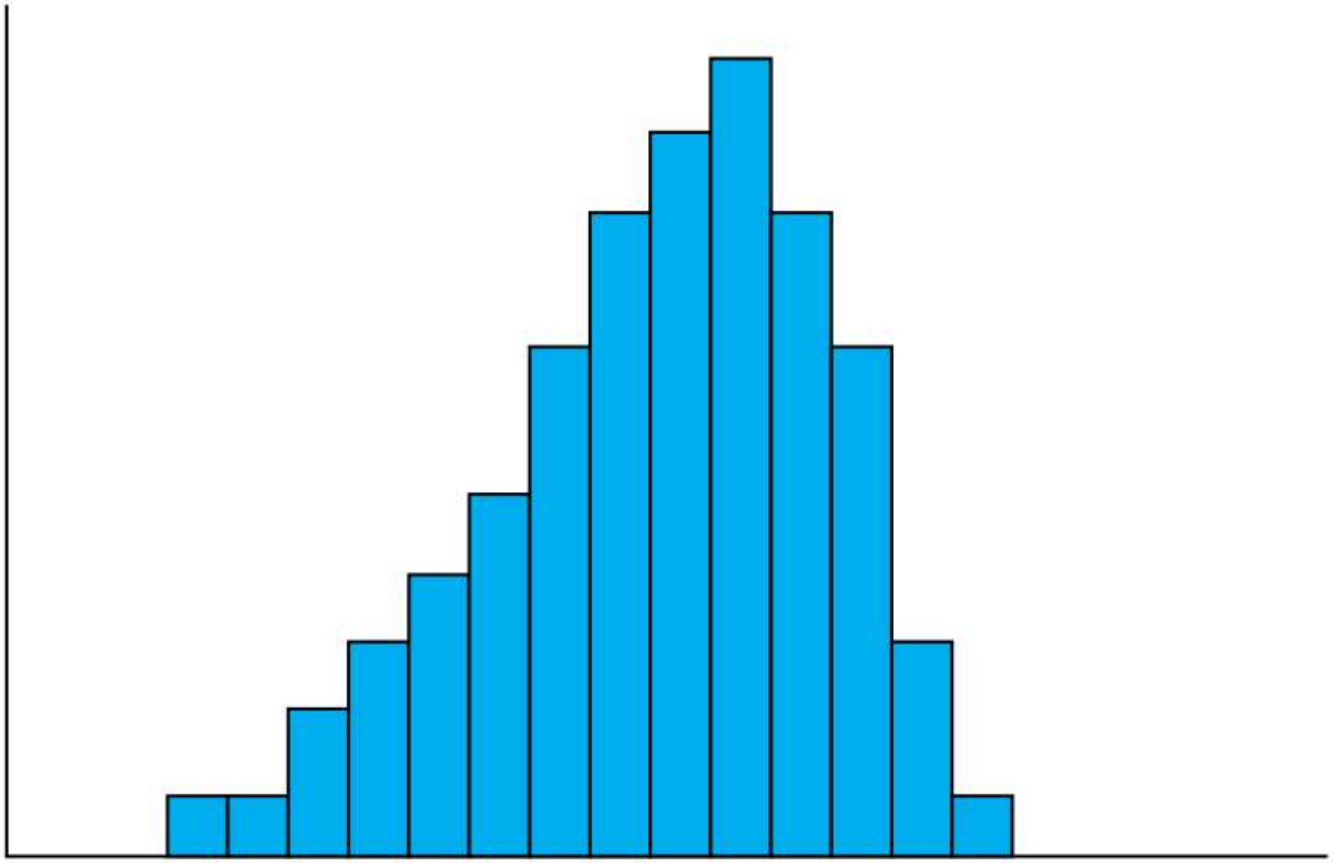


FIGURE 2.10 *Histogram of a data set skewed to the left.*

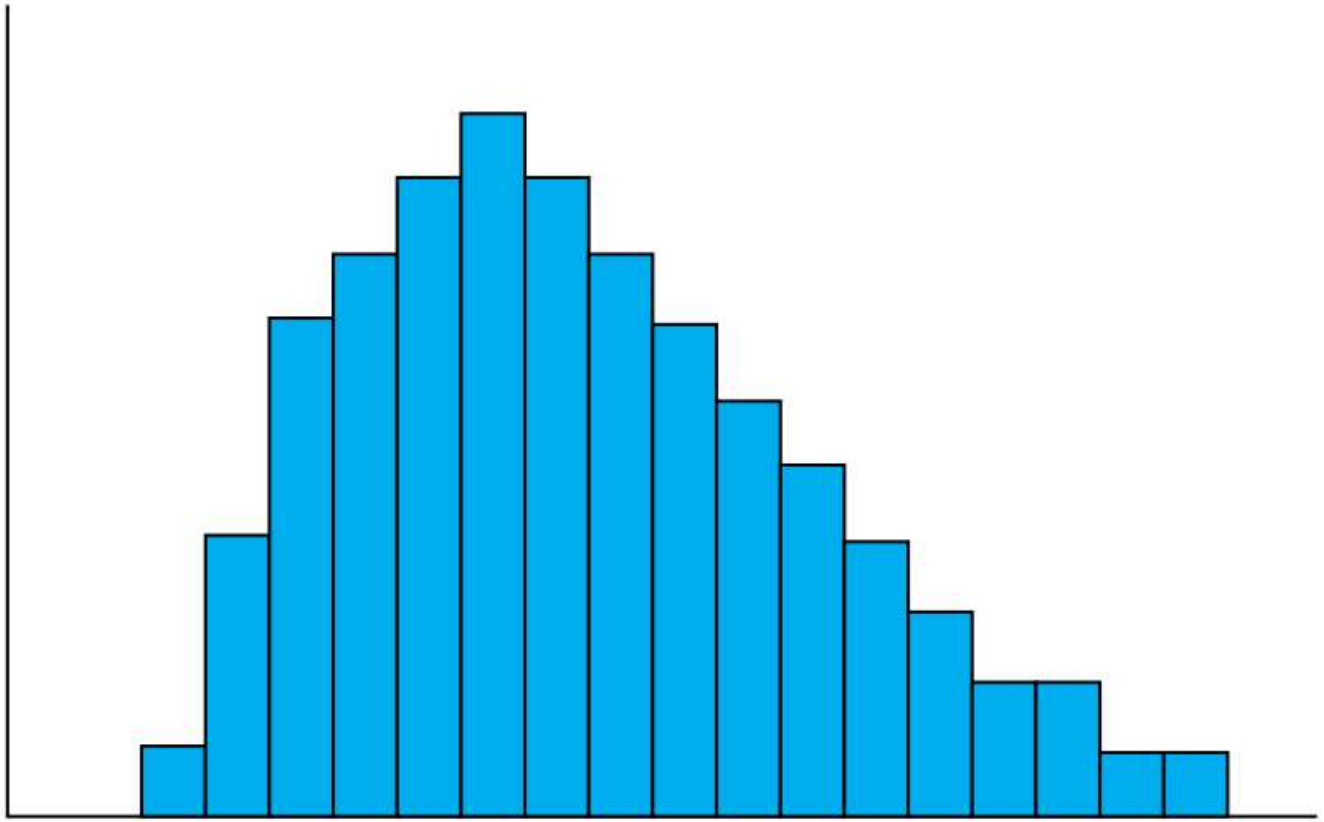


FIGURE 2.11 *Histogram of a data set skewed to the right.*

The Empirical Rule

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then the following statements are true.

1. Approximately 68 percent of the observations lie within

$$\bar{x} \pm s$$

2. Approximately 95 percent of the observations lie within

$$\bar{x} \pm 2s$$

3. Approximately 99.7 percent of the observations lie within

$$\bar{x} \pm 3s$$

Problem

The scores of 25 students on a history examination are listed on the following stem-and-leaf plot.

9	0, 0, 4
8	3, 4, 4, 6, 6, 9
7	0, 0, 3, 5, 5, 8, 9
6	2, 2, 4, 5, 7
5	0, 3, 5, 8

By standing the stem and leaf plot on its side we can see that the corresponding histogram is approximately normal. Use it to assess the empirical rule.

Solution

A calculation yields that the sample mean and sample standard deviation of the data are

$$\bar{x} = 73.68 \quad \text{and} \quad s = 12.80$$

The empirical rule states that approximately 68 percent of the data values are between $\bar{x} - s = 60.88$ and $\bar{x} + s = 86.48$. Since 17 of the observations actually fall within 60.88 and 86.48, the actual percentage is $100(17/25) = 68$ percent.

Similarly, the empirical rule states that approximately 95 percent of the data are between $\bar{x} - 2s = 48.08$ and $\bar{x} + 2s = 96.28$, whereas, in actuality, 100 percent of the data fall in this range.

Paired data sets and the sample correlation coefficient

We are often concerned with data sets that consist of pairs of values that have some relationship to each other. If each element in such a data set has an x value and a y value, then we represent the ith data point by the pair (x_i, y_i).

For instance, in the data set presented in Table 2.12, x_i represents the score on an intelligence quotient (IQ) test, and y_i represents the annual salary (to the nearest \$1000) of the ith chosen worker in a sample of 30 workers from a particular company. Let's effectively display data sets of paired values.

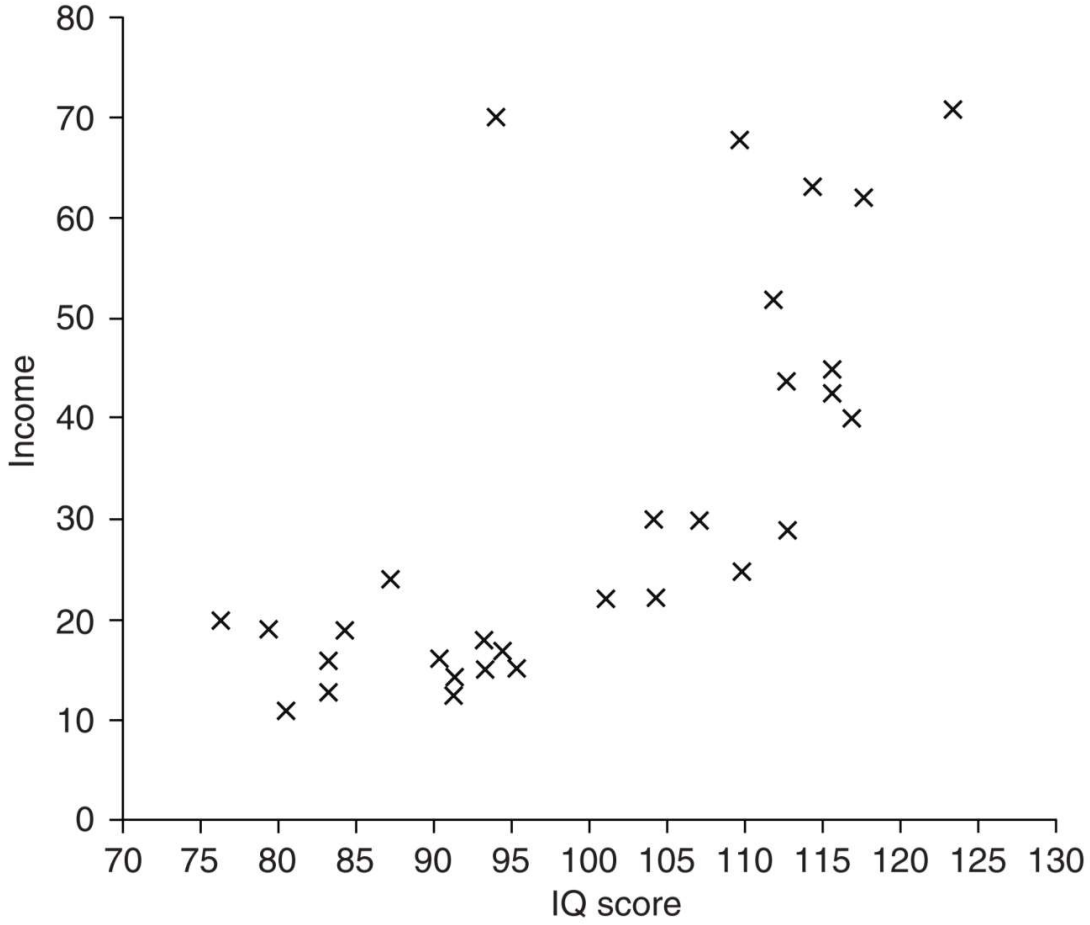
Table 2.12 Salaries versus IQ

Worker i	IQ score x_i	Annual salary y_i (in units of \$1000)	Worker i	IQ score x_i	Annual salary y_i (in units of \$1000)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

A useful way of portraying a data set of paired values is to plot the data on a two dimensional graph, with the x-axis representing the x value of the data and the y-axis representing the y value. Such a plot is called a scatter diagram. Figure 2.13 presents a scatter diagram for the data of Table 2.12.

Scatter diagram of IQ versus income data

FIGURE 2.13



It is clear from Fig. 2.13 that higher incomes appear to go along with higher scores on the IQ test. Note that, not every worker with a high IQ score receives a larger salary than another worker with a lower score (compare worker 5 with worker 29). The scatter diagram of Fig. 2.13 also appears to have some predictive uses. For instance, suppose we want to predict the salary of a worker whose IQ test score is 120. One way to do this is to "fit by eye" a line to the data set, as is done in Fig. 2.14. Since the y value on the line corresponding to the x value of 120 is about 45, this seems like a reasonable prediction for the annual salary of a worker whose IQ is 120.

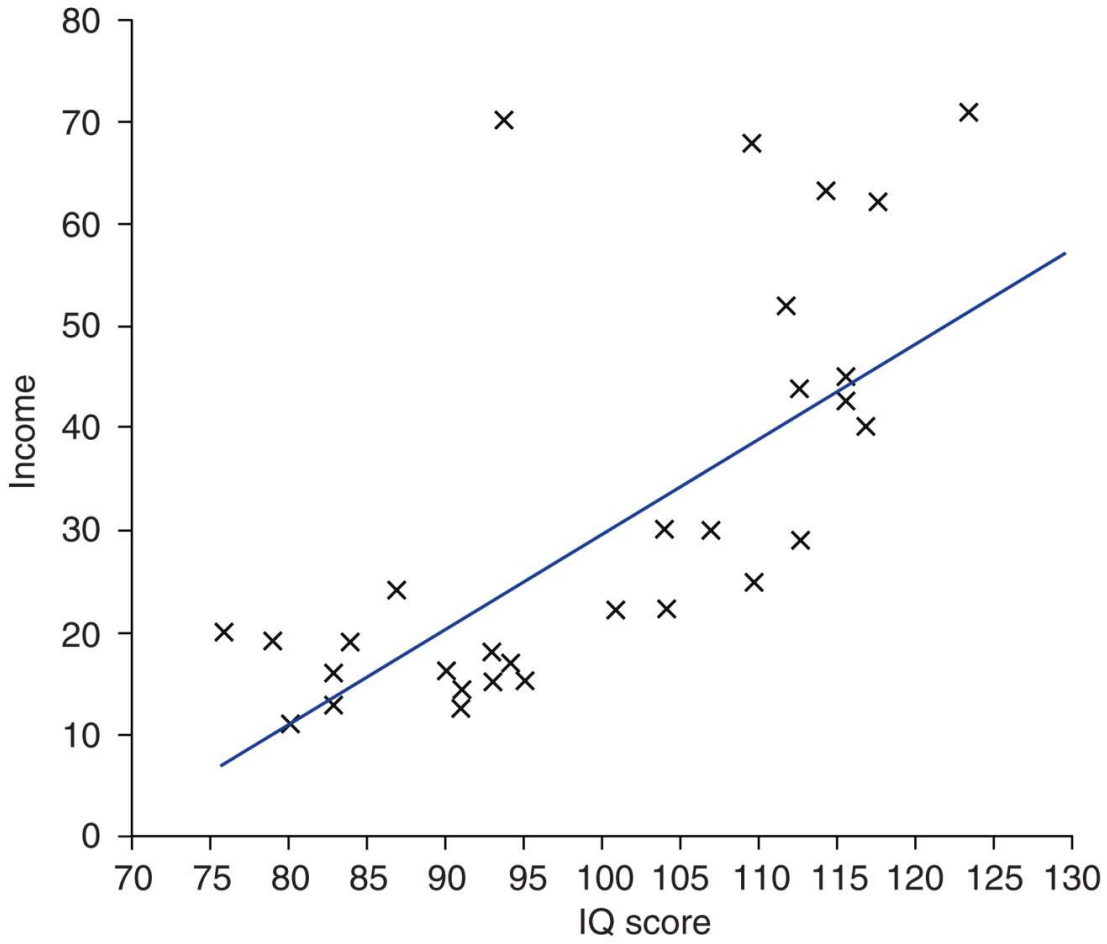


FIGURE 2.14

Scatter diagram for IQ versus income: fitting a straight line by eye.

Sample Correlation Coefficient

Let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values. The *sample correlation coefficient*, call it r, of the data pairs (x_i, y_i), $i = 1, \ldots, n$ is defined by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

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

When $r > 0$ we say that the sample data pairs are *positively correlated*, and when $r < 0$ we say that they are *negatively correlated.*

The following are properties of the sample correlation coefficient.

1. $$-1 \leq r \leq 1$$
2. If for constants a and b, with $b > 0$,

$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

then $r = 1$.

3. If for constants a and b, with $b < 0$,

$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

then $r = -1$.

4. If r is the sample correlation coefficient for the data pairs $x_i, y_i, i = 1, \ldots, n$ then it is also the sample correlation coefficient for the data pairs

$$a + bx_i, \quad c + dy_i, \quad i = 1, \ldots, n$$

provided that b and d are both positive or both negative.

Property 1 says that the sample correlation coefficient r is always between −1 and +1.

Property 2 says that r will equal +1 when there is a straight line (also called a linear) relation between the paired data such that large y values are attached to large x values.

Property 3 says that r will equal −1 when the relation is linear and large y values are attached to small x values.

Property 4 states that the value of r is unchanged when a constant is added to each of the x variables (or to each of the y variables) or when each x variable (or each y variable) is multiplied by a positive constant. This property implies that r does not depend on the dimensions chosen to measure the data. This property implies that r does not depend on the dimensions chosen to measure the data. For instance, the sample correlation coefficient between a person's height and weight does not depend on whether the height is measured in feet or in inches or whether the weight is measured in pounds or in kilograms.

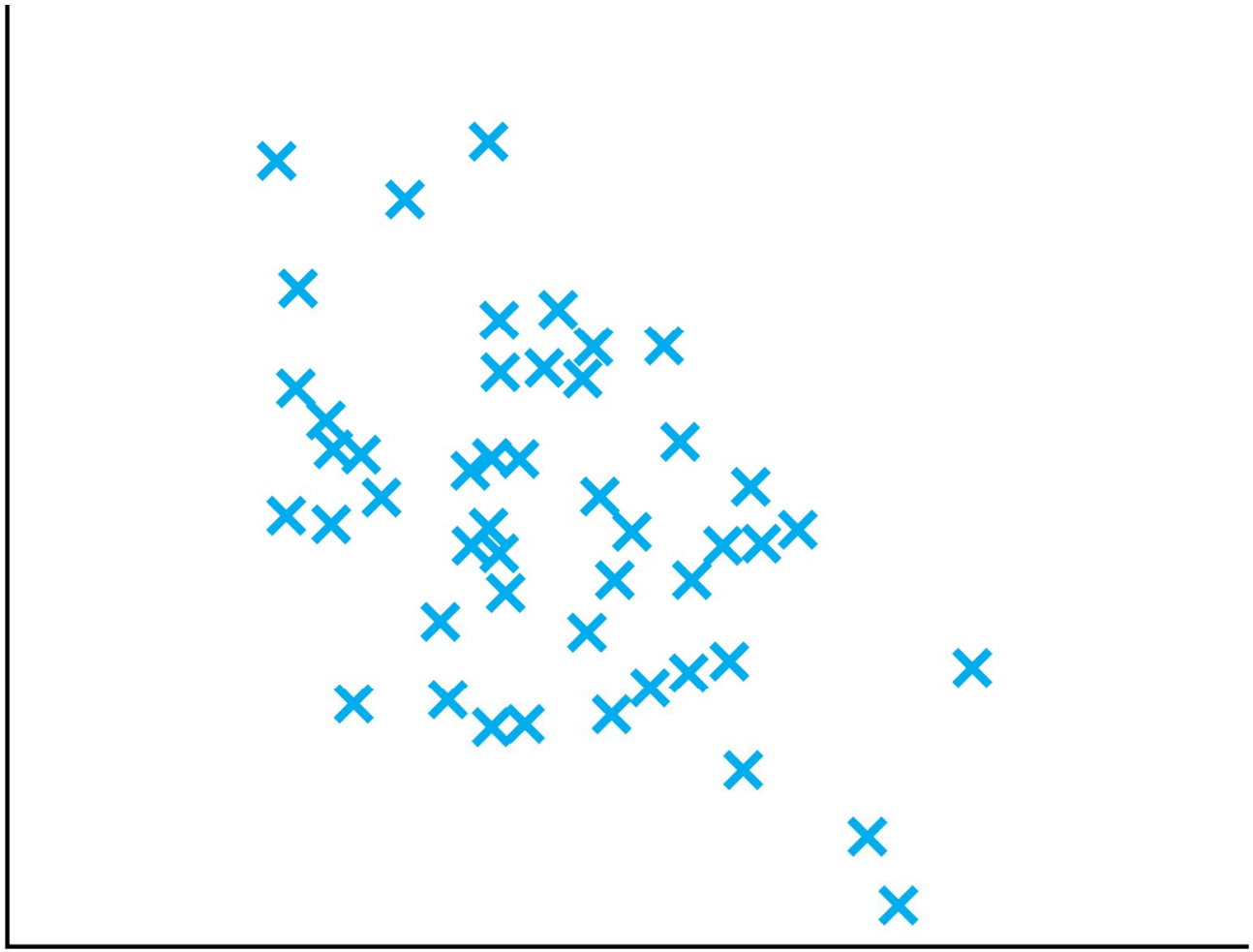
Also, if one of the values in the pair is temperature, then the sample correlation coefficient is the same whether it is measured in Fahrenheit or in Celsius.

Absolute Value of sample Correlation

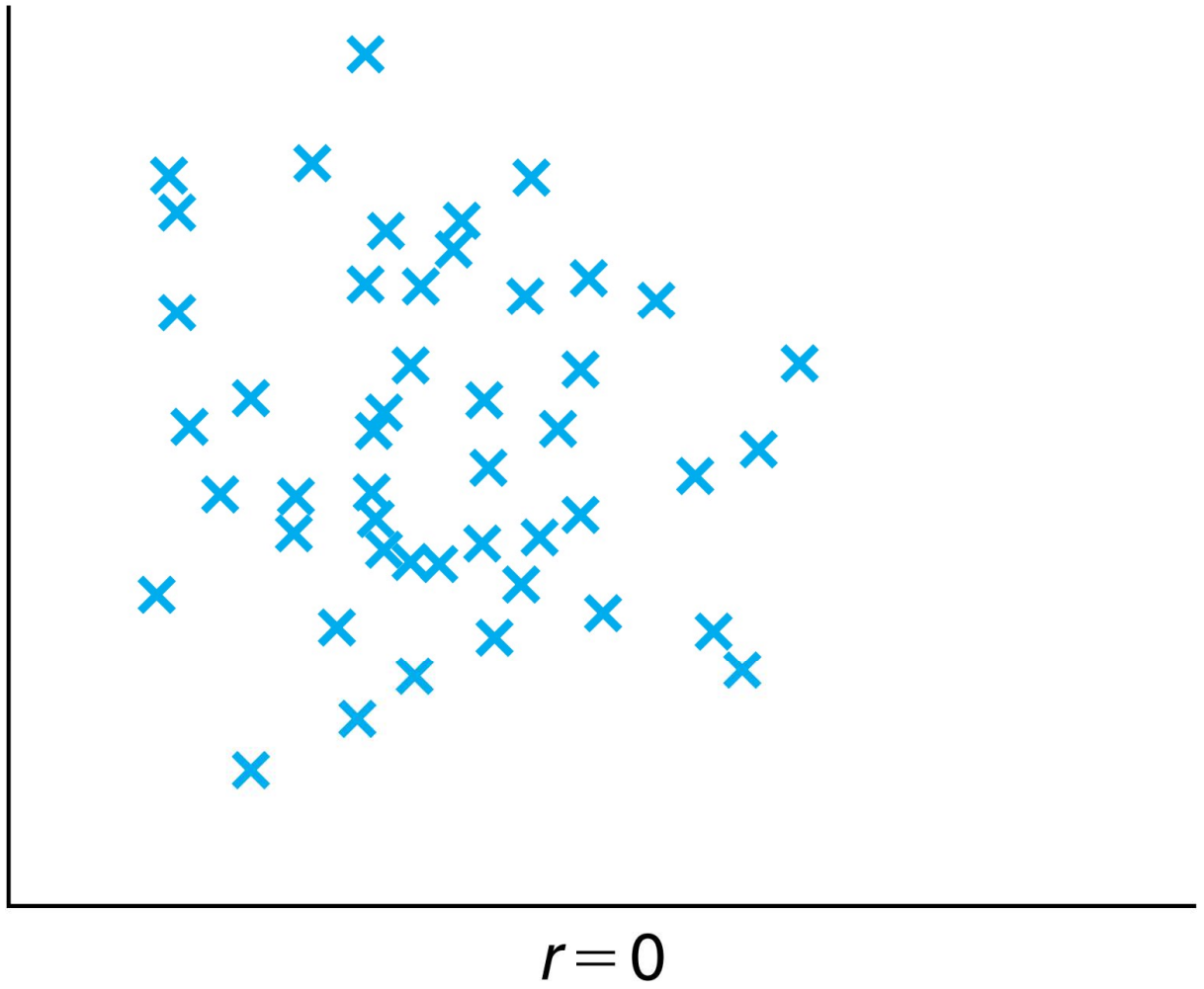
The absolute value of the sample correlation coefficient r (that is, $|r|$, its value without regard to its sign) is a measure of the strength of the linear relationship between the x and the y values of a data pair. A value of $|r|$ equal to 1 means that there is a perfect linear relation — that is, a straight line can pass through all the data points (x_i, y_i), $i = 1, \ldots, n$. A value of $|r|$ of around .8 means that the linear relation is relatively strong; although there is no straight line that passes through all of the data points, there is one that is "close" to them all. A value for $|r|$ of around .3 means that the linear relation is relatively weak.

Sign of "r"

The sign of r gives the direction of the relation. It is positive when the linear relation is such that smaller y values tend to go with smaller x values and larger y values with larger x values (and so a straight line approximation points upward), and it is negative when larger y values tend to go with smaller x values and smaller y values with larger x values (and so a straight line approximation points downward). Figure 2.14 displays scatter diagrams for data sets with various values of r.



$r = -.50$


$r = 0$

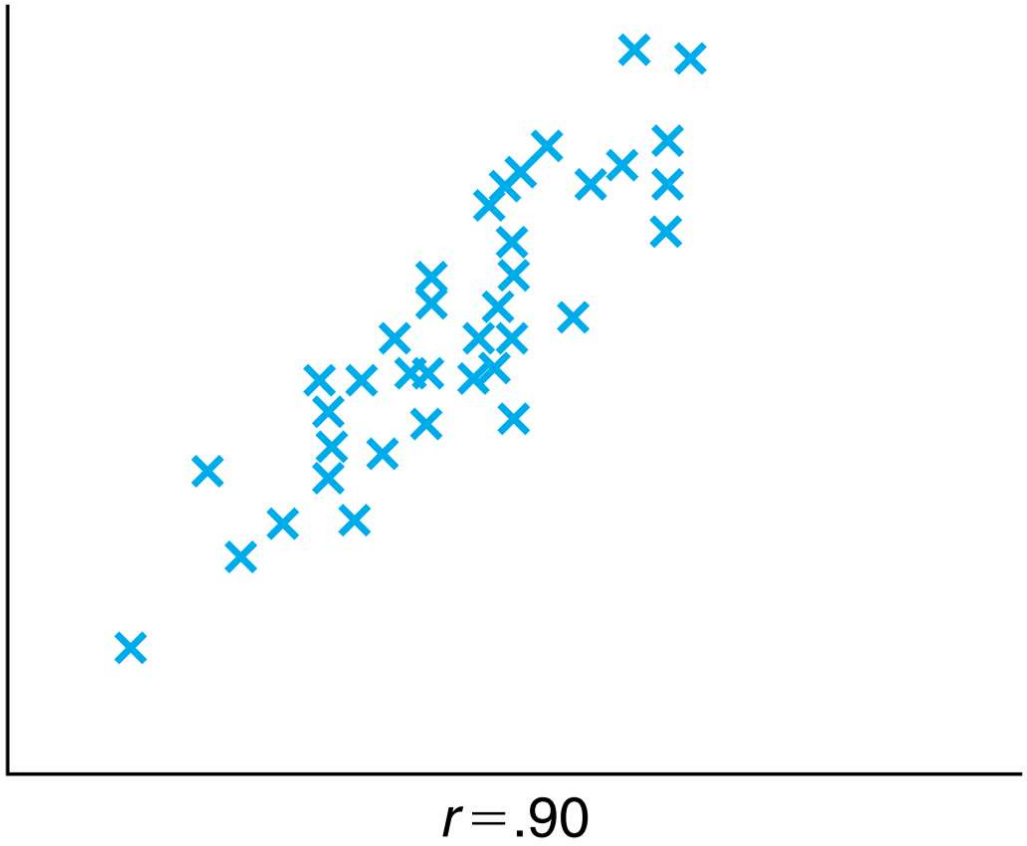


FIGURE 2.14 *Sample correlation coefficients.*

For computational purposes, the following is a convenient formula for the sample correlation coefficient.

Computational Formula for r

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Problem

The following table gives the U.S. per capita consumption of whole milk (x) and of low-fat milk (y) in three different years.

	Per capita consumption (gallons)		
	1980	1984	1988
Whole milk (x)	17.1	14.7	12.8
Low-fat milk (y)	10.6	11.5	13.2

Source: U.S. Department of Agriculture, *Food Consumption, Prices, and Expenditures.*

Find the sample correlation coefficient r for the given data.

Solution

To make the computation easier, let us first subtract 12.8 from each of the x values and 10.6 from each of the y values. This gives the new set of data pairs:

	i		
	1	**2**	**3**
x_i	4.3	1.9	0
y_i	0	0.9	2.6

Now, $\bar{x} = \dfrac{4.3 + 1.9 + 0}{3} = 2.0667$

$\bar{y} = \dfrac{0 + 0.9 + 2.6}{3} = 1.1667$

$\sum_{i=1}^{3} x_i y_i = (1.9)(0.9) = 1.71$ $\sum_{i=1}^{3} x_i^2 = (4.3)^2 + (1.9)^2 = 22.10$

$\sum_{i=1}^{3} y_i^2 = (0.9)^2 + (2.6)^2 = 7.57$

Thus,

$$r = \frac{1.71 - 3(2.0667)(1.1667)}{\sqrt{[22.10 - 3(2.0667)^2][7.57 - 3(1.1667)^2]}} = -0.97$$

Therefore, our three data pairs exhibit a very strong negative correlation between consumption of whole and of low-fat milk.

Practice Problem

The following are the number of traffic deaths in a sample of states, both for 2007 and 2008. Plot a scatter diagram and find the sample correlation coefficient for the data pairs.

2007 and 2008 Traffic Fatalities per State

State	2007	2008
WY	149	159
IL	1248	1044
MA	434	318
NJ	724	594
MD	615	560
OR	452	414
WA	568	504
FL	3221	2986
UT	291	271
NH	129	139