CHAPTER 11 Introduction to Statistics

Statistics is a branch of mathematics that deals with the collection and summarization of data. Methods of statistical analysis make it possible to draw conclusions about a population based on data from a sample of the population. Statistical models have become increasingly useful in manufacturing, government, agriculture, medicine, and the social sciences, and in all types of research. An Indianapolis race-car team is using statistics to improve performance by gathering data on each run around the track. They sample data 300 times a second, and use computers to process the data. In this chapter we give a brief introduction to some of the key topics from statistical theory.

In the previous chapter, we saw that a frequency distribution can be transformed into a probability distribution by using the relative frequency of each value of the random variable as its probability. Sometimes it is convenient to work directly with a frequency distribution.

11.1 FREQUENCY DISTRIBUTIONS; MEASURES OF CENTRAL TENDENCY

Often, a researcher wishes to learn something about a characteristic of a population, but because the population is very large or mobile, it is not possible to examine all of its elements. Instead, a limited sample drawn from the population is studied to determine the characteristics of the population. For example, a book by Frances Cerra Whittelsey, *Why Women Pay More,* published by Ralph Nader's Center for Responsive Law, documents how women are charged more than men for the same service. In the studies cited in this book, the population is U.S. women. The studies involved data collected from a sample of U.S. women.

For these inferences to be correct, the sample chosen must be a **random sample**. Random samples are representative of the population because they are chosen so that every element of the population is equally likely to be selected. For example, a hand dealt from a well-shuffled deck of cards is a random sample.

After a sample has been chosen and all data of interest are collected, the data must be organized so that conclusions may be more easily drawn. One method of organization is to group the data into intervals; equal intervals are usually chosen.

EXAMPLE 1 A survey asked 30 business executives how many college units in management each had. The results are shown below. Group the data into intervals and find the frequency of each interval.

3	25	22	16	0	9	14	8	34	21
15	12	9	3	8	15	20	12	28	19
17	16	23	19	12	14	29	13	24	18

The highest number in the list is 34 and the lowest is 0; one convenient way to group the data is in intervals of size 5, starting with 0–4 and ending with 30–34. This gives an interval for each number in the list and results in seven equal intervals of a convenient size. Too many intervals of smaller size would not simplify the data enough, while too few intervals of larger size would conceal information that the data might provide. A rule of thumb is to use from six to fifteen intervals.

First tally the number of college units falling into each interval. Then total the tallies in each interval, as in the table below. This table is an example of a **grouped frequency distribution.**

College Units	*Tally*	*Frequency*
0–4	III	3
5–9	IIII	4
10–14	𝍸 I	6
15–19	𝍸 III	8
20–24	𝍸	5
25–29	III	3
30–34	I	1
	Total:	30

■ 1

The frequency distribution in Example 1 shows information about the data that might not have been noticed before. For example, the interval with the largest number of units is 15–19, and 19 executives (more than half) had between 9 and 25 units. Also, the frequency in each interval increases rather evenly (up to 8) and then decreases at about the same pace. However, some information has been lost; for example, we no longer know how many executives had 12 units.

The information in a grouped frequency distribution can be displayed in a histogram similar to the histograms for probability distributions in the previous chapter. The intervals determine the widths of the bars; if equal intervals are used, all the bars have the same width. The heights of the bars are determined by the frequencies.

A **frequency polygon** is another form of graph that illustrates a grouped frequency distribution. The polygon is formed by joining consecutive midpoints of the tops of the histogram bars with straight line segments. Sometimes the midpoints of the first and last bars are joined to endpoints on the horizontal axis where the next midpoint would appear. (See Figure 11.1 on the next page.)

EXAMPLE 2 A grouped frequency distribution of college units was found in Example 1. Draw a histogram and a frequency polygon for this distribution.

First, draw a histogram, shown in black in Figure 11.1. To get a frequency polygon, connect consecutive midpoints of the tops of the bars. The frequency polygon is shown in color. ■ 2

1 An accounting firm selected 24 complex tax returns prepared by a certain tax preparer. The number of errors per return were as follows.

8	12	0	6	10	8	0	14
8	12	14	16	4	14	7	11
9	12	7	15	11	21	22	19

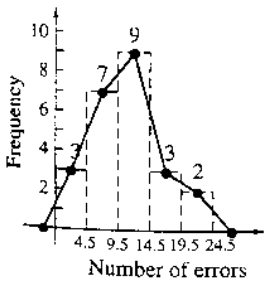
Prepare a grouped frequency distribution for this data. Use intervals 0–4, 5–9, and so on.

Answer:

Interval	*Frequency*
0–4	3
5–9	7
10–14	9
15–19	3
20–24	2
Total:	24

2 Make a histogram and a frequency polygon for the distribution found in Side Problem 1 above.

Answer:



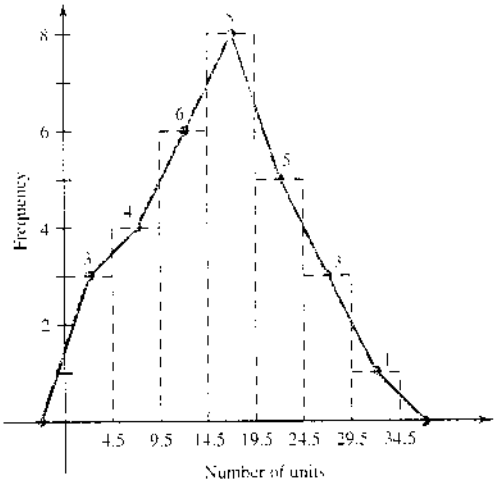


FIGURE 11.1

TECHNOLOGY TIP As noted in Section 10.5, most graphing calculators can display histograms. Many will also display frequency polygons (which are usually labeled LINE or xyLINE in calculator menus). When dealing with grouped frequency distributions, however, certain adjustments must be made on a calculator.

1. *A calculator list of outcomes must consist of single numbers, not intervals.* The table in Example 1, for example, cannot be entered as shown. To convert the first column of the table for calculator use, choose one number in each interval, say 2 in the interval 0–4, 7 in the interval 5–9, 12 in the interval 10–14, etc. Then use 2, 7, 12, . . . as the list of outcomes to be entered into the calculator. The frequency list (the last column of the table) remains the same.
2. *The histogram bar width affects the shape of the graph.* If you use a bar width of 4 in Example 1, the calculator may produce a histogram with gaps in it. To avoid this use the interval $0 \le x < 5$ in place of $0 \le x \le 4$, and similarly for the other intervals and make 5 the bar width.

Following this procedure, we obtain the calculator-generated histogram and frequency polygon in Figure 11.2 for the data from Example 1. Note that the width of each histogram bar is 5. Some calculators cannot display both the histogram and the frequency polygon on the same screen as is done here. ✔

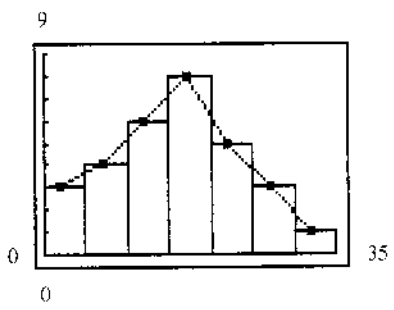


FIGURE 11.2

NOTE The remainder of this section deals with topics that are generally referred to as "measures of central tendency." Computing these various measures is greatly simplified by the statistical capabilities of most scientific and graphing calculators. Calculators vary considerably in how data is entered, so read your instruction manual to learn how to enter lists of data and the corresponding frequencies. On scientific calculators with statistical capabilities, there are keys for finding most of the measures of central tendency discussed below. On graphing calculators, most or all of these measures can be obtained with a single keystroke (look for *one-variable statistics*, which is often labeled 1-VAR, in the STAT menu or its CALC submenu). ◆

MEAN The average value of a probability distribution is the expected value of the distribution. Three measures of central tendency, or "averages," are used with frequency distributions: the mean, the median, and the mode. The most important of these is the mean, which is similar to the expected value of a probability distribution. The **mean** (the arithmetic average) of a set of numbers is the sum of the numbers, divided by the total number of numbers. We write the sum of n numbers $x_1, x_2, x_3, \ldots, x_n$ in a compact way using **summation notation,** also called **sigma notation.** With the Greek letter Σ (sigma), the sum

$$x_1 + x_2 + x_3 + \cdots + x_n$$

is written

$$x_1 + x_2 + x_3 + \cdots + x_n = \sum_{i=1}^{n} x_i.$$

In statistics, $\sum_{i=1}^{n} x_i$ is often abbreviated as just Σx. The symbol $\bar{x}$ (read "x-bar") is used to represent the mean of a sample.

MEAN

The mean of the n numbers $x_1, x_2, x_3, \ldots, x_n$ is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x_i}{n}$$

EXAMPLE 3 The number of bankruptcy petitions (in thousands) filed in the United States in the fiscal years 1988–1993 are given in the table on the next page.* Find the mean number of bankruptcy petitions filed annually during this period.

*Administrative Offices of the U.S. Courts. *Annual Report of the Director.*

Year	Petitions Filed
1988	594
1989	643
1990	725
1991	880
1992	973
1993	919

3 Find the mean of the following list of sales at a boutique.

$25.12	$42.58
$76.19	$32
$81.11	$26.41
$19.76	$59.32
$71.18	$21.03

Answer: $45.47

Let $x_1 = 594$, $x_2 = 643$, and so on. Here, $n = 6$, since there are 6 numbers in the list.

$$\bar{x} = \frac{594 + 643 + 725 + 880 + 973 + 919}{6} = 789$$

The mean number of bankruptcy petitions filed during the given years is 789,000. ■ 3

TECHNOLOGY TIP The mean of the six numbers in Example 3 is easily found by using the $\bar{x}$ key on a scientific calculator or the one-variable statistics key on a graphing calculator. A graphing calculator will also display additional information, which will be discussed in the next section. ✔

The mean of data that have been arranged in a frequency distribution is found in a similar way. For example, suppose the following data are collected.

Value	Frequency
84	2
87	4
88	7
93	4
99	3
	Total: 20

The value 84 appears twice, 87 four times, and so on. To find the mean, first add 84 two times, 87 four times, and so on; or get the same result faster by multiplying 84 by 2, 87 by 4, and so on, and then by adding the results. Dividing the sum by 20, the total of the frequencies, gives the mean.

$$\bar{x} = \frac{(84 \cdot 2) + (87 \cdot 4) + (88 \cdot 7) + (93 \cdot 4) + (99 \cdot 3)}{20}$$

$$= \frac{168 + 348 + 616 + 372 + 297}{20}$$

$$= \frac{1801}{20}$$

$$\bar{x} = 90.05$$

Verify that your calculator gives the same result.

EXAMPLE 4 Find the mean for the data shown in the following frequency distribution.

Value	*Frequency*	*Value* × *Frequency*
30	6	$30 \cdot 6 = 180$
32	9	$32 \cdot 9 = 288$
33	7	$33 \cdot 7 = 231$
37	12	$37 \cdot 12 = 444$
42	6	$42 \cdot 6 = 252$
	Total: 40	Total: 1395

A new column, "Value × Frequency," has been added to the frequency distribution. Adding the products from this column gives a total of 1395. The total from the frequency column is 40. The mean is

$$\bar{x} = \frac{1395}{40} = 34.875. \quad \blacksquare \quad \boxed{4}$$

NOTE The mean in Example 4 is found in the same way as was the expected value of a probability distribution in the previous chapter. In fact, the words *mean* and *expected value* are often used interchangeably. ◆

The mean of grouped data is found in a similar way. For grouped data, intervals are used rather than single values. To calculate the mean, it is assumed that all these values are located at the midpoint of the interval. The letter x is used to represent the midpoints and f represents the frequencies, as shown in the next example.

EXAMPLE 5 Find the mean for the following grouped frequency distribution.

Interval	*Midpoint, x*	*Frequency, f*	*Product, xf*
40–49	44.5	2	89
50–59	54.5	4	218
60–69	64.5	7	451.5
70–79	74.5	9	670.5
80–89	84.5	5	422.5
90–99	94.5	3	283.5
		Total: 30	Total: 2135

A column for the midpoint of each interval has been added. The numbers in this column are found by adding the endpoints of each interval and dividing by 2. For the interval 40–49, the midpoint is $(40 + 49)/2 = 44.5$. The numbers in the product column on the right are found by multiplying frequencies and corresponding midpoints. Finally, we divide the total of the product column by the total of the frequency column to get

$$\bar{x} = \frac{2135}{30} = 71.2 \text{ (to the nearest tenth).} \quad \blacksquare$$

$\boxed{4}$ Find $\bar{x}$ for the following frequency distribution.

Value	*Frequency*
7	2
9	3
11	6
13	4
15	1
17	4

Answer:
$\bar{x} = 12.1$

5 Find the mean of the following grouped frequency distribution.

Interval	*Frequency*
0–4	6
5–9	4
10–14	7
15–19	3

Answer:
8.75

The formula for the **mean of a grouped frequency distribution** is given below.

MEAN OF A GROUPED DISTRIBUTION

The mean of a distribution where x represents the midpoints, f the frequencies, and $n = \Sigma f$, is

$$\bar{x} = \frac{\Sigma(xf)}{n}.$$

CAUTION Note that in the formula above, n is the sum of the frequencies in the entire data set, not the number of intervals. ◆ 5

MEDIAN Asked by a reporter to give the average height of the players on his team the Little League coach lined up his 15 players by increasing height. He picked out the player in the middle and pronounced this player to be of average height. This kind of average, called the **median,** is defined as the middle entry in a set of data arranged in either increasing or decreasing order. If there is an even number of entries, the median is defined to be the mean of the two center entries. The following table shows how to find the median for two sets of data: {8, 7, 4, 3, 1} and {2, 3, 4, 7, 9, 12}.

Odd Number of Entries	*Even Number of Entries*
8	2
7	3
Median = 4	4 } Median = $\frac{4+7}{2} = 5.5$
3	7 }
1	9
	12

NOTE As shown in the table above, when there are an even number of entries, the median is not always equal to one of the data entries. ◆

The procedure for finding the median of a grouped frequency distribution is more complicated. We omit it here because it is more common to find the mean when working with grouped frequency distributions.

EXAMPLE 6 Find the median for the following lists of numbers.

(a) 11, 12, 17, 20, 23, 28, 29

The median is the middle number, in this case 20. (Note that the numbers are already arranged in numerical order.) In this list, three numbers are smaller than 20 and three are larger.

(b) 15, 13, 7, 11, 19, 30, 39, 5, 10

First arrange the numbers in numerical order, from smallest to largest.

5, 7, 10, 11, 13, 15, 19, 30, 39

The middle number can now be determined; the median is 13.

6 Find the median for each of the following lists of numbers.

(a) 12, 15, 17, 19, 35, 42, 58

(b) 28, 68, 7, 15, 47, 59, 13, 74, 32, 25

Answers:

(a) 19

(b) 30

(c) 47, 59, 32, 81, 74, 153

Write the numbers in numerical order.

$$32,\quad 47,\quad 59,\quad 74,\quad 81,\quad 153$$

There are six numbers here; the median is the mean of the two middle numbers, or

$$\text{Median} = \frac{59 + 74}{2} = \frac{133}{2} = 66.5. \quad \blacksquare$$

CAUTION Remember, the data must be arranged in numerical order before locating the median. ◆ 6

TECHNOLOGY TIP Many graphing calculators (including TI-82/83/86, HP-38, Casio 9800/9850) display the median when doing one-variable statistics. You may have to scroll down to a second screen to find it. ✔

In some situations, the median gives a truer representative or typical element of the data than the mean. For example, suppose in an office there are 10 salespersons, 4 secretaries, the sales manager, and Ms. Daly, who owns the business. Their annual salaries are as follows: secretaries, $15,000 each; salespersons, $25,000 each; manager, $35,000; and owner, $200,000. The mean salary is

$$\bar{x} = \frac{(15{,}000)4 + (25{,}000)10 + 35{,}000 + 200{,}000}{16} = \$34{,}062.50.$$

However, since 14 people earn less than $34,062.50 and only 2 earn more, this does not seem very representative. The median salary is found by ranking the salaries by size: $15,000, $15,000, $15,000, $15,000, $25,000, $25,000, . . . , $200,000. There are 16 salaries (an even number) in the list, so the mean of the 8th and 9th entries will give the value of the median. The 8th and 9th entries are both $25,000, so the median is $25,000. In this example, the median is more representative of the distribution than the mean.

MODE Sue's scores on ten class quizzes include one 7, two 8's, six 9's and one 10. She claims that her average grade on quizzes is 9, because most of her scores are 9's. This kind of "average," found by selecting the most frequent entry, is called the **mode.**

7 Find the mode for each of the following lists of numbers.

(a) 29, 35, 29, 18, 29, 56, 48

(b) 13, 17, 19, 20, 20, 13, 25, 27, 13, 20

(c) 512, 546, 318, 729, 854, 253

Answers:

(a) 29

(b) 13 and 20

(c) No mode

EXAMPLE 7 Find the mode for each list of numbers.

(a) 57, 38, **55**, **55**, 80, 87, 98

The number 55 occurs more often than any other, so it is the mode. It is not necessary to place the numbers in numerical order when looking for the mode.

(b) 182, **185**, 183, **185**, **187**, **187**, 189

Both 185 and 187 occur twice. This list has *two* modes.

(c) 10,708, 11,519, 10,972, 17,546, 13,905, 12,182

No number occurs more than once. This list has no mode. ■ 7

The mode has the advantages of being easily found and not being influenced by data that are very large or very small compared to the rest of the data. It is often used in samples where the data to be "averaged" are not numerical. The major disadvantage of the mode is that there may be more than one, in case of ties, or there may be no mode at all when all entries occur with the same frequency.

The mean is the most commonly used measure of central tendency. Its advantages are that it is easy to compute, it takes all the data into consideration, and it is

reliable—that is, repeated samples are likely to give very similar means. A disadvantage of the mean is that it is influenced by extreme values, as illustrated in the salary example above.

The median can be easy to compute and is influenced very little by extremes. Like the mode, the median can be found in situations where the data are not numerical. A disadvantage of the median is the need to rank the data in order; this can be tedious when the number of items is large.

11.1 EXERCISES

For Exercises 1–4, (a) *group the data as indicated;* (b) *prepare a frequency distribution with columns for intervals and frequencies;* (c) *construct a histogram;* (d) *construct a frequency polygon. (See Examples 1 and 2.)*

1. Use six intervals, starting with 0–24.

74	133	4	127	20	30
103	27	139	118	138	121
149	132	64	141	130	76
42	50	95	56	65	104
4	140	12	88	119	64

2. Use seven intervals, starting with 30–39.

79	71	78	87	69	50	63	51	60	46
65	65	56	88	94	56	74	63	87	62
84	76	82	67	59	66	57	81	93	93
54	88	55	69	78	63	63	48	89	81
98	42	91	66	60	70	64	70	61	75
82	65	68	39	77	81	67	62	73	49
51	76	94	54	83	71	94	45	73	95
72	66	71	77	48	51	54	57	69	87

3. Use 70–74 as the first interval.

79	84	88	96	102	104	110	108	106	106	
104	99	97	92	94	90	82	74	72	83	
84	92	100	99	101	107	111	102	97	94	92

4. Use 140–149 as the first interval.

174	190	172	182	179	186	171	152	174	185
180	170	160	173	163	177	165	157	149	167
169	182	178	158	182	169	181	173	183	176
170	162	159	147	150	192	179	165	148	188

5. How does a frequency polygon differ from a histogram?

6. Discuss the advantages and disadvantages of the mean as a measure of central tendency.

Find the mean for each list of numbers. Round to the nearest tenth.

7. 8, 10, 16, 21, 25

8. 44, 41, 25, 36, 67, 51

9. 21,900, 22,850, 24,930, 29,710, 28,340, 40,000

10. 38,500, 39,720, 42,183, 21,982, 43,250

11. 9.4, 11.3, 10.5, 7.4, 9.1, 8.4, 9.7, 5.2, 1.1, 4.7

12. 30.1, 42.8, 91.6, 51.2, 88.3, 21.9, 43.7, 51.2

Find the mean for each of the following. Round to the nearest tenth. (See Example 4.)

13.

Value	*Frequency*
3	4
5	2
9	1
12	3

14.

Value	*Frequency*
9	3
12	5
15	1
18	1

15.

Value	*Frequency*
12	4
13	2
15	5
19	3
22	1
23	5

16.

Value	*Frequency*
25	1
26	2
29	5
30	4
32	3
33	5

Find the median for each of the following lists of numbers. (See Example 6.)

17. 12, 18, 32, 51, 58, 92, 106

18. 596, 604, 612, 683, 719

19. 100, 114, 125, 135, 150, 172

20. 1072, 1068, 1093, 1042, 1056, 1005, 1009

21. 28.4, 9.1, 3.4, 27.6, 59.8, 32.1, 47.6, 29.8

22. .6, .4, .9, 1.2, .3, 4.1, 2.2, .4, .7, .1

Find the mode or modes for each of the following lists of numbers. (See Example 7.)

23. 4, 9, 8, 6, 9, 2, 1, 3
24. 21, 32, 46, 32, 49, 32, 49
25. 74, 68, 68, 68, 75, 75, 74, 74, 70
26. 158, 162, 165, 162, 165, 157, 163
27. 6.8, 6.3, 6.3, 6.9, 6.7, 6.4, 6.1, 6.0
28. 12.75, 18.32, 19.41, 12.75, 18.30, 19.45, 18.33
29. When is the median the most appropriate measure of central tendency?
30. Under what circumstances would the mode be an appropriate measure of central tendency?

For grouped data, the modal class *is the interval containing the most data values. Give the mean and modal class for each of the following collections of grouped data. (See Example 5.)*

31. The distribution in Exercise 3.
32. The distribution in Exercise 4.

For each set of ungrouped data, (a) *Find the mean, median, and mode.* (b) *Discuss which of the three measures best represents the data and why.*

33. The weight gains of 10 experimental rats fed on a special diet were $-1, 0, -3, 7, 1, 1, 5, 4, 1, 4$.
34. A sample of 7 measurements of the thickness of a copper wire were .010, .010, .009, .008, .007, .009, .008.
35. The times in minutes that 12 patients spent in a doctor's office were 20, 15, 18, 22, 10, 12, 16, 17, 19, 21, 23, 13.
36. The scores on a 10-point botany quiz were 6, 6, 8, 10, 9, 7, 6, 5, 6, 8, 3.
37. **Management** A firm took a random sample of the number of days absent in a year for 40 employees, with results as shown below.

Days Absent	*Frequency*
0–2	23
3–5	11
6–8	5
9–11	0
12–14	1

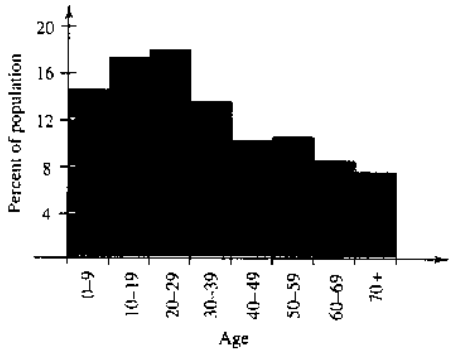
Sketch a histogram and a frequency polygon for the data.

38. **Natural Science** The size of the home ranges (in square kilometers) of several pandas were surveyed over a year's time, with the following results.

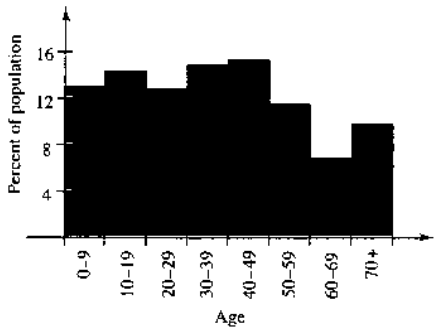
Home Range	*Frequency*
.1–.5	11
.6–1.0	12
1.1–1.5	7
1.6–2.0	6
2.1–2.5	2
2.6–3.0	1
3.1–3.5	1

Sketch a histogram and frequency polygon for the data.

•39. **Social Science** The histogram below shows the percent of the U.S. population in each age group in 1980.* What percent of the population was in each of the following age groups?
(a) 10–19 (b) 60–69
(c) What age group had the largest percent of the population?



40. **Social Science** The histogram below shows the estimated percent of the U.S. population in each age group in the year 2000.* What percent of the population is estimated to be in each of the following age groups then?
(a) 20–29 (b) 70+
(c) What age group will have the largest percent of the population?
(d) Compare the histogram in Exercise 39 with the histogram below. What seems to be true of the U.S. population?



*Data from Census Bureau statistics.