Statistics

May 27, 2022

Second Course-Second Stage

Ferman Ali Ahmed
Department of Chemistry -College of Education
University of Salahaddin

Contents

1 The Nature of Probability and Statistics

Introduction

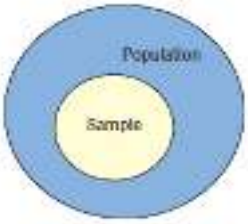
Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

1–1 Descriptive and Inferential Statistics

A **variable** is a characteristic or attribute that can assume different values.

A **population** consists of all subjects (human or otherwise) that are being studied.

A **sample** is a group of subjects selected from a population.



The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are

1. Descriptive statistics
2. Inferential statistics

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

EXAMPLE 1–1 Descriptive or Inferential Statistics

Determine whether descriptive or inferential statistics were used.

a. The average jackpot for the top five lottery winners was $367.6 million.

b. A study done by the American Academy of Neurology suggests that older people who had a high caloric diet more than doubled their risk of memory loss.

c. Based on a survey of 9317 consumers done by the National Retail Federation, the average amount that consumers spent on Valentine's Day in 2011 was $116.

d. Scientists at the University of Oxford in England found that a good laugh significantly raises a person's pain level tolerance.

SOLUTION

a. Descriptive statistics were used because this is an average, and it is based on data obtained from the top five lottery winners at this time.

b. Inferential statistics were used since this is a generalization made from a sample to a population.

c. Descriptive statistics were used since this is an average based on a sample of 9317 respondents.

d. Inferential statistics were used since an inference is made from a sample to a population.

1–2 Variables and Types of Data

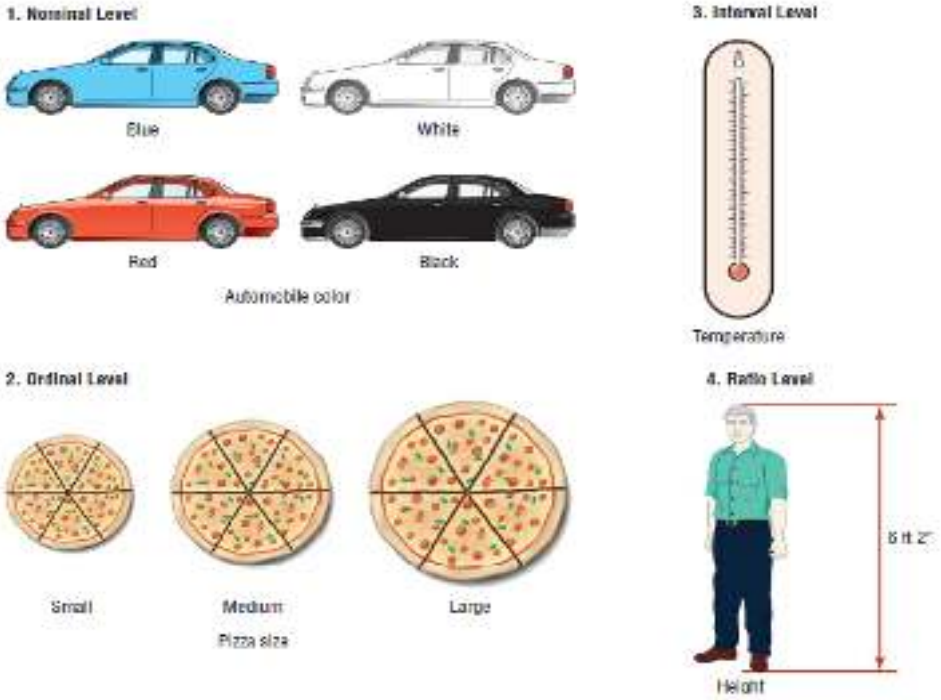
Variables can be classified as qualitative or quantitative.

Qualitative variables are variables that have distinct categories according to some characteristic or attribute.

Quantitative variables are variables that can be counted or measured.

Discrete variables assume values that can be counted.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.



1–3 Data Collection and Sampling Techniques

Random Sampling

A random sample is a sample in which all members of the population have an equal chance of being selected.

Some two-digit random numbers are shown in Table 1–3. To select a random sample of, say, 15 subjects out of 85 subjects, it is necessary to number each subject from 01 to 85. Then select a starting number by closing your eyes and placing your finger on a number in the table. (Although this may sound somewhat unusual, it enables us to find a starting number at random.) In this case suppose your finger landed on the number 12 in the second column. (It is the sixth number down from the top.) Then proceed downward until you have selected 15 different numbers between 01 and 85. When you reach the bottom of the column, go to the top of the next column. If you select a number greater than 85 or the number 00 or a duplicate number, just omit it. In our example, we will use the subjects numbered 12, 27, 75, 62, 57, 13, 31, 06, 16, 49, 46, 71, 53, 41, and 02.

TABLE 1–3 Random Numbers

79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	06	33	15	79	50	28	50	45	45
01	27	92	67	93	31	97	55	29	21	64	27	29
55	75	65	68	65	73	07	95	66	43	43	92	16
84	95	95	96	62	30	91	64	74	83	47	89	71
62	62	21	37	82	62	19	44	08	64	34	50	11
66	57	28	60	13	99	74	31	58	19	47	66	89
48	13	69	97	29	01	75	58	05	40	40	18	29
94	31	73	19	75	76	33	18	05	53	04	51	41
00	06	53	98	01	55	08	38	49	42	10	44	38
46	16	44	27	80	15	28	01	64	27	89	03	27
77	49	85	95	62	93	25	39	63	74	54	82	85
81	96	43	27	39	53	85	61	12	90	67	96	02
40	46	15	73	23	75	96	68	13	99	49	64	11

Systematic Sampling

A systematic sample is a sample obtained by selecting every k^{th} member of the population where k is a counting number.

Researchers obtain systematic samples by numbering each subject of the population and then selecting every k^{th} subject. For example, suppose there were 2000 subjects in the population and a sample of 50 subjects was needed. Since $2000 \div 50 = 40$, then $k = 40$, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at random. Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until 50 subjects were obtained. When using systematic sampling, you must be careful about how the subjects in the population are numbered. If subjects were arranged in a manner such as wife, husband, wife, husband, and every 40th subject were selected, the sample would consist of all husbands. Numbering is not always necessary. For example, a researcher may select every 10th item from an assembly line to test for defects.

Stratified Sampling

A stratified sample is a sample obtained by dividing the population into subgroups or strata according to some characteristic relevant to the study. (There can be several subgroups.) Then subjects are selected from each subgroup.

Samples within the strata should be randomly selected. For example, suppose the president of a two-year college wants to learn how students feel about a certain issue. Furthermore, the president wishes to see if the opinions of first-year students differ from those of second-year students. The president will randomly select students from each subgroup to use in the sample.

Cluster Sampling

A cluster sample is obtained by dividing the population into sections or clusters and then selecting one or more clusters and using all members in the cluster(s) as the members of the sample.

Here the population is divided into groups or clusters by some means such as geographic area or schools in a large school district. Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples. Suppose a researcher wishes to survey apartment dwellers in a large city. If there are 10 apartment buildings in the city, the researcher can select at random 2 buildings from the 10 and interview all the residents of these buildings. Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area. For example, if one wanted to do a study involving the patients in the hospitals in New York City, it would be very costly and time-consuming to try to obtain a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected at random, and the patients in these hospitals would be interviewed in a cluster. See Figure 1–3.

FIGURE 1–3 Sampling Methods

1. Random

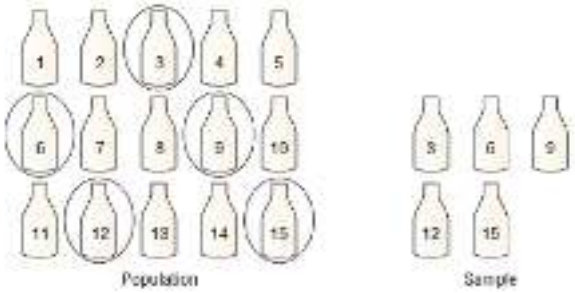
Population

Table D Random Numbers

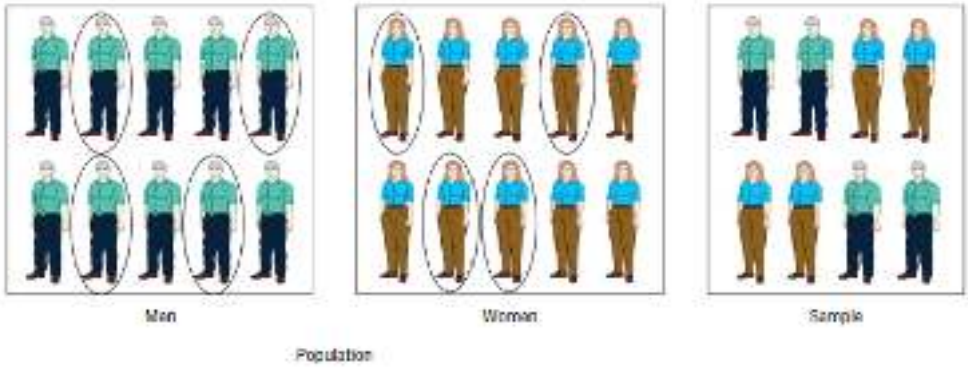
10480	15011	01536	02011	81647
22368	46573	25595	85393	30995
24130	48360	22527	97265	76393
42167	93093	06243	61680	07856
37570	39975	81837	16656	06121
77921	06907	11008	42751	27756
99562	72905	56420	69994	98872
96301	91977	05463	07972	18876
89579	14342	63661	10281	17453
85475	36857	43342	53988	
28918	69578	88231		
63553	40961			

Sample

2. Systematic



3. Stratified



4. Cluster

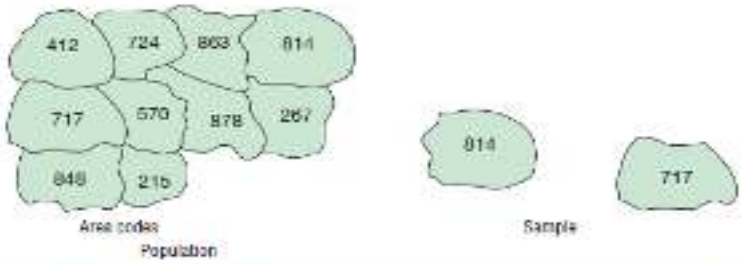


TABLE 1–4 Summary of Sampling Methods

Random	Subjects are selected by random numbers.
Systematic	Subjects are selected by using every *k*th number after the first subject is randomly selected from 1 through *k*.
Stratified	Subjects are selected by dividing up the population into subgroups (strata), and subjects are randomly selected within subgroups.
Cluster	Subjects are selected by using an intact subgroup that is representative of the population.

A nonsampling error occurs when the data are obtained erroneously or the sample is biased, i.e., nonrepresentative.

1–4 Experimental Design

Observational and Experimental Studies

In an **observational study**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

In an **experimental study**, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory variable**. The resultant variable is called the **dependent variable** or the **outcome variable**.

A **confounding variable** is one that influences the dependent or outcome variable but was not separated from the independent variable.

EXAMPLE 1–6 Experimental Design

Researchers randomly assigned 10 people to each of three different groups. Group 1 was instructed to write an essay about the hassles in their lives. Group 2 was instructed to write an essay about circumstances that made them feel thankful. Group 3 was asked to write an essay about events that they felt neutral about. After the exercise, they were given a questionnaire on their outlook on life. The researchers found that those who wrote about circumstances that made them feel thankful had a more optimistic outlook on life. The conclusion is that focusing on the positive makes you more optimistic about life in general. Based on this study, answer the following questions.

a. Was this an observational or experimental study?
b. What is the independent variable?
c. What is the dependent variable?
d. What may be a confounding variable in this study?
e. What can you say about the sample size?
f. Do you agree with the conclusion? Explain your answer.

SOLUTION

a. This is an experimental study since the variables (types of essays written) were manipulated.
b. The independent variable was the type of essay the participants wrote.
c. The dependent variable was the score on the life outlook questionnaire.
d. Other factors, such as age, upbringing, and income, can affect the results; however, the random assignment of subjects is helpful in eliminating these factors. See the answer for the next question.
e. In this study, the sample uses 30 participants total.
f. Answers will vary.

2–1 Organizing Data

1 Suppose a researcher wished to do a study on the ages of the 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine*. When the data are in original form, they are called **raw data** and are listed next.

45	46	64	57	85
92	51	71	54	48
27	66	76	55	69
54	44	54	75	46
61	68	78	61	83
88	45	89	67	55
81	58	55	62	38
55	56	64	81	38
49	68	91	56	68
46	47	83	71	62

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a *frequency distribution*.

A frequency distribution is the organization of raw data in table form, using classes and frequencies.

Categorical Frequency Distributions

The **categorical frequency distribution** is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

EXAMPLE 2–1 Distribution of Blood Types

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

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

Construct a frequency distribution for the data.

SOLUTION

A Class	B Tally	C Frequency	D Percent
A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
		Total 25	100%

Grouped Frequency Distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a **grouped frequency distribution.** For example, a distribution of the blood glucose levels in milligrams per deciliter (mg/dL) for 50 randomly selected college students is shown.

Class limits	Class boundaries	Tally	Frequency
58–64	57.5–64.5	/	1
65–71	64.5–71.5	𝍸 /	6
72–78	71.5–78.5	𝍸𝍸	10
79–85	78.5–85.5	𝍸𝍸 ///	14
86–92	85.5–92.5	𝍸𝍸 //	12
93–99	92.5–99.5	𝍸	5
100–106	99.5–106.5	//	2
			Total 50

The procedure for constructing the preceding frequency distribution is given in Example 2–2; however, several things should be noted. In this distribution, the values 58 and 64 of the first class are called *class limits*. The **lower class limit** is 58; it represents the smallest data value that can be included in the class. The **upper class limit** is 64; it represents the largest data value that can be included in the class. The numbers in the second column are called **class boundaries.** These numbers are used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 64 and 65.

Students sometimes have difficulty finding class boundaries when given the class limits. The basic rule of thumb is that *the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.* For example, if the values in the data set are whole numbers, such as 59, 68, and 82, the limits for a class might be 58–64, and the boundaries are 57.5–64.5. Find the boundaries by subtracting 0.5 from 58 (the lower class limit) and adding 0.5 to 64 (the upper class limit).

$$\text{Lower limit} - 0.5 = 58 - 0.5 = 57.5 = \text{lower boundary}$$
$$\text{Upper limit} + 0.5 = 64 + 0.5 = 64.5 = \text{upper boundary}$$

If the data are in tenths, such as 6.2, 7.8, and 12.6, the limits for a class hypothetically might be 7.8–8.8, and the boundaries for that class would be 7.75–8.85. Find these values by subtracting 0.05 from 7.8 and adding 0.05 to 8.8.

Class boundaries are not always included in frequency distributions; however, they give a more formal approach to the procedure of organizing data, including the fact that sometimes the data have been rounded. You should be familiar with boundaries since you may encounter them in a statistical study.

Finally, the **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example, the class width in the preceding distribution on the distribution of blood glucose levels is 7, found from $65 - 58 = 7$.

The class width can also be found by subtracting the lower boundary from the upper boundary for any given class. In this case, $64.5 - 57.5 = 7$.

Note: Do not subtract the limits of a single class. It will result in an incorrect answer.

The researcher must decide how many classes to use and the width of each class. To construct a frequency distribution, follow these rules:

1. *There should be between 5 and 20 classes.* Although there is no hard-and-fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data.
2. *It is preferable but not absolutely necessary that the class width be an odd number.* This ensures that the midpoint of each class has the same place value as the data. The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the first class in the example with glucose levels is

$$\frac{58 + 64}{2} = 61 \quad \text{or} \quad \frac{57.5 + 64.5}{2} = 61$$

The midpoint is the numeric location of the center of the class.

midpoint is in tenths. For example, if the class width is 6 and the boundaries are 5.5 and 11.5, the midpoint is

$$\frac{5.5 + 11.5}{2} = \frac{17}{2} = 8.5$$

Rule 2 is only a suggestion, and it is not rigorously followed, especially when a computer is used to group data.

3. *The classes must be mutually exclusive.* Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes. Many times, frequency distributions such as this

Age
10–20
20–30
30–40
40–50

are found in the literature or in surveys. If a person is 40 years old, into which class should she or he be placed? A better way to construct a frequency distribution is to use classes such as

Age
10–20
21–31
32–42
43–53

Recall that boundaries are mutually exclusive. For example, when a class boundary is 5.5 to 10.5, the data values that are included in that class are values from 6 to 10. A data value of 5 goes into the previous class, and a data value of 11 goes into the next-higher class.

4. *The classes must be continuous.* Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.
5. *The classes must be exhaustive.* There should be enough classes to accommodate all the data.
6. *The classes must be equal in width.* This avoids a distorted view of the data.

   One exception occurs when a distribution has a class that is **open-ended.** That is, the first class has no specific lower limit, or the last class has no specific upper limit. A frequency distribution with an open-ended class is called an **open-ended distribution.** Here are two examples of distributions with open-ended classes.

Age	Frequency
10–20	3
21–31	6
32–42	4
43–53	10
54 and above	8

Minutes	Frequency
Below 110	16
110–114	24
115–119	38
120–124	14
125–129	5

The frequency distribution for age is open-ended for the last class, which means that anybody who is 54 years or older will be tallied in the last class. The distribution for minutes is open-ended for the first class, meaning that any minute values below 110 will be tallied in that class.

EXAMPLE 2–2 Record High Temperatures

These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data, using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

SOLUTION

The procedure for constructing a grouped frequency distribution for numerical data follows.

Find the highest value and lowest value: $H = 134$ and $L = 100$.

Find the range: R = highest value − lowest value = $H - L$, so

$$R = 134 - 100 = 34$$

Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

Find the class width by dividing the range by the number of classes.

$$\text{Width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Round the answer up to the nearest whole number if there is a remainder: 4.9 = 5.

Then add the width to each upper limit to get all the

$$105 - 1 = 104$$

The first class is 100–104, the second class is 105–109, etc.

Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit:

99.5–104.5, 104.5–109.5, etc.

Tally the data.

Find the numerical frequencies from the tallies.

The completed frequency distribution is

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			Total 50

The frequency distribution shows that the class 109.5–114.5 contains the largest number of temperatures (18) followed by the class 114.5–119.5 with 13 temperatures. Hence, most of the temperatures (31) fall between 110 and 119°F.

The cumulative frequency distribution for the data in this example is as follows:

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

EXAMPLE 2–3 MPGs for SUVs

The data shown here represent the number of miles per gallon (mpg) that 30 selected four-wheel-drive sport utility vehicles obtained in city driving. Construct a frequency distribution, and analyze the distribution.

12	17	12	14	16	18
16	18	12	16	17	15
15	16	12	15	16	16
12	14	15	12	15	15
19	13	16	18	16	14

Source: *Model Year Fuel Economy Guide*, United States Environmental Protection Agency.

SOLUTION

Determine the classes. Since the range of the data set is small ($19 - 12 = 7$), classes consisting of a single data value can be used. They are 12, 13, 14, 15, 16, 17, 18, 19.

The completed ungrouped frequency distribution is

Class limits	Class boundaries	Tally	Frequency
12	11.5–12.5	𝍸 /	6
13	12.5–13.5	/	1
14	13.5–14.5	///	3
15	14.5–15.5	𝍸 /	6
16	15.5–16.5	𝍸 ///	8
17	16.5–17.5	//	2
18	17.5–18.5	///	3
19	18.5–19.5	/	1

In this case, almost one-half (14) of the vehicles get 15 or 16 miles per gallon. The cumulative frequencies are

	Cumulative frequency
Less than 11.5	0
Less than 12.5	6
Less than 13.5	7
Less than 14.5	10
Less than 15.5	16
Less than 16.5	24
Less than 17.5	26
Less than 18.5	29
Less than 19.5	30

Data Description

3–1 Measures of Central Tendency

A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

The Mean

The **mean** is the sum of the values, divided by the total number of values.

The **sample mean**, denoted by $\overline{X}$ (pronounced "*X* bar"), is calculated by using sample data. The sample mean is a statistic.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

where *n* represents the total number of values in the sample.

The **population mean**, denoted by μ (pronounced "mew"), is calculated by using all the values in the population. The population mean is a parameter.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where *N* represents the total number of values in the population.

EXAMPLE 3–1 Police Incidents

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the mean. (Data were obtained by the author.)

475, 447, 440, 761, 993, 1052, 783, 671, 621

SOLUTION

$$\overline{X} = \frac{\Sigma X}{n} = \frac{475 + 447 + 440 + 761 + 993 + 1052 + 783 + 671 + 621}{9}$$

$$= \frac{6243}{9} = 693.7$$

Hence, the mean number of incidents per month to which the police responded is 693.7.

EXAMPLE 3–2 Hospital Infections

The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean.

110 76 29 38 105 31

Source: Pennsylvania Health Care Cost Containment Council.

SOLUTION

$$\overline{X} = \frac{\Sigma X}{n} = \frac{110 + 76 + 29 + 38 + 105 + 31}{6} = \frac{389}{6} = 64.8$$

The mean of the number of hospital infections for the six hospitals is 64.8.

Data Description

3–1 Measures of Central Tendency

A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

The Mean

The **mean** is the sum of the values, divided by the total number of values.

The **sample mean**, denoted by $\overline{X}$ (pronounced "*X* bar"), is calculated by using sample data. The sample mean is a statistic.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

where *n* represents the total number of values in the sample.

The **population mean**, denoted by μ (pronounced "mew"), is calculated by using all the values in the population. The population mean is a parameter.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where *N* represents the total number of values in the population.

EXAMPLE 3–1 Police Incidents

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the mean. (Data were obtained by the author.)

475, 447, 440, 761, 993, 1052, 783, 671, 621

SOLUTION

$$\overline{X} = \frac{\Sigma X}{n} = \frac{475 + 447 + 440 + 761 + 993 + 1052 + 783 + 671 + 621}{9}$$

$$= \frac{6243}{9} = 693.7$$

Hence, the mean number of incidents per month to which the police responded is 693.7.

EXAMPLE 3–2 Hospital Infections

The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean.

110 76 29 38 105 31

Source: Pennsylvania Health Care Cost Containment Council.

SOLUTION

$$\overline{X} = \frac{\Sigma X}{n} = \frac{110 + 76 + 29 + 38 + 105 + 31}{6} = \frac{389}{6} = 64.8$$

The mean of the number of hospital infections for the six hospitals is 64.8.

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

EXAMPLE 3–3 Miles Run per Week

Using the following frequency distribution (taken from Example 2–7), find the mean. The data represent the number of miles run during one week for a sample of 20 runners.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	Total 20

SOLUTION

The procedure for finding the mean for grouped data is given here.

Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13 \quad \text{etc.}$$

For each class, multiply the frequency by the midpoint, as shown, and place the product in column D.

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \quad \text{etc.}$$

The completed table is shown here.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

Find the sum of column D.

Divide the sum by n to get the mean.

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

The Median

The median is the midpoint of the data array. The symbol for the median is MD.

Procedure Table

Finding the Median

Step 1 Arrange the data values in ascending order.

Step 2 Determine the number of values in the data set.

Step 3 *a*. If n is odd, select the middle data value as the median.

b. If n is even, find the mean of the two middle values. That is, add them and divide the sum by 2.

EXAMPLE 3–4 Police Officers Killed

The number of police officers killed in the line of duty over the last 11 years is shown. Find the median.

177 153 122 141 189 155 162 165 149 157 240

Source: National Law Enforcement Officers Memorial Fund.

SOLUTION

Step 1 Arrange the data in ascending order.

122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240

Step 2 There are an odd number of data values, namely, 11.

Step 3 Select the middle data value.

122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240

↑

Median

The median number of police officers killed for the 11-year period is 157.

EXAMPLE 3–5 Tornadoes in the United States

The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

SOLUTION

Step 1 Arrange the data values in ascending order.

656, 684, 702, 764, 856, 1132, 1133, 1303

Step 2 There are an even number of data values, namely, 8.

Step 3 The middle two data values are 764 and 856.

656, 684, 702, 764, 856, 1132, 1133, 1303

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

The Mode

The value that occurs most often in a data set is called the mode.

EXAMPLE 3–6 NFL Signing Bonuses

Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Source: USA TODAY.

SOLUTION

It is helpful to arrange the data in order, although it is not necessary.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Since $10 million occurred 3 times—a frequency larger than any other number—the mode is $10 million.

EXAMPLE 3–8 Accidental Firearm Deaths

The number of accidental deaths due to firearms for a six-year period is shown. Find the mode.

649, 789, 642, 613, 610, 600

Source: National Safety Council.

SOLUTION

Since each value occurs only once, there is no mode.

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

EXAMPLE 3–9 Miles Run per Week

Find the modal class for the frequency distribution of miles that 20 runners ran in one week, used in Example 2–7.

Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5 ← Modal class
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

SOLUTION

The modal class is 20.5–25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 miles per week.

The Midrange

The midrange is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

EXAMPLE 3–12 Bank Failures

The number of bank failures for a recent five-year period is shown. Find the midrange.

3, 30, 148, 157, 71

Source: Federal Deposit Insurance Corporation.

SOLUTION

The lowest data value is 3, and the highest data value is 157.

$$\text{MR} = \frac{3 + 157}{2} = \frac{160}{2} = 80$$

The midrange for the number of bank failures is 80.

EXAMPLE 3–13 NFL Signing Bonuses

Find the midrange of data for the NFL signing bonuses in Example 3–6. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

SOLUTION

The lowest bonus is $10 million, and the largest bonus is $34.5 million.

$$\text{MR} = \frac{10 + 34.5}{2} = \frac{44.5}{2} = \$22.25 \text{ million}$$

Notice that this amount is larger than seven of the eight amounts and is not typical of the average of the bonuses. The reason is that there is one very high bonus, namely, $34.5 million.

Find the modal class for the frequency distribution of times that 20 runners ran in one week, used in Example 2–7.

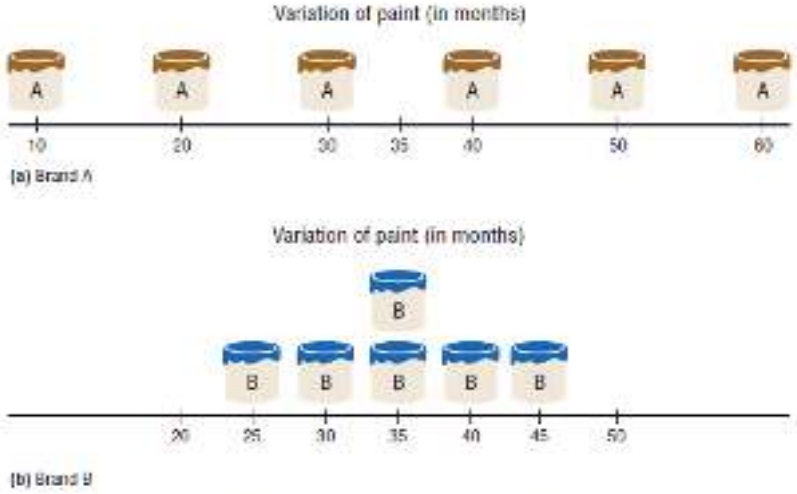
Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5 ← Modal class
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

SOLUTION

The modal class is 20.5–25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 miles per week.

Since the means are equal in Example 3–15, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 3–2.

As Figure 3–2 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 3–2 shows that brand B performs more



consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: *range*, *variance*, and *standard deviation*. Each measure will be discussed in this section.

Range

The **range** is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

EXAMPLE 3–16 Comparison of Outdoor Paint

Find the ranges for the paints in Example 3–15.

SOLUTION

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

Population Variance and Standard Deviation

The **population variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where X = individual value
μ = population mean
N = population size

The **population standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Procedure Table

Finding the Population Variance and Population Standard Deviation

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N}$$

Step 2 Find the deviation for each data value.

$$X - \mu$$

Step 3 Square each of the deviations.

$$(X - \mu)^2$$

Step 4 Find the sum of the squares.

$$\Sigma(X - \mu)^2$$

Step 5 Divide by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Step 6 Take the square root of the variance to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

EXAMPLE 3–18 Comparison of Outdoor Paint

Find the variance and standard deviation for the data set for brand A paint in Example 3–15. The number of months brand A lasted before fading was

10, 60, 50, 30, 40, 20

SOLUTION

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value $(X - \mu)$.

$10 - 35 = -25$ $\quad 50 - 35 = +15$ $\quad 40 - 35 = +5$

$60 - 35 = +25$ $\quad 30 - 35 = -5$ $\quad 20 - 35 = -15$

Step 3 Square each result $(X - \mu)^2$.

$(-25)^2 = 625$ $\quad (+15)^2 = 225$ $\quad (+5)^2 = 25$

$(+25)^2 = 625$ $\quad (-5)^2 = 25$ $\quad (-15)^2 = 225$

Step 4 Find the sum of the squares $\Sigma(X - \mu)^2$.

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

Step 5 Divide the sum by N to get the variance $\dfrac{[\Sigma(X - \mu)^2]}{N}$.

$$\text{Variance} = 1750 \div 6 = 291.7$$

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

A Values X	B $X - \mu$	C $(X - \mu)^2$
10	−25	625
60	+25	625
50	+15	225
30	−5	25
40	+5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

EXAMPLE 3–19 Comparison of Outdoor Paint

Find the variance and standard deviation for brand B paint data in Example 3–15. The months brand B lasted before fading were

35, 45, 30, 35, 40, 25

SOLUTION

Step 1 Find the mean.

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

$35 - 35 = 0$ $\quad 45 - 35 = 10$ $\quad 30 - 35 = -5$
$35 - 35 = 0$ $\quad 40 - 35 = 5$ $\quad 25 - 35 = -10$

Step 3 Square each result and place the squares in column C of the table.

A X	B $X - \mu$	C $(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Step 4 Find the sum of the squares in column C.

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{41.7} \approx 6.5$$

Hence, the standard deviation is 6.5.

Sample Variance and Standard Deviation

Formula for the Sample Variance

The formula for the sample variance (denoted by s^2) is

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

where X = individual value
$\overline{X}$ = sample mean
n = sample size

Formula for the Sample Standard Deviation

The formula for the sample standard deviation, denoted by s, is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

where X = individual value
$\overline{X}$ = sample mean
n = sample size

EXAMPLE 3–20 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and the sample standard deviation.

9 10 14 7 8 3

Source: Pennsylvania School Board Association.

SOLUTION

Step 1 Find the mean of the data values.

$$\overline{X} = \frac{\Sigma X}{n} = \frac{9 + 10 + 14 + 7 + 8 + 3}{6} = \frac{51}{6} = 8.5$$

Step 2 Find the deviation for each data value $(X - \overline{X})$.

$9 - 8.5 = 0.5$ $10 - 8.5 = 1.5$ $14 - 8.5 = 5.5$
$7 - 8.5 = -1.5$ $8 - 8.5 = -0.5$ $3 - 8.5 = -5.5$

Step 3 Square each of the deviations $(X - \overline{X})^2$.

$(0.5)^2 = 0.25$ $(1.5)^2 = 2.25$ $(5.5)^2 = 30.25$
$(-1.5)^2 = 2.25$ $(-0.5)^2 = 0.25$ $(-5.5)^2 = 30.25$

Step 4 Find the sum of the squares.

$$\Sigma(X - \overline{X})^2 = 0.25 + 2.25 + 30.25 + 2.25 + 0.25 + 30.25 = 65.5$$

Step 5 Divide by $n - 1$ to get the variance.

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} = \frac{65.5}{6 - 1} = \frac{65.5}{5} = 13.1$$

Step 6 Take the square root of the variance to get the standard deviation.

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Here the sample variance is 13.1, and the sample standard deviation is 3.6.

Shortcut or Computational Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance	Standard deviation
$s^2 = \dfrac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$	$s = \sqrt{\dfrac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$

EXAMPLE 3–21 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and sample standard deviation.

9, 10, 14, 7, 8, 3

SOLUTION

Step 1 Find the sum of the values:

$$\Sigma X = 9 + 10 + 14 + 7 + 8 + 3 = 51$$

Step 2 Square each value and find the sum:

$$\Sigma X^2 = 9^2 + 10^2 + 14^2 + 7^2 + 8^2 + 3^2 = 499$$

Step 3 Substitute in the formula and solve:

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)} = \frac{6(499) - 51^2}{6(6-1)} = \frac{2994 - 2601}{6(5)} = \frac{393}{30} = 13.1$$

The variance is 13.1.

$s = \sqrt{13.1} \approx 3.6$ (rounded)

Hence, the sample variance is 13.1, and the sample standard deviation is 3.6.

Variance and Standard Deviation for Grouped Data

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Step 6 Take the square root to get the standard deviation.

EXAMPLE 3–22 Miles Run per Week

Find the sample variance and the sample standard deviation for the frequency distribution of the data in Example 2–7. The data represent the number of miles that 20 runners run during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Multiply the frequency by the midpoint for each class, and place the products in column D.

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$$

Multiply the frequency by the square of the midpoint, and place the products in column E.

$$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$$

Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f \cdot X_m$, and the sum of column E is $\Sigma f \cdot X_m^2$. The completed table is shown.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

Substitute in the formula and solve for s^2 to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)} = \frac{20(13{,}310) - 490^2}{20(20-1)}$$

$$= \frac{266{,}200 - 240{,}100}{20(19)} = \frac{26{,}100}{380} \approx 68.7$$

Take the square root to get the standard deviation.

$$s \approx \sqrt{68.7} \approx 8.3$$

Coefficient of Variation

The **coefficient of variation**, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100$$

For populations,

$$\text{CVar} = \frac{\sigma}{\mu} \cdot 100$$

EXAMPLE 3–23 Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

SOLUTION

The coefficients of variation are

$$\text{CVar} = \frac{s}{\bar{X}} = \frac{5}{87} \cdot 100 = 5.7\% \quad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100 = 14.8\% \quad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

EXAMPLE 3–24 Pages in Women's Fitness Magazines

The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

SOLUTION

The coefficients of variation are

$$\text{CVar} = \frac{\sqrt{23}}{132} \cdot 100 = 3.6\% \quad \text{pages}$$

$$\text{CVar} = \frac{\sqrt{62}}{182} \cdot 100 = 4.3\% \quad \text{advertisements}$$

The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.

3–3 Measures of Position

Standard Scores

There is an old saying, "You can't compare apples and oranges." But with the use of statistics, it can be done to some extent. Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a *standard score* or z score. (We also use z scores in later chapters.)

A standard score or z score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean.

A **z score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is z. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$

For populations, the formula is

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

The z score represents the number of standard deviations that a data value falls above or below the mean.

EXAMPLE 3–27 Test Scores

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

SOLUTION

First, find the z scores. For calculus the z score is

$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

EXAMPLE 3–28 Test Scores

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\overline{X} = 40$	$s = 5$
Test B	$X = 94$	$\overline{X} = 100$	$s = 10$

SOLUTION

For test A,

$$z = \frac{X - \overline{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.