INTRODUCTION

Data, Elements and Observations

Suppose we have information on the test scores of students enrolled in a statistics class. In statistical terminology, the whole set of numbers that represents the scores of students is called a **data set**, the name of each student is called an **element**, and the score of each student is called an **observation**.

Many data sets in their original forms are usually very large, especially those collected by federal and state agencies. Consequently, such data sets are not very helpful in drawing conclusions or making decisions. It is easier to draw conclusions from summary tables and diagrams than from the original version of a data set. So, we summarize data by constructing tables, drawing graphs, or calculating summary measures such as averages. The portion of statistics that helps us do this type of statistical analysis is called **descriptive statistics**.

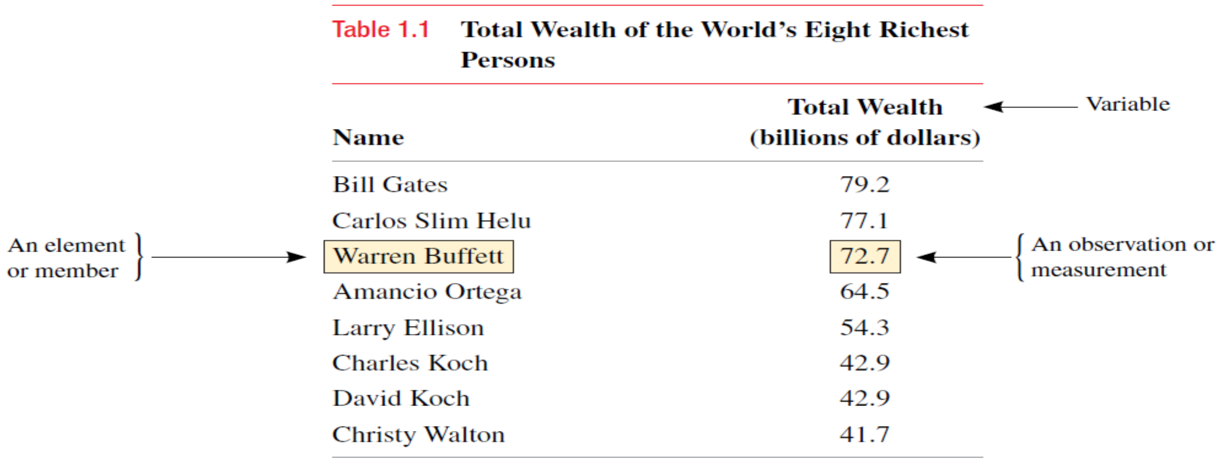

Variable
An element or member
An observation or measurement

Table 1.1 **Total Wealth of the World's Eight Richest Persons**

Name	Total Wealth (billions of dollars)
Bill Gates	79.2
Carlos Slim Helu	77.1
Warren Buffett	72.7
Amancio Ortega	64.5
Larry Ellison	54.3
Charles Koch	42.9
David Koch	42.9
Christy Walton	41.7

Data Set

A **data set** is a collection of observations on one or more variables. Other examples of data sets are a list of the prices of 25 recently sold homes, test scores of 15 students, opinions of 100 voters, and ages of all employees of a company.

The data in Table 2.1 are quantitative raw data.

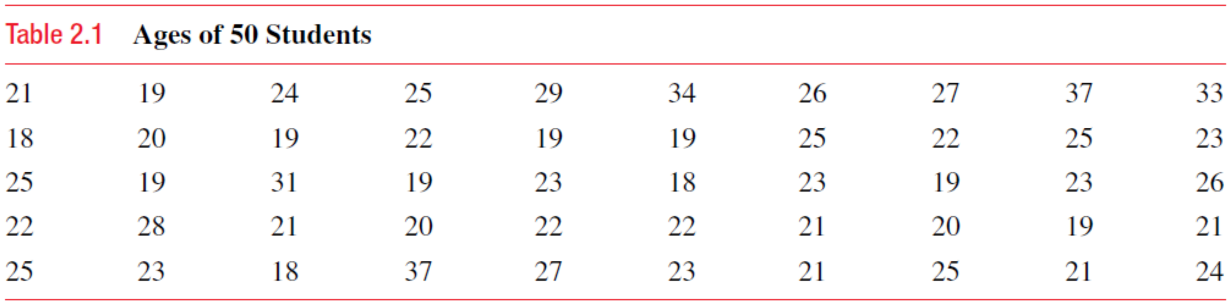


Table 2.1 **Ages of 50 Students**

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

While, suppose we ask the same 50 students about their student status. The responses of the students are recorded in Table 2.2. In this table, F, SO, J, and SE are the abbreviations for freshman, sophomore, junior, and senior, respectively. This is an example of qualitative (or categorical) raw data.

Qualitative (or categorical) raw data

Table 2.2 **Status of 50 Students**

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

Remember:

The data presented in Tables 2.1 and 2.2 are also called **ungrouped data**. An ungrouped data set contains information on each member of a sample or population individually. If we rank the data of Table 2.1 from lowest to the highest age, they will still be ungrouped data but not raw data.

Descriptive Statistics

Descriptive statistics consists of methods for organizing, displaying, and describing data by using tables, graphs, and summary measures.

Population and Sample

In statistics, the collection of all elements of interest is called a **population**. The selection of a portion of the elements from this population is called a **sample**.

Figure 1.1 illustrates the selection of a sample from a population.

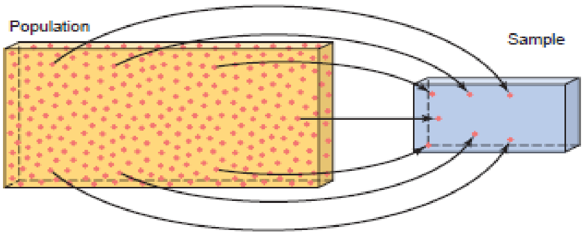


Figure 1.1 Population and sample.

A major portion of statistics deals with making decisions, inferences, predictions, and forecasts about populations based on results obtained from samples. For example, we may make some decisions about the political views of all college and university students based on the political views of 1000 students selected from a few colleges and universities. To make a decision based on this information. The area of statistics that deals with such decision-making procedures is referred to as **inferential statistics**. This branch of statistics is also called *inductive reasoning* or *inductive statistics.*

Inferential Statistics

Inferential statistics consists of methods that use sample results to help make decisions or predictions about a population.

Element or Member

An **element** or **member** of a sample or population is a specific subject or object (for example, a person, firm, item, state, or country) about which the information is collected.

Variable

A **variable** is a characteristic under study that assumes different values for different elements. In contrast to a variable, the value of a *constant* is fixed. A variable is often denoted by x, y, or z

Examples of variables are household incomes, the number of houses built in a city per month during the past year, the makes of cars owned by people, the gross profits of companies, and the number of insurance policies sold by a salesperson per day during the past month.

Types of Variables

A variable is a characteristic under investigation that assumes different values for different elements.

A variable may be classified as quantitative or qualitative. These two types of variables are explained next.

Quantitative Variables

Some variables (such as the price of a home) can be measured numerically, whereas others (such as hair color) cannot. The price of a home is an example of a **quantitative variable** while hair color is an example of a **qualitative variable**.

Quantitative Variable

A variable that can be measured numerically is called a **quantitative variable**. The data collected on a quantitative variable are called **quantitative data**. Income, height, gross sales, price of a home, number of cars owned, and number of accidents are examples of quantitative variables because each of them can be expressed numerically. For instance, the income of a family may be $81,520.75 per year, the gross sales for a company may be $567 million for the past year, and so forth. Such quantitative variables may be classified as either *discrete variables* or *continuous variables.*

Discrete and Continuous variables

The values that a certain quantitative variable can assume may be countable or non-countable. For example, we can count the number of cars owned by a family, but we cannot count the height of a family member, as it is measured on a continuous scale. A variable that assumes countable values is called a **discrete variable**. The variables which are measures on a continuous scale are called continuous variables.

Definition

Discrete Variable

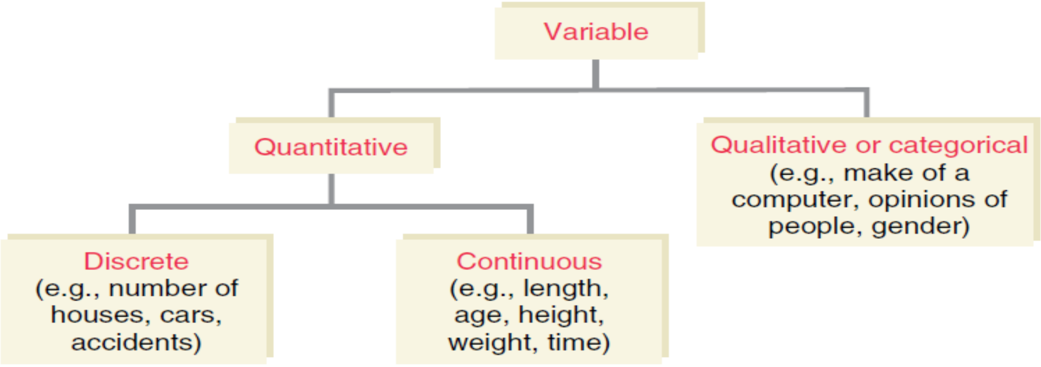
A variable whose values are countable is called a **discrete variable**. In other words, a discrete variable can assume only certain values with no intermediate values. For example, the number of cars sold on any given day at a car dealership is a discrete variable because the number of cars sold must be 0, 1, 2, 3, . . . and we can count it. The number of cars sold cannot be between 0 and 1, or between 1 and 2. Other examples of discrete variables are the number of people visiting a bank on any day, the number of cars in a parking lot, the number of cattle owned by a farmer, and the number of students in a class.

Continuous Variable

A variable that can assume any numerical value over a certain interval or intervals is called a **continuous variable**. For example: Income of families, height of employees, gross sales of stores, prices of homes

Qualitative or Categorical Variable

A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a **qualitative** or **categorical variable.** The data collected on such a variable are called **qualitative data**. For example, the status of an undergraduate college student is a qualitative variable because a student can fall into any one of four categories: freshman, sophomore, junior, or senior. Other examples of qualitative variables are the gender of a person.


Variable
Quantitative
Qualitative or categorical (e.g., make of a computer, opinions of people, gender)
Discrete (e.g., number of houses, cars, accidents)
Continuous (e.g., length, age, height, weight, time)

Organizing and Graphing Data

Organizing and Graphing Qualitative Data

Frequency Distribution of a Qualitative Variable

A *frequency distribution* of a qualitative variable lists all categories and the number of elements that belong to each of the categories.

EXAMPLE 2.1 What Variety of Donuts Is Your Favorite?

A sample of 30 persons who often consume donuts were asked what variety of donuts is their favorite. The responses from these 30 persons are as follows:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

Construct a frequency distribution table for these data.

Table 2.1 Frequency Distribution of Favorite Donut Variety

Donut Variety	Tally	Frequency (f)
Glazed	𝍸 |||	8
Filled	𝍸 ||	7
Frosted	𝍸	5
Plain	|||	3
Other	𝍸 ||	7
		Sum = 30

Relative Frequency and Percentage Distributions

The **relative frequency** of a category is obtained by dividing the frequency of that category by the sum of all frequencies. Thus, the relative frequency shows what fractional part or proportion of the total frequency belongs to the corresponding category. A *relative frequency distribution* lists the relative frequencies for all categories.

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

The **percentage** for a category is obtained by multiplying the relative frequency of that category by 100. A *percentage distribution* lists the percentages for all categories.

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100\%$$

Relative frequency and Percentage frequency distributions for the data in Example 2.1

Table 2.2 Relative Frequency and Percentage Distributions of Favorite Donut Variety

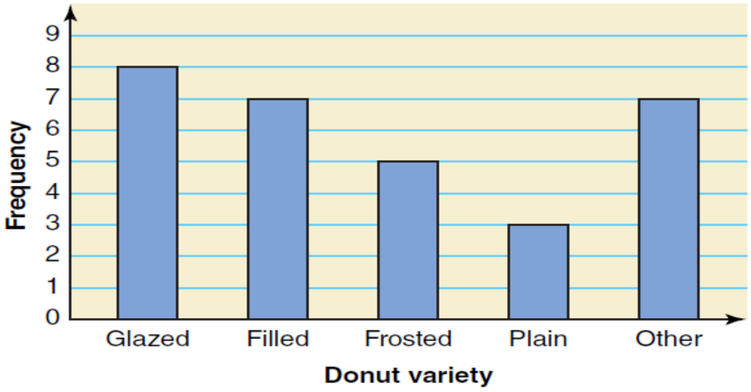
Donut Variety	Relative Frequency	Percentage
Glazed	8/30 = .267	.267(100) = 26.7
Filled	7/30 = .233	.233(100) = 23.3
Frosted	5/30 = .167	.167(100) = 16.7
Plain	3/30 = .100	.100(100) = 10.0
Other	7/30 = .233	.233(100) = 23.3
	Sum = 1.000	Sum = 100%

Graphical Presentation of Qualitative Data

A graphic display can reveal at a glance the main characteristics of a data set. The *bar graph* and the *pie chart* are two types of graphs that are commonly used to display qualitative data.

Bar Graphs

To construct a **bar graph** (also called a *bar chart*), we mark the various categories on the horizontal axis as in Figure 2.1. Note that all categories are represented by intervals of the same width. We mark the frequencies on the vertical axis. Then we draw one bar for each category such that the height of the bar represents the frequency of the corresponding category. We leave a small gap between adjacent bars. Figure 2.1 gives the bar graph for the frequency distribution of Table 2.1.



Bar graph for the frequency distribution in Table 2.1.

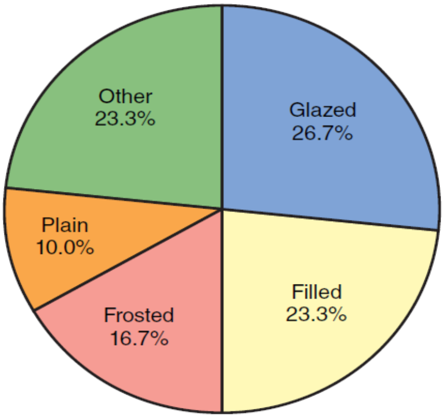
Pie Charts

A **pie chart** is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The whole pie (or circle) represents the total sample or population. Then we divide the pie into different portions that represent the different categories.

Definition:

Pie Chart

A circle divided into portions that represent the relative frequencies or percentages of a population or a sample belonging to different categories is called a *pie chart*.



Pie chart for the percentage distribution of Table 2.2.

Organizing and Graphing Quantitative Data

n the previous section we learned how to group and display qualitative data. This section explains how to group and display quantitative data.

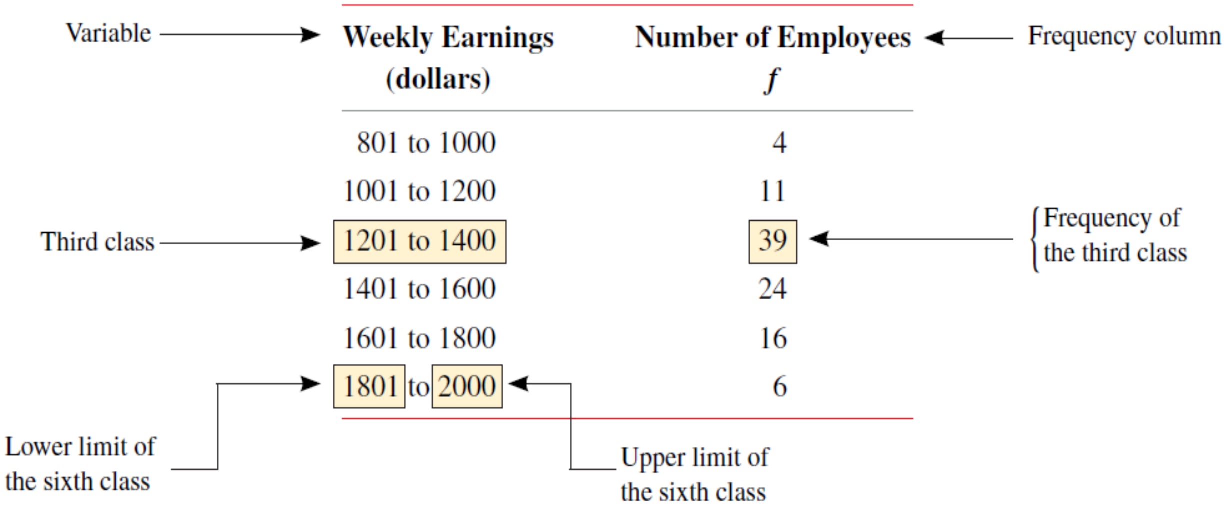
Definition:

Frequency Distribution for Quantitative Data

A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called *grouped data*.

Table 2.3 gives the weekly earnings of 100 employees of a large company. The first column lists the *classes*, which represent the (quantitative) variable *weekly earnings*. For quantitative data, an interval that includes all the values that fall on or within two numbers—the lower and upper limits—is called a **class**. Note that the classes always represent a variable. As we can observe, the classes are non-overlapping; that is, each value for earnings belongs to one and only one class. The second column in the table lists the number of employees who have earnings within each class. For example, 4 employees of this company earn $801 to $1000 per week. The numbers listed in the second column are called the **frequencies,** which give the number of data values that belong to different classes. The frequencies are denoted by *f*. For quantitative data, the frequency of a class represents the number of values in the data set that fall in that class. Table 2.6 contains six classes. Each class has a *lower limit* and an *upper limit*. The values 801, 1001, 1201, 1401, 1601, and 1801 give the lower limits, and the values 1000, 1200, 1400, 1600, 1800, and 2000 are the upper limits of the six classes, respectively. The data presented in Table 2.3 are an illustration of a **frequency distribution table** for quantitative data. Whereas the data that list individual values are called ungrouped data, the data presented in a frequency distribution table are called **grouped data**.

Table 2.3 Weekly Earnings of 100 Employees of a Company



Weekly Earnings (dollars)	Number of Employees f
801 to 1000	4
1001 to 1200	11
1201 to 1400	39
1401 to 1600	24
1601 to 1800	16
1801 to 2000	6

Variable
Frequency column
Third class
Frequency of the third class
Lower limit of the sixth class
Upper limit of the sixth class

Class width/Class Interval

The difference between the lower limits of two consecutive classes gives the **class width**. The class width is also called the **class size**. Width of a class = Lower limit of the next class − Lower limit of the current class.

Thus, in Table 2.3, Width of the first class = 1001 − 801 = 200

The class widths for the frequency distribution of Table 2.3 are listed in the second column of Table 2.4. Each class in Table 2.4 (and Table 2.3) has the same width of 200.

Midpoint or Class Mark

The **class midpoint** or **mark** is obtained by dividing the sum of the two limits of a class by 2.

$$\text{Class midpoint or mark} = \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Thus, the midpoint of the first class is:

$$\text{Midpoint of the first class} = \frac{801 + 1000}{2} = 900.5$$

The class midpoints for the frequency distribution of Table 2.6 are listed in the third column of Table 2.4.

Table 2.4 Class Widths and Class Midpoints for Table 2.3

Class Limits	Class Width	Class Midpoint
801 to 1000	200	900.5
1001 to 1200	200	1100.5
1201 to 1400	200	1300.5
1401 to 1600	200	1500.5
1601 to 1800	200	1700.5
1801 to 2000	200	1900.5

2.2.2 Constructing Frequency Distribution Tables

When constructing a frequency distribution table, we need to make the following three major decisions.

(1) Number of Classes

Usually the number of classes for a frequency distribution table varies from 5 to 20, depending mainly on the number of observations in the data set.1 It is preferable to have more classes as the size of a data set increases. The decision about the number of classes is arbitrarily made by the data organizer.

(2) Class Width

Although it is not uncommon to have classes of different sizes, most of the time it is preferable to have the same width for all classes. To determine the class width when all classes are the same size, first find the difference between the largest and the smallest values in the data. Then, the approximate width of a class is obtained by dividing this difference by the number of desired classes.

$$\text{Approximate class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}$$

Usually this approximate class width is rounded to a convenient number, which is then used as the class width. Note that rounding this number may slightly change the number of classes initially intended.

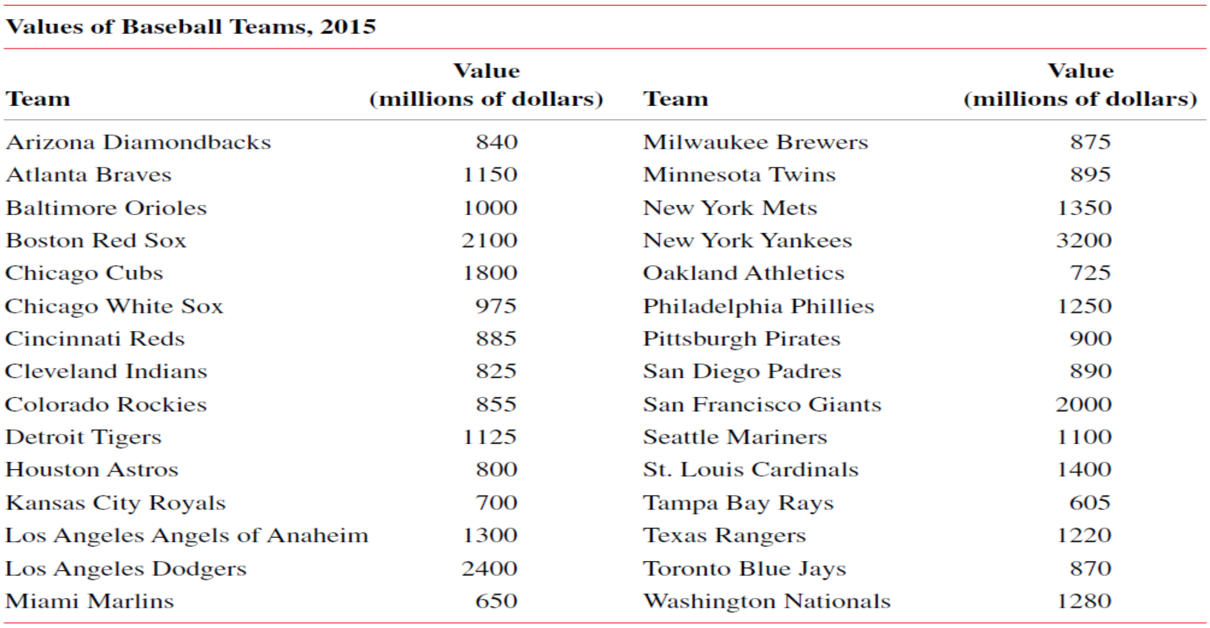
(3) Lower Limit of the First Class or the Starting Point

Any convenient number that is equal to or less than the smallest value in the data set can be used as the lower limit of the first class.

Example 2.3 illustrates the procedure for constructing a frequency distribution table for quantitative data.

EXAMPLE 2.3 Values of Baseball Teams, 2015

The following table gives the value (in million dollars) of each of the 30 baseball teams as estimated by *Forbes* magazine (*source: Forbes* Magazine, April 13, 2015). Construct a frequency distribution table.



Values of Baseball Teams, 2015

Team	Value (millions of dollars)	Team	Value (millions of dollars)
Arizona Diamondbacks	840	Milwaukee Brewers	875
Atlanta Braves	1150	Minnesota Twins	895
Baltimore Orioles	1000	New York Mets	1350
Boston Red Sox	2100	New York Yankees	3200
Chicago Cubs	1800	Oakland Athletics	725
Chicago White Sox	975	Philadelphia Phillies	1250
Cincinnati Reds	885	Pittsburgh Pirates	900
Cleveland Indians	825	San Diego Padres	890
Colorado Rockies	855	San Francisco Giants	2000
Detroit Tigers	1125	Seattle Mariners	1100
Houston Astros	800	St. Louis Cardinals	1400
Kansas City Royals	700	Tampa Bay Rays	605
Los Angeles Angels of Anaheim	1300	Texas Rangers	1220
Los Angeles Dodgers	2400	Toronto Blue Jays	870
Miami Marlins	650	Washington Nationals	1280

Solution In these data, the minimum value is 605, and the maximum value is 3200. Suppose we decide to group these data using six classes of equal width. Then,

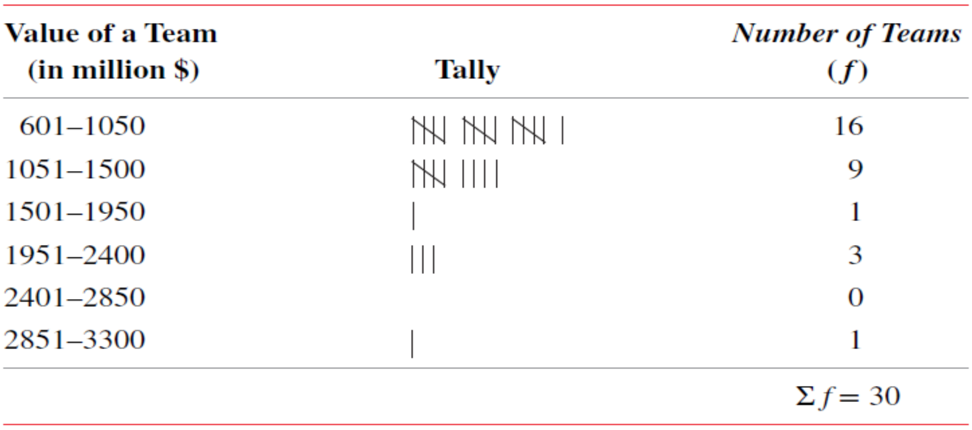
$$\text{Approximate width of each class} = \frac{3200 - 605}{6} = 432.5$$

Now we round this approximate width to a convenient number, say 450. The lower limit of the first class can be taken as 605 or any number less than 605. Suppose we take 601 as the lower limit of the first class. Then our classes will be:

601–1050, 1051–1500, 1501–1950, 1951–2400, 2401–2850, and 2851–3300

We record these five classes in the first column of Table 2.5.

Table 2.5 Frequency Distribution of the Values of Baseball Teams, 2015



Value of a Team		***Number of Teams***
(in million $)	**Tally**	**(*f*)**
601–1050	卌 卌 卌 |	16
1051–1500	卌 ||||	9
1501–1950	|	1
1951–2400	|||	3
2401–2850		0
2851–3300	|	1
		$\Sigma f = 30$

Now we read each value from the given data and mark a tally in the second column of Table 2.5 next to the corresponding class. The first value in our original data set is 840, which belongs to the 601–1050 class. To record it, we mark a tally in the second column next to the 601–1050 class. We continue this process until all the data values have been read and entered in the tally column. Note that tallies are marked in blocks of five for counting convenience. After the tally column is completed, we count the tally marks for each class and write those numbers in the third column. This gives the column of frequencies. These frequencies represent the number of baseball teams with values in the corresponding classes. For example, 16 of the teams have values in the interval $601–$1050 million.

Using the **Σ** notation, we can denote the sum of frequencies of all classes by Σ*f*.

Hence:

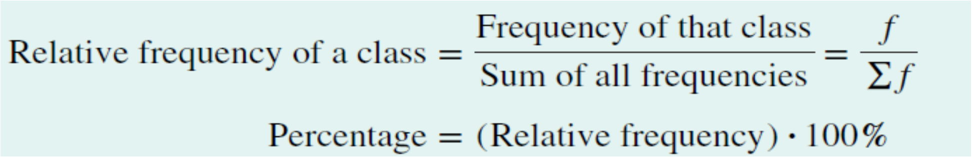
$$\Sigma f = 16 + 9 + 1 + 3 + 0 + 1 = 30$$

The number of observations in a sample is usually denoted by *n*. Thus, for the sample data, **Σ*f*** is equal to *n*.

Note: Note that when we present the data in the form of a frequency distribution table, as in Table 2.5, we lose the information on individual observations. We cannot know the exact value of any team from Table 2.5. All we know is that 16 teams have values in the interval $601–$1050 million, and so forth.

2.2.3 Relative Frequency and Percentage Distributions

Using Table 2.5, we can compute the relative frequency and percentage distributions in the same way as we did for qualitative data in Section 2.1.3. The relative frequencies and percentages for a quantitative data set are obtained as follows. Note that relative frequency is the same as proportion.



$$\text{Relative frequency of a class} = \frac{\text{Frequency of that class}}{\text{Sum of all frequencies}} = \frac{f}{\Sigma f}$$

$$\text{Percentage} = (\text{Relative frequency}) \cdot 100\%$$

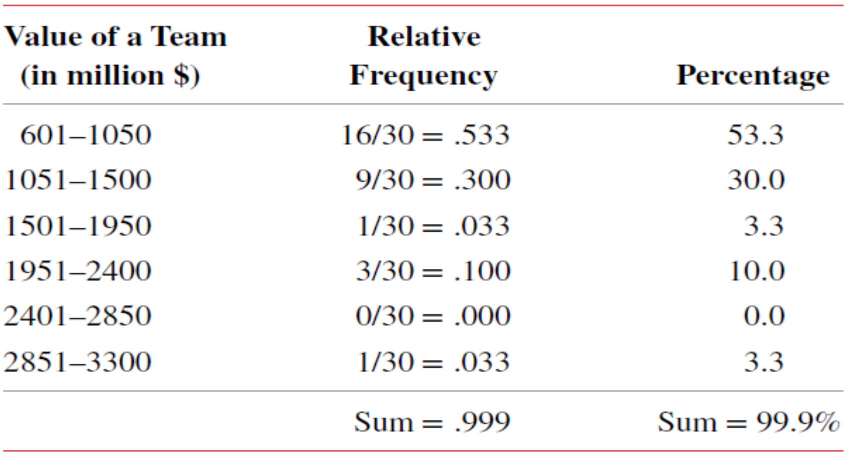
Example 2.4 illustrates how to construct relative frequency and percentage distributions.

EXAMPLE 2.4 Values of Baseball Teams, 2015

Calculate the relative frequencies and percentages for Table 2.5.

Solution The relative frequencies and percentages for the data in Table 2.8 are calculated and listed in the second and third columns, respectively, of Table 2.69.

Table 2.6 Relative Frequency and Percentage Distributions of the Values of Baseball Teams



Value of a Team (in million $)	Relative Frequency	Percentage
601–1050	16/30 = .533	53.3
1051–1500	9/30 = .300	30.0
1501–1950	1/30 = .033	3.3
1951–2400	3/30 = .100	10.0
2401–2850	0/30 = .000	0.0
2851–3300	1/30 = .033	3.3
	Sum = .999	Sum = 99.9%