CHAPTER 1

ATTRIBUTE DESCRIPTORS

The world we live in is a complex and dynamic one. To better understand the world, we often need to reduce it to some simple representative *models*. We often construct models of the real world to decrease the complexity of any problems we study to a manageable level so that we can solve them. We use models of the real world to provide a static version of nature so that we can focus better on the issues at hand.

Geographers typically model the world with objects located at different places on the surface of the world. We use different types of objects to represent the complicated world. We formulate relationships between objects to simulate the dynamics of the world systems. Specifically, we use models to represent the world via *simplification*.

A map is an example of how the real world is modeled. As the map in Figure 1.1 shows, objects in the real world are represented by different *symbols*: *lines* show how rivers run their courses and how roads are connected, while *points* (small, solid circles and squares) and *polygons* (or rectangles of various sizes) show the locations of special interest.

In Figure 1.1, the point representing the county fairground is easily recognized because a text label accompanies it. Similarly, Cuyahoga River, State Route 14, and other highways are identifiable because each of these has a label to identify it. For various buildings represented by squares, however, there is no additional information to help map readers separate one from another to show what they are or what they are for.

We need additional information to give *meaning* to the symbols we use to represent the real world. Like the squares in Figure 1.1, symbols remain only symbols unless we associate them with additional *attribute* information. Lines are

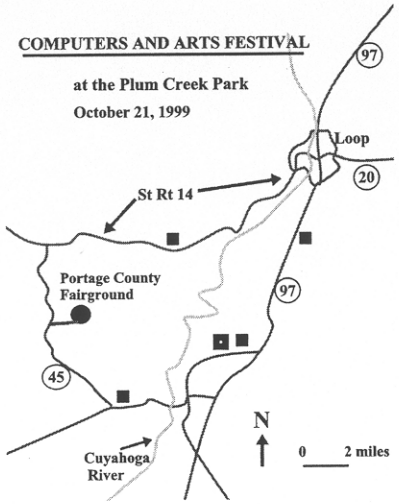
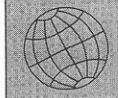


Figure 1.1 A map as a model of the real world.

only lines and points are only points if there is no additional attribute information to describe their properties and characteristics.

In managing geographic information, the conventional approach to structuring spatial data is to have *cartographic data* describing the locations, shapes, or other spatial characteristics of the objects and to have *attribute data* describing other characteristics of the objects. In Figure 1.2, a set of points, representing cities in the three-county area in northeastern Ohio, are shown. To describe each of these points, an *attribute table* records information on their characteristics. In this attribute table, each record is linked to a point. Each record contains a number of *fields* that store attribute data for the associated point. This way, the characteristics of each symbol we use in a map that represents geographic objects in the world can be described in detail in the attribute table.

ArcView Notes



The data structure described in this section is commonly known as *relational data structure*. Typically, a GIS database contains layers of thematic data on the area of interest. Each layer, represented as a shapefile in ArcView, has a map view and an attribute table. The map view is the cartographic data of the thematic layer where coordinates of points, lines, and polygons are used in displays. The attribute table, on the other hand, stores additional attribute information describing various aspects of the objects in the map view.

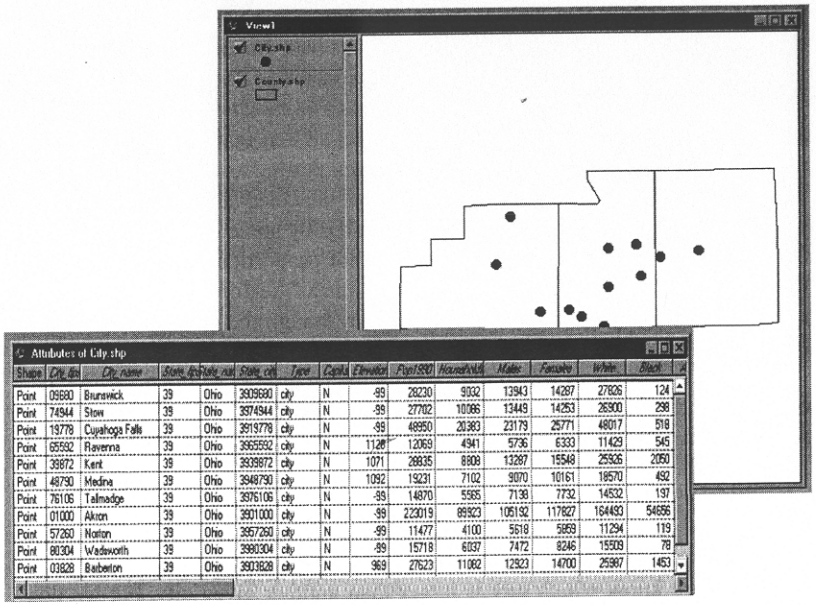


Figure 1.2 Map view and attribute table.

Let's first turn our attention to the attribute tables in geographic information systems (GIS) databases. As mentioned earlier, each record in an attribute table contains a number of fields. Since each record is associated with an object—or, in statistical terms, an observation in the map view—the data stored in the fields in a record are the information describing the associated object or observation in as many ways as the number of fields.

There are several types of data an attribute table can store. *Numerical data* are measured quantitatively. We can easily find examples of this type of data: areas of farms, precipitation amount at each monitoring station, population count for each county or city, and so on. This type of information is normally referred to as being measured at *ratio scale*. Data measured at ratio scale typically have a real zero value. For example, temperature at 0° Kelvin means no energy. By definition, there is no situation with a temperature below 0° Kelvin. Therefore, temperature measured in Kelvin is of ratio scale. Another obvious example is any data value measured in proportion, such as population densities, male/female ratios in high school classrooms, cost/benefit ratios, and so on. The absolute minimum for a proportion is 0. A proportion below 0 is not interpretable. Mathematical operations such as addition (+), subtraction (−), multiplication (∗), and division (/) can be applied to data that are measured at ratio scale.

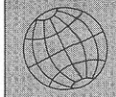
When there is no real zero value for the phenomenon being measured but the data are on a continuous scale, the data are measured at *interval scale*. Examples of this type of measurements include temperatures and elevations. A temperature

of 0°C does not mean that there is no temperature. It simply means that the temperature is at the position where it was defined as 0°C. In fact, it is 32°F when the temperature is 0°C. For elevation, 0 meter above sea level does not mean that there is no elevation. It simply means that it is at the same elevation as the average elevation of oceans. For data measured at this scale, all four mathematical operations are applicable. With both ratio and interval data, putting aside the preciseness of the measurement, we know the exact position of the observation along the continuous value line. While interval data are measured by some defined intervals, the differences between intervals sometimes are not proportional. For example, the difference between 80°F and 90°F is not the same as the difference between 80°F and 70°F in terms of how warm you feel, even though the difference is 10°F in both cases.

There are situations in which data simply give the order of the measured phenomenon. In this case, the data are said to be measured at *ordinal scale*. We can use 1, 2, 3, . . . , to represent the order or the ranking of cities in a state according to their population sizes. We can use descriptions or terms such as *high*, *medium*, or *low altitude* to represent the heights of mountains in a rough sense. Then observations are grouped into classes, and the classes follow an order. With ordinal data, mathematical operations such as $+$, $-$, $*$, or / cannot be applied. With only their ranks, we know which city is larger than another given city, but we don't know by how much. For ordinal data, we know the order of measured phenomena but we cannot add two measures to get another measure, that is, 1st + 2nd = 3rd.

Finally, we can measure phenomena in categorical form. This is known as measuring at the *nominal scale*. For this scale, no mathematical operations can be applied because nominal data only identify individual objects being measured. We don't even have the order between these objects, which we would know if they were measured at ordinal scale. We can easily think of many examples of data at this scale: street numbers of the houses along a street, telephone numbers of friends, flight numbers, zoning codes for different types of land use, and so on. Please note that the numbers at nominal scale, simply represent different things. They cannot be added or multiplied. Adding two telephone numbers will not result in another telephone number. Dividing house numbers by another house number is meaningless.

ArcView Notes



In working with an ArcView Table document, users need to understand the attribute data they use. For convenience, categorical data are often coded with numbers, such as 1, 2, 3, . . . for different types of land use. In those cases, the numbers should be treated as characters. FIPS and ZIP codes also consist of numbers, but these numbers should not be used in any numerical calculations. In ArcView, data measured at ratio or interval scales are of the type `number`, while data measured at ordinal or nominal scales are of the type `string`. There are ArcView functions such as `AsString` or `AsNumber` that can be used to convert attribute data between numerical and string forms.

GIS data sets are often large. A thematic GIS data layer of land use can easily contain more than 2,000 polygons for a large area. In a study matching potential customers and a newspaper's distribution system, the associated attribute table can easily have over 10,000 records in a moderately sized metropolitan region. Therefore, understanding the data will not be simple. For meaningful analysis of attribute data associated with map views, statistics are needed to describe, to summarize, or to find the relationships between attributes and geographical features.

In the subsequent sections, we will first look at ways to calculate descriptive statistics of attribute data using ArcView. These descriptive statistics indicate various statistical properties, such as central tendency and dispersion of the data. Statistics depicting the relationship between attributes will also be discussed.

1.1 CENTRAL TENDENCY

Often the first step in analyzing a set of numerical data is to measure their *central tendency*. The concept of central tendency uses a representative value to summarize the set of numerical data. For example, an *average* family income from a census tract gives an idea of the economic status of families in that census tract. Using the average family income to represent all income figures in that census tract allows us to quickly get an overall impression of its economic status.

In surveying students at Kent State University about their means of commuting, we found that most of them drive to the university. That specific type of commuting choice as nominal data, therefore, is the *mode* of commuting of those students. When comparing housing prices between neighborhoods, we often use the housing price in each neighborhood that stands close to the middle of the range of prices. Comparing the *middle* prices between neighborhoods allows us to avoid the pitfall of comparing the highest housing price in one neighborhood to the lowest housing price in another neighborhood.

The concept of central tendency is applied in everyday life. People use average Scholastic Aptitude Test (SAT) scores of freshman classes to compare how well a college does between years or to compare how colleges stand in relation to other colleges. We use phrases such as *typical weather* or *typical traffic pattern* to describe phenomena that happen most often. These are just a few of the many examples we can find everywhere.

1.1.1 Mode

The *mode* is the simplest measure of central tendency. It is the value that occurs most frequently in a set of data; consequently, that specific value is also known as

the *modal value*. For categorical or nominal data, the category that has the most observations or highest frequency is regarded as the mode. When working with ordinal data, the mode is usually the rank shared by two or more observations.

The modal value for interval or ratio data may not be very useful because that value may not occur more than once in a data set. Alternatively, researchers often *degrade* or *simplify* interval or ratio data to nominal scale by assigning individual data to one of the categories that they set up based on ranges of data values.

In Table 1.1, a total of 51 countries and territories in North and South America are listed by their population counts in 1990, areas in square miles, population densities, and categories of low/medium/high population density. The locations of the listed countries are shown in Figure 1.3.

To illustrate the use of mode and the effect of degrading interval/ratio data to nominal data, Table 1.1 first calculates the population density of each country by dividing the population count by its area. When examining the derived population densities, we cannot find a mode because no two or more countries have the same

TABLE 1.1 Population Density of Countries in the Americas

Country	Population	Area in Sq Miles	Population Density	Category
Anguilla	9,208	33	276	Low
Antiuga and Barbuda	65,212	179	365	Medium
Argentina	33,796,870	1,073,749	31	Low
Aruba	67,074	71	950	High
Bahamas, The	272,209	4,968	55	Low
Barbados	260,627	170	1,534	High
Belize	207,586	8,562	24	Low
Bermuda	59,973	15	3,941	High
Bolivia	7,648,315	420,985	18	Low
Brazil	151,525,400	3,284,602	46	Low
British Virgin Islands	18,194	63	290	Low
Canada	28,402,320	3,824,205	7	Low
Cayman Islands	31,777	107	297	Low
Chile	13,772,710	286,601	48	Low
Columbia	34,414,590	440,912	78	Low
Costa Rica	3,319,438	19,926	167	Low
Cuba	11,102,280	42,642	260	Low
Dominica	70,671	283	250	Low
Dominican Republic	759,957	18,705	415	Medium
Ecuador	10,541,820	99,201	106	Low
El Salvador	5,752,470	7,991	720	High
Falkland Islands	2,136	4,446	0	Low
French Polynesia	217,000	1,167	186	Low
Grenada	95,608	142	675	High
Guadeloupe	410,638	673	610	High
Guatemala	10,321,270	42,279	244	Low

TABLE 1.1 ***Continued***

Country	Population	Area in Sq Miles	Population Density	Category
Guyana	754,931	81,560	9	Low
Haiti	7,044,890	10,485	672	High
Honduras	5,367,067	43,572	123	Low
Jamaica	2,407,607	4,264	565	Medium
Martinique	374,574	425	881	High
Mexico	92,380,850	757,891	122	Low
Montserrat	12,771	41	314	Medium
Netherlands Antilles	191,572	311	617	High
Nicaragua	4,275,103	49,825	86	Low
Panama	2,562,045	28,841	89	Low
Paraguay	4,773,464	154,475	31	Low
Peru	24,496,400	500,738	49	Low
Pitcairn Islands	56	21	3	Low
Puerto Rico	3,647,931	3,499	1,043	High
St. Kitts and Nevis	42,908	106	404	Medium
St. Lucia	141,743	234	606	High
St. Pierre and Miquelon	6,809	94	72	Low
St. Vincent and the Grenadines	110,459	150	734	High
Suriname	428,026	56,177	8	Low
Trinidad and Tobago	1,292,000	1,989	650	High
Turks and Caicos Islands	14,512	189	77	Low
United States	258,833,000	3,648,923	71	Low
Uruguay	3,084,641	68,780	45	Low
Venezuela	19,857,850	353,884	56	Low
Virgin Islands	101,614	135	755	High

population density value. If we really want to identify the mode for this data set, the data have to be degraded from ratio scale to nominal scale.

If we define population densities below 300 persons per square mile as low, those between 300 and 600 persons per square mile as medium, and those over 600 persons per square mile as high, we can see from the last column in Table 1.1 that low density is the mode of this set of population densities. With the mode, we now have an overall impression of the levels of population density in these countries.

1.1.2 Median

The *median* is another measure of central tendency. In a set of data, the median is the *middle* value when all values in the data set are arranged in ascending or descending order.

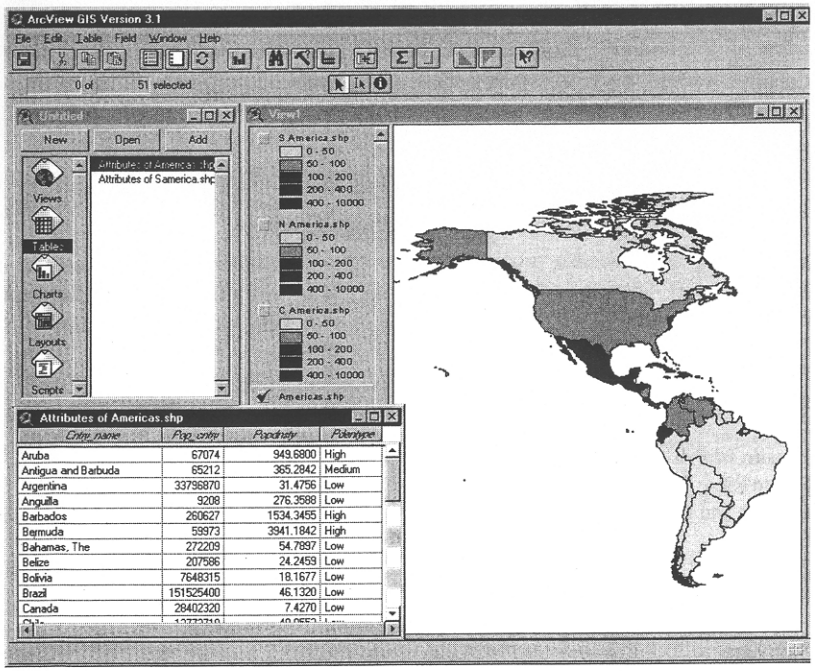


Figure 1.3 Population densities of the Americas.

To find the median of the population densities listed in Table 1.2, we first sort the table by population densities. Since 31 countries are listed in this table, the 16th value in the sorted sequence will be our median. The 16th entry in the list is 314 persons/mile^2 (Montserrat).

When the number of observations in a data set is odd, it is relatively simple to work out the median of the set. For a set of data containing an even number of values, the median is simply the value midway between the two middle values. For example, there are 12 countries listed in Table 1.3. The middle two values are 45 persons/mile^2 (Uruguay) and 46 persons/mile^2 (Brazil). The median of the set of 12 population densities will therefore be 45.5 persons/mile^2 since $(45 + 46)/2 = 45.5$ (persons/mile^2).

In general, a median can be found in any data set containing interval or ratio data. The median of a data set gives a value that is at the middle of the set. This median value is not severely affected by the inclusion of extremely large or extremely small values in the data set since it is defined by its position in the ordered sequence of data values.

1.1.3 Mean

The *mean* is the most commonly used measure of central tendency. It is the *average* value in a data set. This average is also known as *arithmetic mean* because of

TABLE 1.2 Population Density of Countries in Central America

Country	Population	Area in Sq Miles	Population Density	Category
Belize	207,586	8,562	24	Low
Bahamas, The	272,209	4,968	55	Low
United States	258,833,000	3,648,923	71	Low
Turks and Caicos Islands	14,512	189	77	Low
Nicaragua	4,275,103	49,825	86	Low
Panama	2,562,045	28,841	89	Low
Mexico	92,380,850	757,891	122	Low
Honduras	5,367,067	43,572	123	Low
Costa Rica	3,319,438	19,926	167	Low
Guatemala	10,321,270	42,279	244	Low
Dominica	70,671	283	250	Low
Cuba	11,102,280	42,642	260	Low
Anguilla	9,208	33	276	Low
British Virgin Islands	18,194	63	290	Low
Cayman Islands	31,777	107	297	Low
Montserrat	12,771	41	314	Medium
Antigua and Barbuda	65,212	179	365	Medium
St. Kitts and Nevis	42,908	106	404	Medium
Dominican Republic	7,759,957	18,705	415	Medium
Jamaica	2,407,607	4,264	565	Medium
St. Lucia	141,743	234	606	High
Guadeloupe	410,638	673	610	High
Netherlands Anitlles	191,572	311	617	High
Haiti	7,044,890	10,485	672	High
Grenada	95,608	142	675	High
El Salvador	5,752,470	7,991	720	High
St. Vincent and the Grenadines	110,459	150	734	High
Martinique	374,574	425	881	High
Aruba	67,074	71	950	High
Puerto Rico	3,647,931	3,499	1,043	High
Barbados	260,627	170	1,524	High

the way it is calculated. The mean is calculated by adding together all the values in a data set and then dividing the sum by the number of values. The equation for calculating the mean is

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where $\overline{X}$ (read as "X bar") denotes the mean of a group of values: $x_1, x_2, \ldots, x_n$. If there were 5 values in the data set, n would be 5. The symbol, $\sum_{i=1}^{n} x_i$, means

TABLE 1.3 Population Density of Countries in South America

Country	Population	Area in Sq Miles	Population Density	Category
Argentina	33,796,870	1,073,749	31	Low
Bolivia	7,648,315	420,985	18	Low
Brazil	151,525,400	3,284,602	46	Low
Chile	13,772,710	286,601	48	Low
Columbia	34,414,590	440,912	78	Low
Ecuador	10,541,820	99,201	106	Low
Guyana	754,931	81,560	9	Low
Suriname	428,026	56,177	8	Low
Paraguay	4,773,464	154,475	31	Low
Peru	24,496,400	500,738	49	Low
Uruguay	3,084,641	68,780	45	Low
Venezuela	19,857,850	353,884	56	Low

adding all 5 values as follows:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + x_5.$$

As an example, even though it is simple, Table 1.4 lists the levels of population density for Canada and the United States. The mean can be calculated as

$$\overline{X} = \frac{\sum_{i=1}^{2} x_i}{2} = \frac{7 + 71}{2} = \frac{78}{2} = 39 \text{ (persons/mile}^2\text{)}.$$

There are two density values, so $n = 2$. The mean is simply the average of the two values.

In Table 1.1, 51 countries are listed, so the mean population density is

$$\overline{X} = \frac{\sum_{i=1}^{51} x_i}{51} = \frac{x_1 + x_2 + \cdots + x_{51}}{51} = \frac{276 + 365 + \cdots + 755}{51} = 385.79.$$

TABLE 1.4 Population Density of Canada and the United States

Country	Population	Area in Sq Miles	Population Density	Category
Canada	28,402,320	3,824,205	7	Low
United States	258,833,000	3,648,923	71	Low

Similarly, in Table 1.2, the mean population density for Central American countries is

$$\overline{X} = \frac{\sum_{i=1}^{31} x_i}{31} = \frac{x_1 + x_2 + \cdots + x_{31}}{31} = \frac{24 + 55 + \cdots + 1534}{31} = 446.56.$$

For the South American countries, the mean population density can be calculated from Table 1.3 as

$$\overline{X} = \frac{\sum_{i=1}^{12} x_i}{51} = \frac{x_1 + x_2 + \cdots + x_{12}}{12} = \frac{8 + 9 + \cdots + 106}{12} = 43.82.$$

The above calculations of the mean of interval or ratio data are appropriate if all values are counted individually. But if observations are grouped into classes and all observations within each group are represented by a value, the calculation of the mean will be slightly different. The mean derived from the grouped data is usually called the *grouped mean* or *weighted mean*. Assuming that the value midway between the upper bound and the lower bound of each class is the representative value, x_i, and f_i represents the number of observations in the ith class, the weighted mean, $\overline{X}_w$, can be calculated as

$$\overline{X}_w = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i},$$

where k is the number of classes.

Before computers were widely available, the grouped mean was used to estimate the overall mean in a very large data set. In this procedure, observations are divided into groups according to their values. A value from each group, typically the midpoint between the lower and upper bounds of the group, is used to represent the group. When calculating the grouped mean, the number of observations in each group is used as the weight. This is also the reason why the grouped mean is often called the weighted mean.

Compared to the median, the mean is very sensitive to the inclusion of extreme values. Even if only one extremely large value is added to the data set, the average of all values in the data set will be pulled toward a larger value. As a result, the mean may be overestimated.

It is important to note that mode, median, and mean are three different measure of central tendency. When applied to a common data set, these three measures will give three different values. They differ in their definitions and in how they are calculated, so they have different meanings.

1.2 DISPERSION AND DISTRIBUTION

While the mean is a good measure of the central tendency of a set of data values, it does not provide enough information to describe how the values in a data set

are distributed. With the central tendency, we know what the average value is but we do not know how the values scatter around this average. Are the values similar to the mean, with only small differences? Do the values vary very differently from the mean? We don't know for sure unless we can measure how these values disperse or concentrate around the mean.

To illustrate the need for more information than the mean can give, let us use an example of the following series of numbers to compute their mean:

$$x_a: \quad 2, 5, 1, 4, 7, 3, 6$$
$$\overline{X}_a = \frac{2+5+1+4+7+3+6}{7} = 4.$$

The mean, 4, seems to be reasonably representative of these numbers. However, the following series of numbers also yields a mean of 4, with quite a different composition:

$$x_b: \quad 24, -18, 21, -43, 2, 33, -23$$
$$\overline{X}_b = \frac{24+(-18)+21+(-43)+2+33+(-23)}{7} = 4.$$

If we only know the means of these two sets of numbers, and have no further information, we might speculate that the two data sets are very similar to each other because their means are identical. However, by briefly examining the two number series, we know that the first series has a relatively narrow range centered at the mean, while the second series has a very wide range, that is, a highly dispersed set of values. Relying on the mean alone to compare these two series of values will yield misleading results. The truth is concealed by the large positive and negative values offsetting each other in the second series.

To better understand how values in a data set distribute, a number of descriptive statistics can be used. These include mean deviations, standard deviations, skewness, and kurtosis. These measures provide information about the degree of dispersion among the values and the direction in which the values cluster. Together they describe the distribution of numeric values in a data set so that analysts can understand the distribution or compare it with other distributions.

1.2.1 Mean Deviation

The first measure of dispersion is the *mean deviation*. It takes into account every value in the data set by calculating and summing the deviation of each value from the mean, that is, the difference between each value and the mean. The equation for calculating the mean deviation is

$$\text{Mean deviation} = \frac{\sum_{i=1}^{n} |x_i - \overline{X}|}{n}.$$

For data series x_a, the mean deviation is

$$\text{Mean deviation}_a = \frac{|2-4|+|5-4|+|1-4|+|4-4|+|7-4|+|3-4|+|6-4|}{7}$$
$$= \frac{2+1+3+0+3+1+2}{7} = 1.71.$$

The symbol $|x - \overline{X}|$ denotes the absolute difference between each value of x and the mean. So the equation first adds up all the absolute differences and then divides this number by the number of values to get the *average* of all absolute differences. This average absolute difference is the mean deviation. For the other series, x_b, the mean deviation is 20.57, which is quite different from 1.71 of series x_a.

This measure is simple to calculate and easy to understand. It provides a convenient summary of the dispersion of a set of data based on all values. In this manner, each value influences the mean deviation. A value that is close to the mean contributes little to the mean deviation. A value that is further away from the mean contributes more. With this measure, the presence of extremely large or extremely small values can be shown.

1.2.2 Variation and Standard Deviation

In calculating the mean deviation, we use the absolute values of the differences between data values and the mean as deviations because we need to make sure that positive deviations are not offset by negative deviations. Another way to avoid the offset caused by adding positive deviations to negative deviations is to square all deviations before summing them. The *variance* is one such measure. It can be calculated as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n},$$

where σ^2 is the variance. The i, n, and $\overline{X}$ are the same as those defined earlier. The equation for the variance actually calculates the average squared deviation of each value from the mean. While it is easier to understand it is not efficient in computation. A more computationally efficient formula for variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{X}^2.$$

This formula is more efficient because it minimizes the rounding error introduced by taking the differences and then squaring them.

Although variance measures dispersion in a data set, it is not commonly used because of its large numeric value. The deviations are squared before they are averaged. The process of squaring the deviations often leads to large numbers that cannot be compared directly to the original data values. As a remedy, the

square root of the variance is often used to describe the dispersion of a data set. This measure is known as the *root mean square deviation*, or simply *standard deviation*. It can be calculated as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}}$$

or

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{X}^2}.$$

The standard deviation value is similar in numeric range to the data values. It is used more often than variance because taking the root of the squared deviation returns the magnitude of the value to that of the data set.

As an example, Table 1.5 shows the calculation of the standard deviation of the population densities from the 12 South American countries. For these 12 population densities, the mean is 43.916 (rounded to 44). The variance is 748. Therefore, the standard deviation is 27.35 because $\sqrt{748} = 27.35$.

Similarly, the variance for the population density values of all countries in the Americas is 372,443.36, and for the Central American countries it is 122,734.90. The standard deviations are 350.34 for the Central American countries and 610.28 for all countries in the Americas.

TABLE 1.5 Variance and Standard Deviation

Country	Population Density x	$x - \bar{X}$	$(x - \bar{X})^2$
Argentina	31	−13	169
Bolivia	18	−26	676
Brazil	46	2	4
Chile	48	4	16
Colombia	78	34	1156
Ecuador	108	64	4096
Guyana	9	−35	1225
Suriname	8	−36	1296
Paraguay	31	−13	169
Peru	49	5	25
Uruguay	45	1	1
Venezuela	56	12	144
$\sum$	527		8977
$\bar{X}$	44		748

The standard deviation has another useful property to help describe how values in a data set distribute. Statistically, the inclusion of data values in a value range bounded by standard deviations results in a well-known relationship if the distribution of the data closely resembles that of a normal distribution:

1. About 68% of the data values are within 1 standard deviation on either side of the mean, that is, values within an interval bounded by $\overline{X} - \sigma$ and $\overline{X} + \sigma$.
2. About 95% of the data values are within 2 standard deviations on either side of the mean, that is, values within an interval bounded by $\overline{X} - 2\sigma$ and $\overline{X} + 2\sigma$.
3. About 99% of the data values are within 3 standard deviation on either side of the mean, that is, values within an interval bounded by $\overline{X} - 3\sigma$ and $\overline{X} + 3\sigma$.

Similar to the calculation of the mean, a weighted variance and the associated weighted standard deviation can be derived from data representing observations grouped into classes. Adopting the same notations used before, the weighted variance is defined as

$$\sigma_w^2 = \frac{1}{\sum_{i=1}^{k} f_i}\left[\sum_{i=1}^{k} f_i(x_i - \overline{X}_w)^2\right].$$

This intuitively meaningful formula also has its computational counterpart. For more efficient computation of the grouped variance, the following formula should be used:

$$\sigma_w^2 = \frac{1}{\sum_{i=1}^{k} f_i}\left[\sum_{i=1}^{k} f_i x_i^2 - \sum_{i=1}^{k} f_i(\overline{X}_w)^2\right].$$

Then the standard deviation for the grouped data is the square root of the weighted variance.

The mean and the standard deviation describe where the center of a distribution is and how much dispersion a distribution has. Together they provide a sketch of the distribution as a basis for understanding a data set or comparing multiple data sets.

1.2.3 Skewness and Kurtosis

For a set of values, the mean gives its central tendency. The standard deviation suggests how much the values spread over the numeric range around the mean. There are also other characteristics of a numeric distribution that can be described by using additional measures. These include *skewness*, which measures the *directional bias* of a numeric distribution in reference to the mean, and *kurtosis*, which measures the *peakness* of a numeric distribution. Combining the mean, the

standard deviation, the skewness, and the kurtosis, we have a set of descriptive statistics that can give rather detailed information about a given numeric distribution.

To understand how the skewness and kurtosis of a numeric distribution are calculated, it is necessary to discuss the concept of *frequency distribution*. The frequency distribution is often shown in a *histogram* in which the horizontal axis shows the numeric range of the data values and the vertical axis shows the frequency, that is, the number of values in each interval. Figure 1.4 shows five examples of frequency distributions with different levels of skewness and kurtosis. At the top is a symmetric distribution with low skewness and medium kurtosis. The two skewed distributions in the middle row show distributions with directional bias but low kurtosis. The two distributions in the bottom row show the

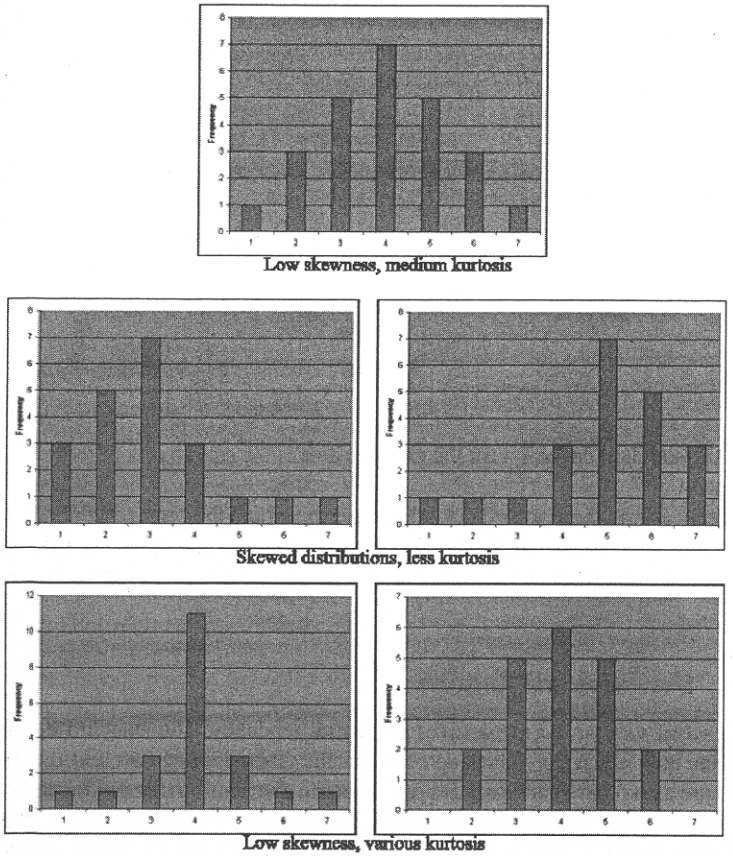


Figure 1.4 Frequency distribution: skewness and kurtosis.

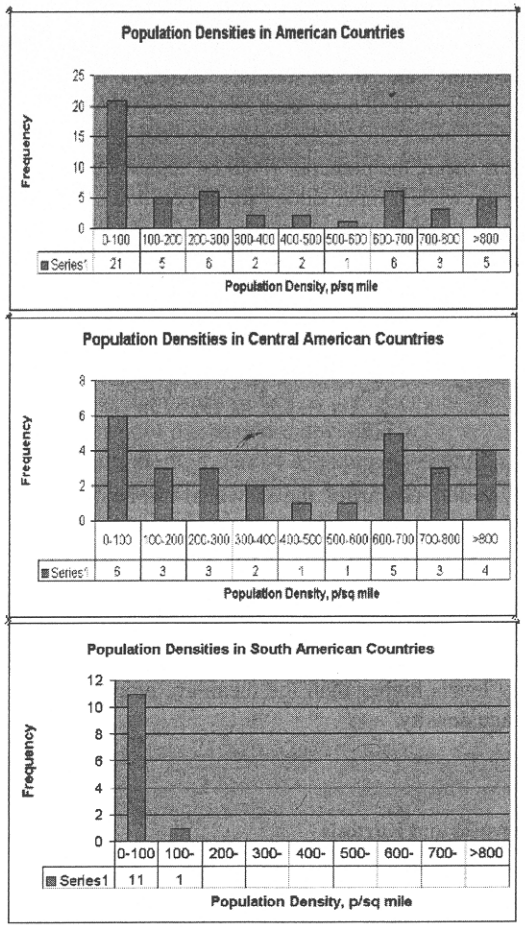


Figure 1.5 Frequency distribution of population densities of American countries.

difference between two kurtosis values. Figure 1.5 shows the frequency distributions of population density of the three America regions.

Skewness measures the extent to which the bulk of data values in a distribution are clustered to one side or the other side of the mean. When most values are less than the mean, the distribution is said to be *positively skewed*. Alternatively, a *negatively skewed* distribution is a distribution in which the bulk of the values are greater than the mean. Specifically, skewness can be calculated by

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^3}{n\sigma^3}.$$

where x, $\overline{X}$, σ, and n are the same as defined earlier. Notice that the measure of skewness is based on the cubic value of the deviations from the mean (or mean deviation) and the cubic value of the standard deviation, σ.

Because σ is positive, the denominator of the skewness formula is always positive. The numerator, however, can be positive or negative. If most of the values are smaller than the mean, the numerator will be negative and thus the distribution will be positively skewed. If most values are larger than the mean, the numerator will be positive and the skewness measure will be negative. The skewness of a symmetric distribution is 0 (zero).

Take the data set of population densities in South American countries as an example. Table 1.6 shows that

$$\sum_{i=1}^{n}(x_i - \overline{X})^3 = 191873.$$

Since σ is 27.35, as derived earlier, the skewness can be calculated as follows:

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^3}{n\sigma^3} = \frac{191{,}873}{12 \times 27.35^3} = \frac{191{,}873}{245{,}501} = 0.7816.$$

The distribution is thus moderately skewed in the positive direction, that is, countries with density levels higher than the mean are more frequent than countries with below-average density.

TABLE 1.6 Skewness and Kurtosis

Country	Population Density x	$x - \overline{X}$	$(x - \overline{X})^2$	$(x - \overline{X})^3$	$(x - \overline{X})^4$
Argentina	31	−13	169	−2197	28561
Bolivia	18	−26	676	−17576	456976
Brazil	46	2	4	8	16
Chile	48	4	16	64	256
Colombia	78	34	1156	39304	1336336
Ecuador	108	64	4096	262144	16777216
Guyana	9	−35	1225	−42875	1500625
Suriname	8	−36	1296	−46656	1679616
Paraguay	31	−13	169	−2197	28561
Peru	49	5	25	125	625
Uruguay	45	1	1	1	1
Venezuela	56	12	144	1728	20736
$\sum$	527		8977	191873	21829525
$\overline{X}$	44		748		

Skewness is most useful when it is used to compare distributions. For example, two distributions can have similar means and similar standard deviations, but their skewness can be very different if there are different directional biases.

With kurtosis, the extent to which values in a distribution are concentrated in one part of a frequency distribution can be measured. If the bulk of values in a distribution have a high degree of concentration, the distribution is said to be very *peaky*. Alternatively, a *flat* distribution is one without significant concentration of values in one part of the distribution.

The kurtosis is usually measured by the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^4}{n\sigma^4} - 3.$$

The kurtosis is based on the fourth power of deviations of the values from their mean and the fourth power of the standard deviation, σ.

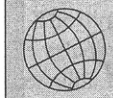
By subtracting 3 from the first part of the kurtosis equation, we structure the calculation of kurtosis so that a symmetrical, bell-shaped distribution has a value close to 0. In this way, a peaky distribution will have a positive kurtosis value and a flat distribution will have a negative kurtosis value.

Still using the population density values in South America, Table 1.6 gives $\sum_{i=1}^{n}(x_i - \overline{X})^4 = 21829525$ and $\sigma^2 = 748$. Therefore,

$$\begin{aligned}\text{Kurtosis} &= \frac{\sum_{i=1}^{12}(x_i - \overline{X})^4}{n\sigma^4} - 3 \\ &= \frac{21{,}829{,}525}{12 \times 748^2} - 3 = 3.25 - 3 \\ &= 0.25\end{aligned}$$

giving a distribution that is slightly peaky.

ArcView Notes



When using ArcView to calculate descriptive statistics, the procedures are conducted in Table documents. There is a set of basic descriptive statistics users can derive by using the following steps:

1. Open the Table document from the Project window or open a theme table by selecting the menu item **Theme/Table** after adding a theme to a View document and have the theme active.
2. Click the title button of the field that contains the numerical values for which the statistics are needed.
3. Select the menu item **Field/Statistics** to have ArcView calculate the statistics, including sum (or total), count (i.e., n),

mean, maximum, minimum, range, variance, and standard deviation of the field.

While these basic descriptive statistics provided by the original ArcView are useful, they are not complete because the **Field/Statistics** does not produce skewness, kurtosis, or other values discussed in this chapter. To find these values, you can use a project file available on the companion website to this book that was developed to provide all descriptive statistics discussed in this chapter.

The project file, `ch1.apr`, which can be found in the `AVStat\Chapter1\Scripts\` directory in the archive downloaded from the support website, can be used to calculate additional statistics using the following steps:

1. Start a new project by selecting the menu item **File/Open Project** from the Project window.
2. Navigate to `\AVStat\Chapter1\Scripts\` and select `Ch1.apr`.
3. Add themes from your data directory to the view. You can add more than one theme, but make the theme for which you want to calculate additional statistics active.
4. Open the Table document of the active theme and click the title button of the field that contains the numerical values you want to use. Be sure that the field is a numeric field, not a string field.
5. Select the menu item **Field/Other Statistics** to have ArcView calculate additional statistics.

Figure 1.6 shows the location of the menu item and the results of calculating additional statistics, including skewness and kurtosis. Please note that results from the calculation using the **Field/Other Statistics** on data accompanied by the book will be different from the results reported in the text and in Tables 1.5–1.6. These discrepancies are rounding errors introduced when rounded values were used in the text and tables for illustrations.

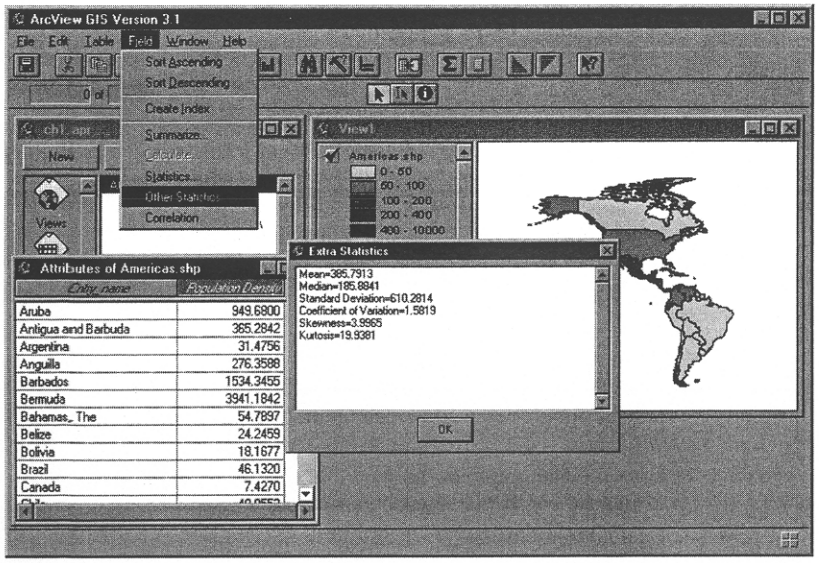


Figure 1.6 Function for calculating additional statistics.

1.3 RELATIONSHIP

The descriptive statistics discussed in the previous sections are useful for understanding and comparing how values distribute within one data set or between data sets. The mean, standard deviation, skewness, and kurtosis, although providing a basis for comparing different distributions, cannot measure the relationship between distributions quantitatively. To do so, we will need to apply the technique that this section discusses. This technique is based on the concept of *correlation*, which measures statistically the *direction* and the *strength* of the relationship between two sets of data or two variables describing a number of observations.

Given two counties in the same region where a similar tax code and similar developmental strategies have been applied, a comparison of their average family income figures will give us an impression of how well each county performs economically. If we consider subscribing to the concept that more spending on higher education will result in better economic progress, a look at the relationship between spending on higher education and some indicators of economic status will provide a potential answer. For this type of comparison, we typically measure how strongly the values of these two variables are related and the direction of their relationship.

The direction of the relationship of two variables is positive (or *direct*) if one of the values in a variable behaves similarly to another variable. For example, when the value of one variable for a particular observation is high, the value of the other variable for that observation is likely to be high. Alternatively, a *negative* (or *inverse*) relationship between two variables indicates that the value of one variable increases when the value of the other variable decreases. Of course, the stronger the relationship is, the more predictable this pattern will be.

In Figure 1.7, there are three diagrams that plot pairs of values as points in what are called *scatterplots*. In the top diagram, the relationship between the total length of motorways in the United Kingdom in 1993 is positively related to the total number of vehicles by region in 1991. Notice that the points show a pattern

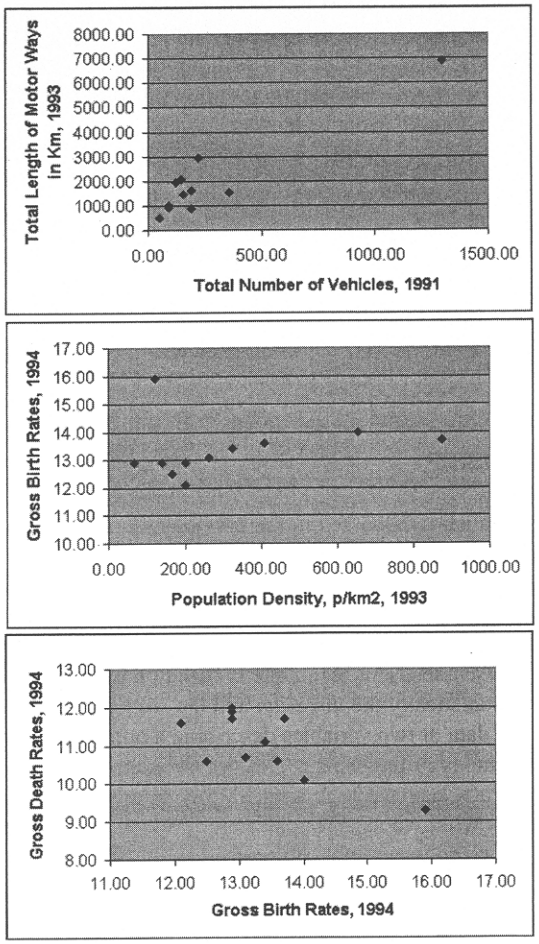


Figure 1.7 Sample relationships.

that runs from lower left to upper right. Taking any region as an example, it seems that a higher total number of vehicles in a region is often associated with a higher total length of motorways. In the lower diagram, the gross death rates and gross birth rates in regions of the United Kingdom in 1994 clearly show an inverse relationship. The pattern, as illustrated by the points in this scatterplot, indicates that a region with a high birth rate have a low death rate. The middle diagram shows the relationship between birth rate and population density. In this case, there does not seem to be a relationship at all; the trend as depicted by the points is flat, forming an almost horizontal line.

Beyond the direction of the relationship between two sets of data values, the strength of the relationship can be estimated quantitatively, and of course visually, from the scatterplots. In Figure 1.7, the top plot indicates a stronger relationship than the relationship described by the bottom plot. This is because the point distribution in the top plot is less scattered than that of the bottom plot.

Quantitatively, we can measure the direction and the strength of the relationship between two sets of data values by calculating the correlation coefficient. In this section, we will discuss only what is known as the *product-moment correlation coefficient* (or *Pearson's correlation coefficient*). This coefficient works best for *interval/ratio scale* data. For data measured at nominal or ordinal scales, other coefficients (χ^2 and Spearman's rank coefficient) should be applied.

The Pearson's correlation coefficient, r, between two variables, x_i and y_i, $i = 1, 2, \ldots, n$, can be calculated by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{(n-1)S_x S_y},$$

where S_x and S_y are the standard deviations of x and y, respectively.

The numerator is essentially a covariance, indicating how the two variables, x and y, vary together. Each x_i and y_i is compared with its corresponding mean. If both x_i and y_i are below their means, the product of the two negatives will be a positive value. Similarly, if both are large, the product will be positive, indicating a positive correlation. If x_i is larger than the mean but y_i is smaller than the mean, the product will be negative, indicating an inverse relationship. The sum of all these covariations reflects the overall direction and strength of the relationship. For computational efficiency, we prefer to use the formula

$$r = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{X} \bullet \overline{Y}}{S_x S_y},$$

where $\overline{X}$ and $\overline{Y}$ are means for x and y, respectively, and S_x and S_y are standard deviations for x and y, respectively, defined as

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} x^2}{n} - \overline{X}^2}$$

and

$$S_y = \sqrt{\frac{\sum_{i=1}^{n} y^2}{n} - \overline{Y}^2}.$$

This coefficient is structured so that the sign of the value of r indicates the direction of the relationship as:

$r > 0$ when the relationship between the two variables is a direct (or positive) one,

$r < 0$ when the relationship between the two variables is an inverse (or negative) one, and

$r \approx 0$ when there is no relationship between the two variables.

The absolute value of r indicates the strength of the relationship with a numeric range of:

$r = -1$ for the strongest or perfectly inverse relationship and

$r = 1$ for the strongest or perfectly direct relationship.

TABLE 1.7 Data for Pearson's Product-Moment Correlation Coefficient

Regions ($n = 11$)	Total Length of Motorways x	Total No. of Vehicles y	x^2	y^2	xy
North	152	923	23,104	851,929	140,296
Yorkshire and Humberside	289	1629	83,521	2,653,641	470,781
East Midlands	185	1438	34,225	2,067,844	266,030
East Anglia	22	890	484	792,100	19,580
South East	919	6893	844,561	47,513,449	6,334,667
South West	302	1960	91,204	3,841,600	591,920
West Midlands	378	2066	142,884	4,268,356	780,948
North West	486	2945	236,196	8,673,025	1,431,270
Wales	120	979	14,400	958,441	117,480
Scotland	285	1545	81,225	2,387,025	440,325
Northern Ireland	113	483	12,769	233,289	54,579

$n = 11$

$\sum x = 3251$

$\overline{X} = 295.55$

$\sum x^2 = 1{,}564{,}573$

$\overline{X}^2 = 87{,}349.81$

$S_x = \sqrt{\frac{1{,}564{,}753}{11} - 87{,}347.12} = 234.31$

$\sum y = 21751$

$\overline{Y} = 1{,}977.36$

$\sum y^2 = 74{,}240{,}699$

$\overline{Y}^2 = 3{,}909{,}952.57$

$S_y = \sqrt{\frac{74{,}240{,}699}{11} - 3{,}939{,}952.57} = 1{,}684.99$

$\sum xy = 10{,}647{,}876$

$$r = \frac{\frac{10{,}647{,}876}{11} - 295.55 \times 1{,}977.36}{234.31 \times 1{,}684.99} = \frac{967{,}988.73 - 584{,}408.75}{394{,}810} = \frac{383{,}579.98}{394{,}810} = 0.97$$

Table 1.7 shows an example of the calculation of Pearson's product-moment correlation coefficient. The two variables are:

x: Total number of private vehicles (except tractors and motorcycles) in 1991

y: Total length of motorways in kilometers in 1993.

The resulting correlation coefficient 0.97, indicating a very strong, positive relationship between the two variables. Specifically, a region that has a larger number of vehicles also has a longer total motorway.

ArcView Notes

`Ch1.apr` also allows calculation of the correlation coefficient between any two fields in an attribute table. The procedure for doing so with ArcView is as follows:

1. Start ArcView and select the menu item **File/Open Project** from the Project window.
2. Navigate to the directory where you might have copied the project file to open `Ch1.apr` by highlighting it and clicking the **OK** button.
3. Use the **Add Theme** button to bring in any theme that contains the attribute fields you want to calculate the correlation coefficients. Make sure that the theme is the active theme and then open its attribute table by clicking the **Table** button under the **Theme** menu.
4. Select the **Field/Correlation** menu item.
5. From the pop-up dialogue box, select the first variable from the drop-down list of attributes. Click the **OK** button to proceed.
6. From the next pop-up dialogue box, select the second variable and then click the **OK** button to calculate the correlation coefficient.

As shown in Figures 1.8 and 1.9, the steps for calculating a correlation coefficient are in fact quite straightforward. A word of caution: this procedure should not be applied to fields in an attribute table that contain either nominal or ordinal data or strings. Note that the correlation function is also accessible under the **Statistics/Correlation** menu item when the View is active.

1.4 TREND

The previous section focused on the technique for measuring the direction and strength of the relationship between two variables. In this section, we will discuss the technique for measuring the trend, as shown by the relationship between two

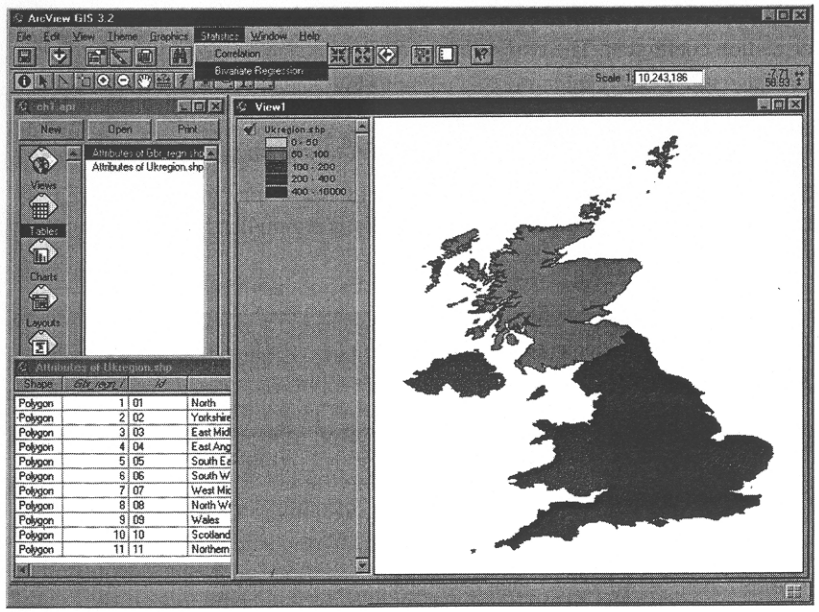
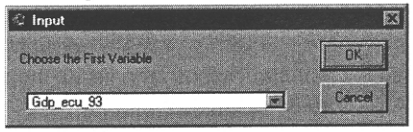
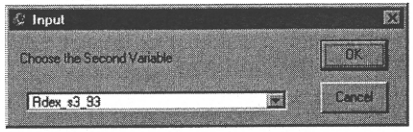


Figure 1.8 Statistical function of correlation.

Selecting the first variable:



Selecting the second variable:



The resulting correlation coefficient:

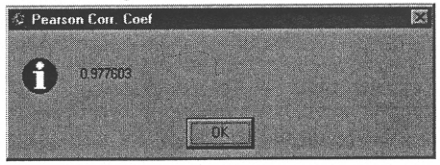


Figure 1.9 Steps for calculating the correlation coefficient.

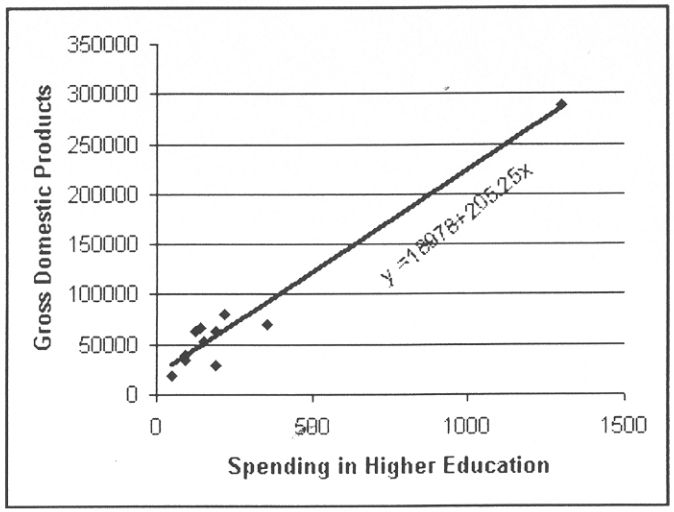


Figure 1.10 Simple linear regression model.

variables. The technique shows the *dependence* of one variable on another. Going beyond knowing the strength and direction of the relationship, the technique for measuring the trend allows us to estimate the likely value of one variable based on a known value of another variable. This technique is known as the *regression* model. Although the regression model does not imply a *causal relationship*, it provides the information necessary for the *prediction* of one variable by another.

To illustrate what measuring the trend of the relationship between two variables means, Figure 1.10 shows the relationship between expenditure on higher education in 1993 and gross domestic product (GDP) (in million European Community Units (ECU)) in 1993 by region in the United Kingdom. It can be seen that there is a positive, strong correlation between the two variables. For regions where GDP was high, spending on higher education was also high. Alternatively, the relationship shows that regions with less spending on higher education are also regions with lower GDP. The straight line running between the data points is the regression line. Because this trend line is a linear one, this type of regression model is also called *simple linear regression* or *bivariate regression* because it involves only two variables.

The simple linear regression model is the simplest form of regression model. It is generally represented as

$$Y = a + bX,$$

where

Y is the dependent variable,

X is the independent variable,

a is the intercept, and

b is the slope of the trend line.

The variable that is used to predict another variable is the independent variable (X). The variable that is predicted by another variable is the dependent variable (Y). The intercept is the value of the dependent variable when the value of the independent variable is zero or the value of Y when the regression line intersects the y-axis. Finally, the slope is the rate of change of the dependent variable's value as a per unit change of the independent variable.

The procedure for finding the trend line is to *fit a regression line* among all data points. Specifically, the procedure is to find the value of a (intercept) and the value of b (slope) in the regression line $Y = a + bX$:

$$b = \frac{\sum xy - n\overline{XY}}{\sum x^2 - n\overline{X}^2}$$

and

$$a = \overline{Y} - b\overline{X},$$

where $\sum xy$ is the sum of the products $x_i y_i$, $i = 1, 2, \ldots, n$; $\sum x^2$ is the sum of squared values of the independent variables; and $\overline{X}$ and $\overline{Y}$ are the means of the independent and dependent variables, respectively.

As an example, Table 1.8 shows the steps for calculating the values of a and b to construct the trend line in a simple linear regression model. The resulting regression model is

$$\hat{y} = 18,977.96 + 205.25x.$$

The intercept is 18,977.96 and the slope is 205.25. With this model, we can calculate a set of estimated values for the dependent variable using the values we have for the independent variable. The results of this estimation are listed in Table 1.9. As shown in this table, there are some deviations between the observed values and the estimated values of the dependent variable. The deviations between the observed and predicted values are known as *residuals*. Ideally, a perfect regression model will have zero residuals. The larger the residuals, the less powerful the regression model is. When the residuals are small, we typically say that the regression line is a *good fit*.

Since regression models are not always equally powerful, how do we know if a regression model is a good enough fit of the data? To answer this question, we need to calculate a *coefficient of determination*, usually denoted as r^2. The coefficient of determination is the ratio between the variance of the predicted values of the dependent variable and the variance of the observed values of the

TABLE 1.8 Data for Simple Linear Regression Model

Name	1993 GDP in Million ECU y	1993 Expenditure on Higher Education x	x^2	xy
North	38,494.4	92.3	8,519.29	3,553,033.12
Yorkshire and Humberside	63,701.6	191.0	36,481.0	12,167,005.60
East Midlands	52,937.1	155.1	24,056.01	8,210,544.21
East Anglia	29,552.9	188.5	35,532.25	5,570,721.65
South East	288,479.1	1,297.5	1,683,506.25	374,301,632.25
South West	62,739.0	123.1	15,153.61	7,723,170.90
West Midlands	67,161.5	142.3	20,249.29	9,557,081.45
North West	80,029.7	219.2	48,048.64	17,542,510.24
Wales	34,028.7	91.0	8,281.0	3,096,611.70
Scotland	69,601.0	355.1	126,096.01	24,715,315.10
Northern Ireland	18,033.3	48.7	2,371.69	878,221.71

$\sum y = 804,758.30$ | $\sum y = 2,903.80$

$n = 10$ | $n = 10$

$\overline{Y} = 73,159.85$ | $\overline{X} = 263.98$

$\sum xy = 467,315,847.93$ | $\sum x^2 = 2,008,295.04$

$$b = \frac{467,315,847.93 - 11(263.98)(73,159.85)}{2,008,295.04 - (11)(263.98)(263.98)} = \frac{254,875,738.7}{1,241,755.2} = 205.25$$

$$a = 73,159.85 - (205.5)(263.98) = 18,977.96$$

$$\hat{y} = 18,977.96 + 205.25x$$

dependent variable. Specifically, the coefficient of determination can be calculated as

$$r^2 = \frac{S_{\hat{y}}^2}{S_y^2},$$

where $S_{\hat{y}}^2$ is the variance of the predicted values of the dependent variable, also known as *regression variance*, and S_y^2 is the variance of the observed values of the dependent variable, also known as *total variance*.

In the lower part of Table 1.9, the coefficient of determination is calculated as $r^2 = 0.9799$. Converting this ratio to a percentage, we can say that 97.99% of the variance of the dependent variable is accounted for or captured by the regression. Consequently, the higher the r^2 value is, the better the regression model's fit to the data values. Also note that the square root of the coefficient of determination is r, which is the Pearson's product-moment correlation coefficient discussed in the previous section.

TABLE 1.9 Regression Model: Observed Values, Predicted Values, and Residuals

Regions	x	Observed y	Predicted $\hat{y}$	Residuals $y - \hat{y}$
North	92.3	38,494.4	39,722.54	−1,228.14
Yorkshire and Humberside	191.0	63,701.6	58,180.71	5,520.89
East Midlands	155.1	52,937.1	50,812.24	2,124.87
East Anglia	188.5	29,552.9	57,667.59	−28,114.69
South East	1,297.5	288,479.1	285,289.84	3,189.26
South West	123.1	62,739.0	44,244.24	18,494.77
West Midlands	142.3	67,161.5	48,185.04	18,976.47
North West	219.2	80,029.7	63,968.76	16,060.94
Wales	91.0	34,028.7	37,655.71	−3,627.01
Scotland	355.1	69,601.0	91,862.24	−22,261.24
Northern Ireland	48.7	18,033.3	28,973.64	−10,940.34

$$\sum y^2 = 1.13615 \times 10^{11}$$

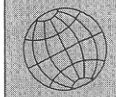
$$\sum \hat{y}^2 = 1.11328 \times 10^{11}$$

Total variance: $S_y^2 = \dfrac{\sum y^2}{n} = 1.0329 \times 10^{10}$

Regression variance: $S_{\hat{y}}^2 = \dfrac{\sum \hat{y}^2}{n} = 1.0121 \times 10^{10}$

Coefficient of determination: $r^2 = \dfrac{1.0121 \times 10^{10}}{1.0329 \times 10^{10}} = 0.9799$

ArcView Notes



Similar to the calculation of the correlation coefficient, the project file, `Ch1.apr`, can be used to load themes and to calculate the simple linear regression model. The menu item for calculating regression can be found in the Statistics menu for View documents.

To use `Ch1.apr` for calculating a regression model, do the following:

1. Start ArcView and use **File/Open Project** to load `Ch1.apr`, as before.
2. Start a new View document and use the **Add Theme** button to bring in themes you want to use.
3. From the menu of the View document, choose **Statistics/Bivariate Regression** to initiate the procedure.
4. In the pop-up dialog box, choose the field for the independent variable and then click **OK** to proceed.

5. In the next pop-up dialog box, choose the field for the dependent variable and then click **OK** to proceed.
6. The results are shown in another pop-up dialog box, including the values for the coefficient of determination, intercept, and slope of the regression model.
7. After you click the **OK** button in the dialog box that shows the results, another dialog box appears to ask if a scatterplot of the data should be created. Click **Yes** to create it or **No** to end the calculation. When the scatterplot appears, you can expand the window horizontally to extend the plot in the X direction.

As an example, Figure 1.11a shows the **Statistics/Bivariate Regression** menu item. Figure 1.11b shows the steps in performing the calculation of bivariate regression and its results.

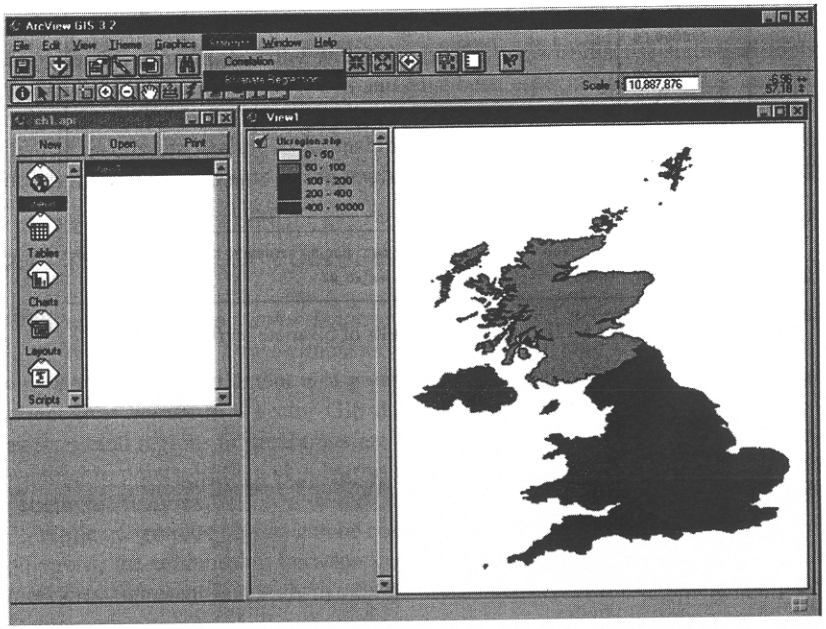


Figure 1.11a The bivariate regression menu item.