BASIC STATISTICS:

A USER ORIENTED APPROACH

(Manuscript)

Spyros MAKRIDAKIS
and
Robert L. WINKLER

CHAPTER 3

CHAPTER 3

SUMMARY MEASURES

3.1 Introduction

In descriptive statistics, the objective is to describe sets of data. At one extreme, we can simply present an entire set of data. As noted in Chapter 2, however, this is usually very impractical. To reduce the burden placed on individuals who would like to know something about a set of data, methods of summarizing the data have been developed. The frequency distributions discussed in Chapter 2 represent summarizations of data, and one can quickly obtain some information about a set of observations by looking at a frequency distribution. Graphical presentations such as histograms are especially effective in conveying information.

Often we want to summarize the data further, considering not the entire frequency distribution but some particular aspect of the distribution. For example, in terms of the data in Table 2.1, we might be interested primarily in the average age of the students. When you get an examination back, you might like to know not just your score but also the average score so that you can see if you are above or below the average. If you can also find out the highest and lowest scores, you will have an even better idea of how you performed relative to the rest of the class. In newspapers and magazines we read about the average cost of houses in various cities, median incomes for different professions, batting averages for baseball players, the average mileage per gallon obtained with certain cars, the highest and lowest temperatures ever recorded in our community on July 4, and so on.

A summary measure such as an average provides information about a set of data in terms of a single number. Obviously some information is lost - a single number cannot tell us all there is to know about a large set of data. However, summary measures can be quite informative while still being very easy to communicate. Knowing the average score, the low score, and the high score on a test is not the same as knowing the entire distribution of scores, but these three summary measures are easy to understand and they can tell you quite a bit about the test scores.

In this chapter, we focus on two types of summary measures: measures of location and measures of dispersion. A measure of location tells us something about the "typical" member of a data set (for example, a "typical" age or a "typical" test score). In terms of a frequency distribution, a measure of location tells us something about the center of the distribution. The mean, median, and mode are commonly-used measures of location.

A measure of dispersion tells us something about the variation in a set of data. The average income provides information about the center of a distribution of incomes, but how concentrated or spread out is this distribution? Are most of the incomes right around the average, or are there many incomes that are quite a bit lower or higher than the average? The range, variance, and standard deviation are commonly-used measures of dispersion.

Measures of location are presented in Section 3.2, and measures of dispersion are covered in Section 3.3. In Section 3.4, we discuss the distinction between a population and a sample and between parameters (summary measures of a population) and statistics (summary measures of a sample). Various formulas for means, variances, and standard deviations of both populations and samples are given in Section 3.5. Section 3.6 involves making some statements about frequencies (for example, what proportion of the data are within a specified distance from the mean) just on the basis of the mean and standard deviation.

3.2 Measures of Location

In everyday language, the term "average" is used to indicate an average, or typical, value in a set of data. Sometimes different people talk about different types of averages. The mayor of a city may report that the average family income in the city is $16,200, and the Chamber of Commerce may, at the same time, release a figure of $19,600 for the average family income. It is possible that the mayor and the Chamber of Commerce are using the same set of data concerning incomes but that they are choosing different measures to call the "average". In this section we present three types of "averages", or measures of location, and we try to indicate the differences among these measures. An understanding of such differences will make you a more knowledgeable consumer of statistics in the sense of being able to interpret published numerical claims.

3.2.1 The Mean

Probably the most common use of the term "average" is to represent an arithmetic average of a set of numbers. In statistics, we call an arithmetic average a *mean*. The mean of a set of numbers is found by adding up all of the numbers and dividing this sum by how many numbers there are in the set. For example, if we consider the 230 students whose ages are given in Table 2.1, the mean age is

$$\frac{27 + 25 + 25 + 32 + 30 + 29 + \ldots + 29 + 29 + 25 + 37 + 34 + 29.}{230}$$

The sum of the 230 ages is 6403, and the mean is therefore

$$\frac{6403}{230} = 27.84.$$

In order to express the mean in terms of a formula, we need to introduce some notation. Let x_1 represent the first number in the set of data, so that $x_1 = 27$ for the data on student ages. Similarly, let x_2 represent the second number in the data set ($x_2 = 25$ for the ages), let x_3 represent the third number ($x_3 = 25$ for the ages), and so on. For our age data, the last number in data set is $x_{230} = 29$. Thus, the ages can be represented in symbols as

$$x_1, x_2, x_3, \ldots, x_{229}, x_{230}.$$

The mean can be written in the form

$$\frac{x_1 + x_2 + x_3 + \ldots + x_{229} + x_{230}}{230}.$$

Other sets of data may have more or less than 230 numbers. We denote the size of the data set (that is, the number of ages, or incomes, or whatever we happen to be measuring) by the letter N. Now we can express the mean in terms of a general formula:

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_{N-1} + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

To avoid having to write out the sum $x_1+x_2+x_3+ \ldots x_{N-1} + x_N$, we employ the Greek capital Σ (sigma) to denote summation. Thus $\sum_{i=1}^{N} x_i$ tells us to add up the x_is (that is, the numbers in the data set), starting with x_1 (i=1) and ending with x_N (i=N).

The mean is the most widely used summary measure in statistics. It is easy to compute, and it is a familiar measure that is not difficult to interpret. Moreover, means can be combined without difficulty. If we know the mean score on an examination for one class of 65 students and the mean score of another class of 83 students on the same examination, we can find the overall mean score of the 65 + 83 = 148 students:

$$\text{Overall Mean} = \frac{65\ (\text{Mean for First Class}) + 83\ (\text{Mean for Second Class})}{148}.$$

Finally, the mean has certain desirable properties which make it an attractive measure to use in statistical inference. For inferential purposes, the mean is used much more often than other measures of lacation, as you shall see when inferential statistics is discussed in later chapters.

Purely as a descriptive measure of location, the mean sometimes has drawbacks. Consider the following set of nine incomes (in thousands of dollars):

14, 16, 12, 865, 14, 18, 14, 20, 17.

The mean income is:

$$\text{Mean} = \frac{14+16+12+865+14+18+14+20+17}{9} = \frac{990}{9} = 110.$$

This mean value, 110, hardly seems typical of the set of data. Eight out of the nine incomes are less than or equal to 20. Unfortunately, the one very large income, 865, has a considerable influence on the mean. In general, the mean is unduly influenced by extreme values.

When there are no extreme values and the frequency distribution is relatively symmetric (that is, the left-hand side of the distribution looks roughly like a mirror image of the right-hand side), the mean will be in the center of the distribution. The mean age for our 230 students is 27.84, and this seems to be in the center of the histogram of ages (Figure 2.1). When extreme values on one side of the histogram are not balanced out by extreme values on the other side, then the mean moves away from the center and toward the extreme value (as in the income example). Then the median appears to be a more "typical" value.

3.2.2 The Median

A way of avoiding the influence of extreme values is by not being concerned with how large or small each of the numbers is, but only with the middle value when the numbers are put in order. For instance, we can order the nine incomes given above from the smallest to the largest. This will give us the following list:

12, 14, 14, 14, 16, 17, 18, 20, 865
↑
Middle of the nine incomes when they are ordered from smallest to largest.

The number 16 is the center, or middle number, of the nine ordered numbers. Precisely half of the remaining eight incomes are below 16, and half are above 16. Here 16 is the *median* income.

Median = the middle value when a set of data is ordered from smallest to largest.

The biggest advantage of the median is that it is not influenced by extreme values. For instance, if the largest income, 865, doubles and becomes 1730, the median will not change. It will still be 16. The mean, on the other hand, become 206.1, a large increase from the previous mean of 110. This further illustrates how the mean is influenced by extreme values.

How can the median be found when the number of data is even instead of odd as in the case we just looked at? Suppose that we have the ten numbers 9, 6, 12, 15, 7, 18, 13, 5, 3, and 5. First we need to order them. This results in the following list:

$$3, 5, 5, 6, \underline{|7, 9,|}\ 12, 13, 15, 18$$

↑

Middle two numbers

↑

$$\text{Mean of } 7 + 9 = \frac{7+9}{2} = 8$$

↑ ↑

8 ← Median

Here there are *two* numbers in the middle, since there are four numbers below 7 and four numbers above 9. The median is the average of the middle two numbers - which is 8. Note that 8 is a middle value in the sense that 5 numbers are below 8 and 5 numbers are above 8.

The median gives us a "typical" value in the center of the frequency distribution, since it is not affected by extreme values. However, there are some disadvantages to using the median. For one thing, it is much more difficult to calculate the median than the mean. In order to find the median of a large set of data the numbers must first be ordered from smallest to largest. This can take time even when a computer is used. Another disadvantage is that medians cannot be combined. Thus, you may know that the median of a set of 50,000 numbers is 160, and the median of another set of 50,000 numbers is 190. You cannot find the combined median of the 100,000 numbers unless you order all 100,000 numbers. This is not the case, however, with the mean, as we have already indicated. If the mean of the first set of 50,000 numbers is 120 and the mean of

the second set of 50,000 numbers is 180, then the overall mean of the 100,000 numbers is simply:

$$\frac{120 + 180}{2} = 150.$$

Finally, the median tends to be less reliable for inferential purposes than the mean. This is a statement we cannot elaborate upon at this point, but we will discuss it further in subsequent chapters. This is probably the biggest advantage of the mean over the median (and other statistical measures), which makes the mean much more useful, overall, than other statistical measures of location.

3.2.3 The Mode

In the set of nine incomes we used earlier in this chapter, the number 14 (corresponding to an income of $14,000) appears three times. This is more often than any other number appears. The number occurring must often in a data set is called the *mode*.

Mode = the value that occurs with the greatest frequency in a set of data.

The best way to find the mode is to construct a frequency distribution. The mode is simply the number which has the largest frequency, either in terms of absolute frequency or relative frequency. The frequency distribution of the the nine incomes is given in Table 3.1.

Table 3.1 : Frequency Distribution of Incomes (in thousands of dollars)

Income				Absolute Frequency
12				1
14	←	Mode	←	3
16				1
17				1
18				1
20				1
865				1

The mode is not difficult to find, requiring only a frequency distribution. Moreover, in some situations the mode is of particular interest. For instance, a shoe manufacturer might want to know which size of women's shoes is sold most often. However, in many cases these advantages

are outweighed by some disadvantages. For one thing, the mode is not always in the middle of the distribution. In an examination, more students may earn a perfect score of 100 than any other single score, but we would not think of 100 as being a central, or middle, value. For another thing, the mode may not exist. In some data sets, no numbers appears more than once, in which case there is no mode. If a small class takes an exam, the students may all wind up with different scores. Alternatively, there could be two or more modes. Perhaps three students receive scores of 94, three other students receive scores of 78, and no other score occurs more than twice. Then 78 and 94 are both modes, and we say that the frequency distribution is bimodal. Finally, the mode, like the median, is less reliable for inferential purposes than is the mean, and inferential procedures involving modes are more difficult to deal with than inferential procedures involving means.

3.2.4 Comparing the Mean, Median, and Mode

We have said that the mean, median, and mode are summary measures describing the position or location of the central values of the data. A choice among these measures depends upon your objectives and upon the nature of the data. If the histogram of the frequency distribution is roughly symmetric with its high point near the middle, the three measures will be very similar. The age data, represented in histogram form in Figure 2.1, illustrate this point. The histogram is roughly symmetric, and the mean (27.84) is close to the median (28) and the mode (28).

The income data given earlier in Section 3.2 demonstrate that the three measures need not always be similar. Here the mean is 110, the median is 16, and the mode is 14. The mean is influenced by extreme values, and the income data provide an (admittedly extreme) example of this phenomenon. In other cases, the mode may deviate from the other measures or there may even be several modes, as discussed above.

One way to visualize the mean is to think of a histogram being balanced on a wedge. Suppose that we record the number of days a patient is hospitalized. To keep the example simple, we assume that this information has been recorded for only ten patients, with the following results:

$$6, 3, 9, 6, 4, 8, 1, 4, 3, 6.$$

A modified histogram for the hospitalization data is shown in Figure 3.1, with each number depicted as a weight. Because there were three patients who were in the hospital 6 days, the bar above 6 consists of three weights. Where will this histogram balance? It turns out that the balancing point (in the terminology of physics, the center of gravity) is right at the mean. Here the mean number of days in the hospital is

$$\frac{6+3+9+6+4+8+1+4+3+6}{10} = \frac{50}{10} = 5,$$

so that the histogram will balance at 5, as shown in Figure 3.1.

Why does the histogram balance at 5? To the left, we have two weights one position away from 5 (at 4), two weights two positions away (at 3), and one weight four positions away (at 1). A weight two positions away is equivalent to two weights one position away. Thus, the weights to the left of 5 are equivalent to 2(1) + 2(2) + 1(4) = 10 weights one position away, at 4. To the right of 5, we have three weights one position away (at 6), one weight three positions away (at 8), and one weight four positions away (at 9). The weights to the right of 5 are equivalent to 3(1) + 1(3) + 1(4) = 10 weights one position away, at 6. But we also said the weights on the left are equivalent to 10 weights one position away, at 4. Thus, the histogram balances at 5.

If we think of the mean as the balancing point of a histogram such as in Figure 3.1, we can realize why the mean is influenced by values which are far away from the middle of the histogram. The further away they are, the more they will weight the board on their side and shift the balancing point unless a similar value on the opposite side exerts a counteracting effect.

The median, on the other hand, is not influenced by extreme values. The weights are divided into two piles: one pile has the weights for values smaller than the median, and the other has the larger values. Each pile then is put at the same distance from the median, as can be seen in Figure 3.2. The median balances the data, but the absolute value of each data point is not important. On the right-hand side, for instance, the weight corresponding to 9 and the weights corresponding to 6 are in the same pile.

The mode, finally, ignores all points except those of the highest frequency. Weights are only given to the value with the highest frequency, and the rest of the values are unimportant (see Figure 3.3). The balancing is trivial since there is only one pile of weights.

Measures of location other than the mean, median, and mode are available. As noted earlier, the choice of a descriptive measure of location depends on one's objectives and on the nature of the data (that is, the shape of the frequency distribution). When we discuss inferential statistics, however, the mean will generally be used. From an inferential viewpoint, the mean has important advantages over other measures of location.

3.3 Measures of Dispersion

Measures of location provide information about the "typical" member of a data set, or about the center of a frequency distribution. Location is not the only interesting aspect of a set of data, however. Variation within the data set can also be important. If a country is, on the average, 400 feet above see level, it could be 400 feet above sea level everywhere (that is, perfectly flat) or it could vary a lot, with many mountains and valleys. This might make a big difference to you if you planned to tour the country by bicycle.

Suppose that on a particular day, we measure the amount of rainfall (in hundreds of millimeters) hourly in Paris and Kuala Lumpur, Malaysia. Since there are 24 hours in a day, we have 24 pieces of data for each location. The data is shown in Table 3.2.

Table 3.2 : Hourly Amount of Rainfall
(Hundreds of Millimeters)

Observation	RAIN IN PARIS	RAIN IN K-L
1	14.000	0.000
2	13.000	0.000
3	15.000	0.000
4	9.000	52.000
5	9.000	0.000
6	8.000	0.000
7	8.000	0.000
8	13.000	0.000
9	15.000	0.000
10	3.000	0.000
11	5.000	0.000
12	10.000	0.000
13	18.000	0.000
14	12.000	108.000
15	20.000	0.000
16	16.000	0.000
17	9.000	0.000
18	4.000	0.000
19	0.000	0.000
20	0.000	0.000
21	0.000	94.000
22	29.000	0.000
23	17.000	0.000
24	17.000	0.000
---	---	---
Sum	264.0000	264.0000
N =	24	24
Mean	11.0000	11.0000

From Table 3.2 it can be seen that the mean hourly amount of rainfall on this day happens to be 11 (that is, 0.11 millimeters) in both cities. Nonetheless, the hourly distribution of rain is very different. In Paris it rained fairly steadily most of the day (and night). In Kuala Lumpur it only rained during three hours, but the rain during these hours was much heavier than for any single hour in Paris. It would be wrong, then, to imply that the weather in Paris and Kuala-Lumpur was

identical because the mean amount of rainfall is the same for both cities. With the actual data or frequency distributions of the amount of rainfall this is obvious. However, saving such data involves storing a lot of information, which we would like to avoid if we could. An alternative is to compute measures of dispersion, or variation, in addition to a measure of location such as the mean. This will tell us how "spread out" our data are. We consider three measures of dispersion: the range, the variance, and the standard deviation.

3.3.1 The Range

An easy to compute and useful measure of variation is the *range*. The range is simple: the difference between the largest and the smallest data value. The range of the Paris data is

$$\text{Range}_P = \underset{\substack{\uparrow \\ \text{Largest amount} \\ \text{of rainfall}}}{29} - \underset{\substack{\uparrow \\ \text{Smallest amount} \\ \text{of rainfall}}}{0} = 29.$$

The range of the Kuala-Lumpur data is

$$\text{Range}_{K\text{-}L} = 108 - 0 = 108.$$

The fact that the range of rainfall in Kuala Lumpur is much bigger than that of Paris indicates that the Kuala Lumpur rainfall data is much more dispersed than the Paris data. The range provides information about dispersion, but its weakness is that it depends only on the most extreme data values, one at each end of the frequency distribution. It tell us nothing about whether the rest of the values tend to be close to the mean, close to the extremes, distributed fairly evenly, and so on. As a result, it is used primarily for small data sets as a quick, rough measure of dispersion.

3.3.2 The Variance

Another measure of dispersion is the *variance*. To find the variance, we find out how far each value is from the mean. If a particular value, x_i, is 12 and the mean is 10, then the deviation of x_i from the mean is 12 - 10 = 2. Values above the mean have positive deviations, while values below the mean have negative deviations.

When we are interested in measuring dispersion, we want to know how far the individual data values are from the mean. We do not, however, care on which side of the mean a value falls. Thus, the sign of the deviation from the mean is not of interest.

Since the sign is not of interest, we square the deviations to get rid of any negative signs. Then we find the average of these squared deviations and call this the variance.

$$\text{Variance} = \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^2}{N}.$$

The term $(x_i - \text{Mean})^2$ is the square of the deviation of the ith piece of data, x_i, from the mean. As in the calculation of the mean, Σ denotes summation, and we add up all of the squared deviations, from the first, $(x_i - \text{Mean})^2$, to the last, $(x_N - \text{Mean})^2$.

Table 3.3 : Computation of the Variance for the Hospitalization Data.

Observation	X	MEAN	X-MEAN	(X-MEAN)^2
1	6.000	5.000	1.000	1.000
2	3.000	5.000	-2.000	4.000
3	9.000	5.000	4.000	16.000
4	6.000	5.000	1.000	1.000
5	4.000	5.000	-1.000	1.000
6	8.000	5.000	3.000	9.000
7	1.000	5.000	-4.000	16.000
8	4.000	5.000	-1.000	1.000
9	3.000	5.000	-2.000	4.000
10	6.000	5.000	1.000	1.000
Sum	50.0000	50.0000	0.0000	54.0000
N =	10	10	10	10
Mean	5.0000	5.0000	0.0000	5.4000
				↑
				Variance

The hospitalization data of Section 3.2 are shown in the second column of Table 3.3. Here x represents the number of days of hospitalization. The next column just lists the mean. Column four is the difference between our data (the values of x) and their mean of 5. Finally, column five shows the square of each of the differences in column four. The variance is the sum of all these squared differences divided by 10, the number of observations. In symbols,

$$\text{Variance} = \frac{\sum_{i=1}^{N} (x_i - \text{Mean})^2}{N} = \frac{54}{10} = 5.4 .$$

Notice that in applying the variance formula, there are ten terms (that is, the summation goes from 1 to 10). These terms are precisely the ten numbers of column five (the last column) of Table 3.3. Computing the variance is straightforward, and it can be accomplished routinely by a computer program or on a hand calculator.

Figure 3.4 shows a plot of the data and their mean. The vertical lines are the differences between each data point and the mean (that is, the equivalent of column four of Table 3.4). They show the extent to which each data point varies from the mean of 5. The variance is the average of the squares of all of these differences. This is an alternative way of visualizing the variance and what it actually means.

3.3.3 The Standard Deviation

The variance is the average of the *squared* differences between the data values and the mean. The important point here is that they have been *squared*, which means that the variance is expressed in squared units. For the hospitalization data, the variance is in "days squared."

To avoid squared units the square root of the variance can be taken. the result is called the standard deviation.

The formula for the standard deviation is simply the square root of the formula for the variance.

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \text{Mean})^2}{N}} .$$

The square root of the variance of the hospitalization data is simply $\sqrt{5.4} = 2.324$. This is in terms of days, not "days squared."

The variance and the standard deviation provide us with equivalent information. The only difference between the two is that they are expressed in different units of measurement. Apart from that, their interpretation is similar. The larger the standard deviation is the more dispersed the data are around their mean; the smaller it is the less spread out the data are.

3.4 Populations and Samples

So far we have worked with some data sets without making a distinction as to whether these data refer to a population or a sample.

A *population* contains *all* possible units (observations) of interest. For instance, for a candidate running for election, the "population" consists of all persons of voting age in his or her district; for the director of a school, the "population" is all the students in the school. The words "of interest" are important here because in the same school the population for a teacher may be only those students in that teacher's classes; for an auditor the "population" may be all accounts payable and accounts receivable in a company's books; for the marketing manager of a pharmaceutical company wishing to introduce a new drug the "population" might be all present and future patients with the ailment the drug is supposed to cure; and so on.

There are several aspects that need further explanation. First of all, the units of a population do not need to be people or physical objects. For example, we might be interested in the population of the daily high temperatures at Los Angeles during a particular ten-year period. Or we might be interested in the per-family income in a given country. A member of the "population" is then the yearly amount of money that a particular family in this country earns.

It is also important to understand that in order to define a population we usually have an objective in mind. What is a population in one case might not be in another. Thus, for the baby-food department of a big company, the "population" is, say, all children 18 months of age or younger. For the cosmetics division, however, the "population" may be all women at least 15 years of age.

A *sample* is a set of data which is only a part of the population. A pre-election poll is generally a sample of the population of all registered voters. The set of accounts payable on the 36th page of an accounting book is a sample of all accounts payable. The women, older than 15 years, who walk between 56th and 57th street on 5th Avenue in New York is a sample of all possible women.

In statistics the words "population" and "sample" are of extreme interest. As a matter of fact, the major objective of statistics is to be able to generalize (in statistical terminology, to make inferences) from a sample to the entire population. Thus, when the Harris or Gallup organization takes a sample of around 1,600 people of voting age, the purpose is to be able to make an inference about who will be elected in the next election from the information obtained from the sample. Similarly, when a food company tests the potential market for a new dessert with a few hundred families, its purpose is to infer how all potential buyers will like and consequently buy the new product.

The benefits of being able to make inferences from a sample to the entire population are enormous. Collecting data from a sample instead of from the entire population is not only a matter of less cost and work but also a matter of speed. The information from a sample usually can be collected and processed easily. To do the same for the entire population can be very costly and very time-consuming. Furthermore, the chances of making mistakes in counting increase when large populations are involved.

In statistics it is customary to distinguish populations from samples. The size of a population is denoted by N, while that of a sample is denoted by n. Also, summary values computed from a population are usually denoted by Greek letters. For example, μ is used for the population mean, σ^2 for the population variance, and σ for the population standard deviation. Measures that describe some aspects of a population are called *parameters*.

Parameter: A summary measure of a population.

In contrast, values computed from a sample are generally denoted by Roman letters. We use $\bar{x}$ for a sample mean, s^2 for a sample variance, and s for a sample standard deviation. Measures that describe some aspects of a sample are called *statistics*.

Statistic: A summary measure of a sample.

In inferential statistics, we are often concerned with a particular population parameter. In making inferences, we may calculate one or more statistics from the sample. For instance, suppose that we are interested in μ, the mean life span of all light bulbs manufactured by Company XYZ. It is unfeasible to test every light bulb in the population, but we can take a sample of light bulbs and test them to see how long they last. The average life span in the sample, which is the sample mean $\bar{x}$, can be computed. We might use this sample mean $\bar{x}$ as an estimate of the population μ. Moreover, as you shall see later in the book, methods are available to evaluate the accuracy of our estimate.

If we just want to describe a data set, it doesn't make much difference whether the data represent a population or a sample. In inferential statistics, however, this distinction is important. Table 3.4 summarizes various differences in notation and terminology between populations and samples and gives several examples of each.

Table 3.4 : Populations and Samples

POPULATION	SAMPLE
1. Definition	
All conceivable possible units of interest.	*A part* of the population.
2. Symbols used	
μ = mean	$\bar{x}$ = mean
σ^2 = variance	s^2 = variance
σ = standard deviation	s = standard deviation
N = total number of units in the population	n = total number of units in the sample
μ, σ^2, and σ are called parameters.	$\bar{x}$, s^2, and s are called statistics.
3. Examples	
The number of people in the U.S. found by the population census taken by the Bureau of the Census once every ten years.	The number of families whose income is above $30,000 in a sample of 1,000 families.
All those who voted in an election.	A sample (opinion poll) of 1,600 registered voters taken before the election
The total number of unemployed that have been registered with the unemployment insurance office.	The number of unemployed in a panel of 1,000 families used to monitor unemployment.
All potential buyers of a new product.	The buyers of the new product in test marketing.
The life span of all light bulbs manufactured by Company XYZ.	The life span of 300 light bulbs tested to see how long they would last.

3.5 Formulas for the Mean, Variance, and Standard Deviation

So far we have presented formulas for computing the population mean, the variance, and the standard deviation. These are:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}, \tag{3.1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}, \tag{3.2}$$

$$\text{and } \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}. \tag{3.3}$$

Unfortunately, (3.1), (3.2) and (3.3) are not the only formulas for the calculation of means, variances, and standard deviations. There are other formulas which we need to understand and be able to use under different circumstances.

Suppose we want to calculate the mean of the data shown in Table 2.1. The mean is (see Section 3.2.1):

$$\mu = \frac{\sum_{i=1}^{230} x_i}{230} = \frac{6403}{230} = 27.84.$$

Now suppose that the data of Table 2.1 are not available, but the frequency distribution shown in Table 2.5 is known instead. The mean cannot be found by using expression (3.1), which is reserved for raw data, not frequency distributions.

In order to find the mean of a frequency distribution, this alternative formula should be used:

$$\mu = \Sigma f_i x_i, \tag{3.4}$$

where x_i is the middle point of the ith class interval and f_i is the relative frequency of the ith class interval (that is, column four in Table 2.5).

The rationale behind formula (3.4) is that some values (hence some class intervals) occur more frequently than others. They should, therefore, be given more weight. Thus, in the first class interval there are only two students; in the second there are 10, in the third 29, and so on. The relative frequencies of these classes are 2/230 = 0.0087, 10/230 = 0.0435, 29/230 = 0.1261, and so on. These relative frequencies, which are shown in the fourth column of Table 2.5, are what we denote by f_i in formula (3.4).

Table 3.5 shows the computations involved in finding the mean. First the product of each x_i and f_i is found (see the last column of Table 3.5), and then their sum is calculated. This sum (see the bottom of column four) of 27.8467 is the mean. Thus, the mean is found as follows:

$$\mu = \Sigma f_i x_i = 20.5(.0087) + 22.5(.0435) + 24.5(.1261) + \ldots 38.5(.0043) = 27.8467.$$

In order to see clearly the relationship between formulas (3.4) and (3.5), imagine that each class interval has only one observation (e.g., student). Then, f_i will be a constant always equal to 1. In this case, formula (3.4) can be rewritten as

$$\mu = \Sigma f x_i, \qquad (3.5)$$

where f is a constant (this is why it is not written as f_i).

Now if there are, say 230 observations and each interval contains only one, the value of f will be 1/230; in general, it will be 1/N. If this 1/N is substituted in (3.5) we get

$$\mu = \Sigma \frac{1}{N} x_i,$$

or

$$\mu = \frac{\Sigma x_i}{N},$$

which is precisely (3.1). Therefore, (3.1) is a special case of (3.4) which is used when each f_i is a constant equal to 1/N.

Table 3.5 : The Mean of the Frequency Distribution of Ages (Table 2.5)

x	f	fx
20.5	0.0087	0.1784
22.5	0.0435	0.9788
24.5	0.1261	3.0895
26.5	0.2957	7.8361
28.5	0.2870	8.1795
30.5	0.1522	4.6421
32.5	0.0478	1.5535
34.5	0.0217	0.7487
36.5	0.0130	0.4745
38.5	0.0043	0.1656
		27.8467 ← $\Sigma f_i x_i$

Table 3.6 : Calculation of the Variance of the Frequency Distribution of Ages (Table 2.5)

x	Mean	x-Mean	(x-Mean)^2	f	f(x-Mean)^2
20.5	27.8467	-7.3467	53.9740	0.0087	0.4696
22.5	27.8467	-5.3467	28.5874	0.0435	1.2435
24.5	27.8467	-3.3467	11.2004	0.1261	1.4124
26.5	27.8467	-1.3467	1.8136	0.2957	0.5363
28.5	27.8467	0.6533	0.4268	0.2870	0.1225
30.5	27.8467	2.6533	7.0400	0.1522	1.0715
32.5	27.8467	4.6533	21.6532	0.0478	1.0350
34.5	27.8467	6.6533	44.2664	0.0217	0.9606
36.5	27.8467	8.6533	74.8796	0.0130	0.9734
38.5	27.8467	10.6533	113.4928	0.0043	0.4880
					8.3128 ↑ Variance

Note that there is a very small difference in the means calculated by formulas (3.4) and (3.1). In the first case, the mean is

$$\mu = \Sigma f_i x_i = 27.8467$$

In the second case, the mean is

$$\mu = \frac{\Sigma x_i}{N} = 27.84.$$

Why this difference? The reason is that in constructing a frequency distribution we have lost some information by grouping observations together in intervals. The difference in means is usually small, however, and in this case it is incredibly small. The correct value of the mean is 27.84, and the value of 27.8467 is a little different because the data were grouped first and then their mean was computed.

The formula for the variance from a frequency distribution is

$$\sigma^2 = \Sigma f_i (x_i - \mu)^2 \qquad (3.6)$$

Formula (3.6) is equivalent to (3.2). If each frequency is one, then $f_i = f = 1/N$, and (3.6) becomes the same as (3.2).

In order to use (3.6) we need to know the mean, then find $x_i - \mu$ for each class interval, square this difference, and multiply $(x_i - \mu)^2$ by the relative frequency f_i. This is done in Table 3.6. The sum of the last column of this table is the variance, which is equal to 8.3128.

The standard deviation is simply the square root of the variance. Thus, for a frequency distribution,

$$\sigma = \sqrt{\Sigma f_i (x_i - \mu)^2}. \qquad (3.7)$$

For the frequency distribution of ages, $\sigma^2 = 8.3128$, and

$$\sigma = \sqrt{8.3128} = 2.8832.$$

In order to use the formulas (3.2) or (3.6) for the variance, several steps are required. The mean of the data must be found, each difference $(x_i - \mu)$ needs to be found and squared, and so on. Equivalent formulas for the variance, formulas which are a bit easier to use from a computational viewpoint, are available:

$$\sigma^2 = \frac{N\Sigma x_i^2 - (\Sigma x_i)^2}{N^2} \qquad (3.8)$$

and $$\sigma^2 = \sum f_i x_i^2 - \mu^2. \qquad (3.9)$$

Formula (3.8) is the equivalent of (3.2) while (3.9) is the equivalent of (3.6). In order to arrive at (3.8) and (3.9) we only need to make some algebraic manipulations on (3.2) or (3.6).

Finally, there is a slight difference in the formulas for the variance of a sample and the variance of a population. For the population variance, we divide the sum of squared deviations by N, the population size. In the case of the sample, instead of dividing by the sample size, n, we divide by n-1. The formula for the sample variance is therefore

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}.$$

*(3.10)

The reason for dividing by n-1 instead of n relates to inferential statistics. The sample variance s^2 is used to estimate the population variance σ^2, as you shall see later in the book. If we divided by n in computing s^2, then our sample variance would tend, on average, to be less than the population variance. Dividing by n-1 instead of n exactly corrects this problem.

You might feel a little confused at this point because of all of the formulas that have been presented. Unfortunately, statistics is a technical subject which makes some difficulties of this kind unavoidable.

Table 3.7 : Formulas for the Mean and Variance

POPULATION	SAMPLE
1. Raw Data : Constant Frequency of 1/N	
(a) Definitional Formulas	
$\mu = \frac{\sum x_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$
$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$
$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$
(b) Computational Formulas	
μ : same	$\bar{x}$: same
$\sigma^2 = \frac{N\sum x_i^2 - (\sum x_i)^2}{N^2}$	$s^2 = \frac{n\sum x_i^2 - (\sum x_i)^2}{n(n - 1)}$
$\sigma = \sqrt{\frac{N\sum x_i^2 - (\sum x_i)^2}{N^2}}$	$s = \sqrt{\frac{n\sum x_i^2 - (\sum x_i)^2}{n(n - 1)}}$
2. Grouped Data : Unequal Frequencies	
(a) Definitional Formulas	
$\mu = \sum f_i x_i$	$\bar{x} = \sum f_i x_i$
$\sigma^2 = \sum f_i (x_i - \mu)^2$	$s^2 = \left(\frac{n}{n - 1}\right) \sum f_i (x_i - \bar{x})^2$
$\sigma = \sqrt{\sum f_i (x_i - \mu)^2}$	$s = \sqrt{\frac{n}{n - 1} \sum f_i - (x_i - \bar{x})^2}$
(b) Computational Formulas	
μ : same	$\bar{x}$: same
$\sigma^2 = \sum f_i x_i^2 - \mu^2$	$s^2 = \frac{n}{n - 1}\left[\sum f_i x_i^2 - \bar{x}^2\right]$
$\sigma = \sqrt{\sum f_i x_i^2 - \mu^2}$	$s = \sqrt{\left(\frac{n}{n - 1}\right)\left[\sum f_i x_i^2 - \bar{x}^2\right]}$

You should try to understand the rationale behind each summary measure. Table 3.7 can help you keep track of the various formulas. The definitional formulas in Table 3.7 come directly from our definitions of the mean, variance, and standard deviation. The computational formulas are exactly what the name implies - they are given because they are somewhat easier to use from a computational standpoint than are the definitional formulas.

3.6 Chebyshev's Theorem

So far we have said that the variance or standard deviation tells us how spread out our data are. Furthermore, we said that the larger the variance and standard deviation are, the greater the dispersion. This by itself is important information but it is only part of the whole story. Obviously we would like to be more precise than simply saying that a larger variance or standard deviation means more spread out data. A mean has an intuitively appealing interpretation as an average value, but the variance and standard deviation are not so easy to interpret. A result known as Chebyshev's Theorem helps somewhat by giving us some idea as to what percentage of the data might be within one, two, or any other number of standard deviations from the mean.

Chebyshev's Theorem: For any set of data (population or sample) or any distribution (frequency or otherwise), the proportion of data which lies within k standard deviations of the mean is *at least*

$$1 - \frac{1}{k^2}.$$

Thus we can tell at least what percentage of the data are in the regions $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 2.35\sigma$, $\mu \pm 3.19\sigma$, and so forth. The "$\pm$" symbol is interpreted as follows: the region $\mu \pm \sigma$ is the interval from $\mu - \sigma$ to $\mu + \sigma$.

As an example of how Chebyshev's Theorem works, let us use the data on ages in Table 2.1. We have already found the mean to be 27.84. Table 3.8 shows how the standard deviation can be calculated by the computational formula (3.8). Its value is 2.843. Notice that this value is a little different than 2.8832 we found by using formula (3.7), but the difference is small and is - as in the case of the mean - due to losing some information while grouping the data into class intervals of two years.

Table 3.8 : Computations for the Standard Deviation of the Population of Ages

Observation	x	x^2
1	27.000	729.000
2	25.000	625.000
3	25.000	625.000
4	32.000	1024.000
5	30.000	900.000
6	29.000	841.000
7	26.000	676.000
8	28.000	784.000
.	.	.
.	.	.
.	.	.
223	34.000	1156.000
224	34.000	1156.000
225	29.000	841.000
226	29.000	841.000
227	25.000	625.000
228	37.000	1369.000
229	34.000	1156.000
230	29.000	841.000
Sum	6403.0000	180113.0000
N =	230	230
Mean	27.8391	

$$\sigma = \sqrt{\frac{N\sum x_i^2 - (\sum x_i)^2}{N^2}}$$

$$= \sqrt{\frac{230(180113) - (6403)^2}{230^2}} = 2.843$$

We can now use Chebyshev's Theorem since we have the mean and standard deviation. Table 3.9 shows the results for some different values of k, and the Chebyshev proportion is graphed as a function of k in Figure 3.5. Note, first of all, that the theorem does not work for values of k less than 1. The resulting proportions are negative, which makes no sense. When k = 1 the value of $1-(1/k^2)$ is zero. The result is correct since it means that at least zero percent of the data lies between $\mu-\sigma$ and $\mu+\sigma$, but it is also trivial. We do not need Chebyshev to tell us that.

Chevyshev's Theorem becomes useful when k is greater than one. For example, it tells us that at least .75, or 75% of the ages are between 22.16 and 33.52.

Table 3.9 : Applying Chebyshev's Theorem when $\mu = 27.84$ and $\sigma = 2.84$

Value of k	$\mu \pm k\sigma$	Value of $\mu \pm k\sigma$	Interval $\mu - k\sigma$ $\mu + k\sigma$	$1 - 1/k^2$	Value of $1 - 1/k^2$
0.8	$\mu \pm .8\sigma$	27.84 ± 2.27	25.57 - 30.11	$1 - 1/.8^2$	- .56*
1.0	$\mu \pm \sigma$	27.84 ± 2.84	25.00 - 30.68	$1 - 1/1^2$	0**
1.5	$\mu \pm 1.5\sigma$	27.84 ± 4.26	23.58 - 32.10	$1 - 1/1.5^2$	.56
2.0	$\mu \pm 2\sigma$	27.84 ± 5.68	22.16 - 33.52	$1 - 1/2^2$	.75
2.3	$\mu \pm 2.3\sigma$	27.84 ± 6.53	21.31 - 34.37	$1 - 1/2.3^2$	.81
2.7	$\mu \pm 2.7\sigma$	27.84 ± 7.67	20.00 - 35.68	$1 - 1/2.7^2$	.86
3.0	$\mu \pm 3\sigma$	27.84 ± 8.52	19.32 - 36.36	$1 - 1/3^2$	.89
3.75	$\mu \pm 3.75\sigma$	27.84 ± 10.65	17.19 - 38.49	$1 - 1/3.75^2$	.93
4.0	$\mu \pm 4\sigma$	27.84 ± 11.36	16.48 - 39.20	$1 - 1/4^2$	.94
5.0	$\mu \pm 5\sigma$	27.84 ± 14.2	13.64 - 42.04	$1 - 1/5^2$	.96
6.0	$\mu \pm 6\sigma$	27.84 ± 17.04	10.80 - 44.88	$1 - 1/6^2$	.97
7.0	$\mu \pm 7\sigma$	27.84 ± 19.88	7.96 - 47.72	$1 - 1/7^2$	.98
8.0	$\mu \pm 8\sigma$	27.84 ± 22.72	5.12 - 50.56	$1 - 1/8^2$	.984
9.0	$\mu \pm 9\sigma$	27.84 ± 25.56	2.28 - 53.40	$1 - 1/9^2$	.989
10.0	$\mu \pm 10\sigma$	27.84 ± 28.4	- .56 - 56.24	$1 - 1/10^2$	.99

* Cannot be defined for values of $k < 1$.

** Not useful. It is only useful for values of $k > 1$.

A problem which limits the practical value of Chebyshev's Theorem is its generality. Because it is very general and aims at all possible cases, it sometimes might not be powerful enough. For instance, we are told by Chebyshev's Theorem that the interval $\mu \pm 3\sigma$ contains *at least* 89% of the students' ages. If we look at Table 2.5, however, we see that the interval $\mu \pm 3\sigma$ (or from 19.32 to 36.36) contains more than 98% of all ages. That is, we can be more precise by looking at Table 2.5 rather than using Chebyshev's Theorem. On the other hand, we get a great deal of information from this theorem by just knowing two numbers: the mean and standard deviation.

If we do not know exactly how many pieces of data are within an interval such as $\mu \pm 2\sigma$, but we have some idea as to the shape of the histogram, we may be able to do a little better than

Chebyshev's Theorem by using an approximation. If the histogram is roughly bell-shaped (that is, if it is similar to the curve given in Figure 3.6), then we can get a rough idea of how close the data are to the mean. This approximation is often called the Empirical Rule.

Empirical Rule: If a histogram is roughly bell-shaped, then
- approximately 68% of the data are within one standard deviation of the mean;
- approximately 95% of the data are within two standard deviations of the mean;
- all or almost all of the data are within three standard deviations of the mean.

For example, the histogram of ages in Figure 2.1 is roughly bell-shaped. For the age data, $\mu = 27.84$ and $\sigma = 2.88$. Thus, $\mu \pm \sigma$ is the interval from 27.84 - 2.88 = 24.96 to 27.84 + 2.88 = 30.72. Exactly 173, or 75%, of the 230 ages are in this interval (see Table 2.2). The interval $\mu \pm 2\sigma$, which goes from 22.08 to 33.60, includes 216, or 94% of the ages. The interval $\mu \pm 3\sigma$, from 19.20 to 36.48, includes 228, or over 99% of the ages. These figures (75% for $\mu \pm \sigma$, 94% for $\mu \pm 2\sigma$, and 99% for $\mu \pm 3\sigma$) are close to the values given in the Empirical Rule.

The Empirical Rule is based on a theoretical distribution called the normal distribution, and the bell-shaped curve in Figure 3.6 is called a normal curve. We will discuss the normal distribution in Chapter 5 and use it extensively later in the book. If a set of data follows, approximately, a theoretical curve such as the normal curve, we can make statements like those given in the Empirical Rule. These statements are more precise than statements based on Chebyshev's Theorem. The Empirical Rule does not apply to all data sets, but when it does apply, it provides more accurate information than does Chebyshev's Theorem.

Figure 3.1 : Balancing a Board with the Ten Numbers as weights - The Case of the Mean.

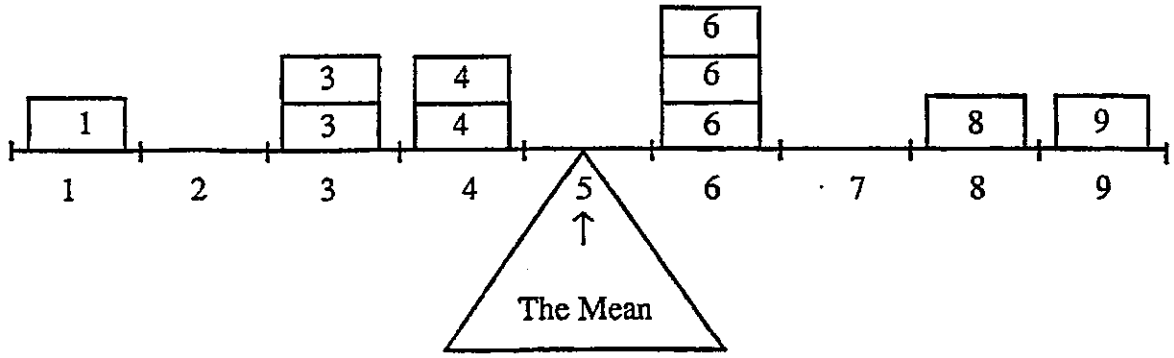

Figure 3.2 : Balancing a Board with the Ten Numbers - The Case of the Median.

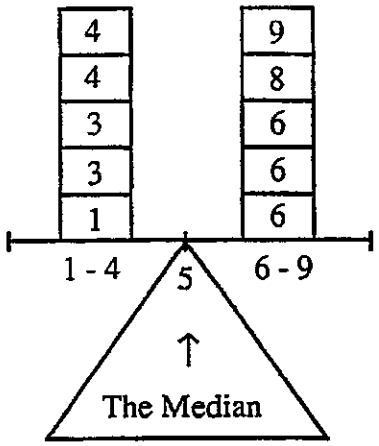

Figure 3.3 : Balancing a Board with the Highest Frequency Numbers only, Ignoring the Rest - The Case of the Mode.

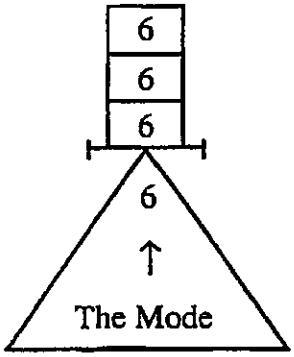

Figure 3.4 : Variation of each of the Ten Data Points from their Mean of 5.

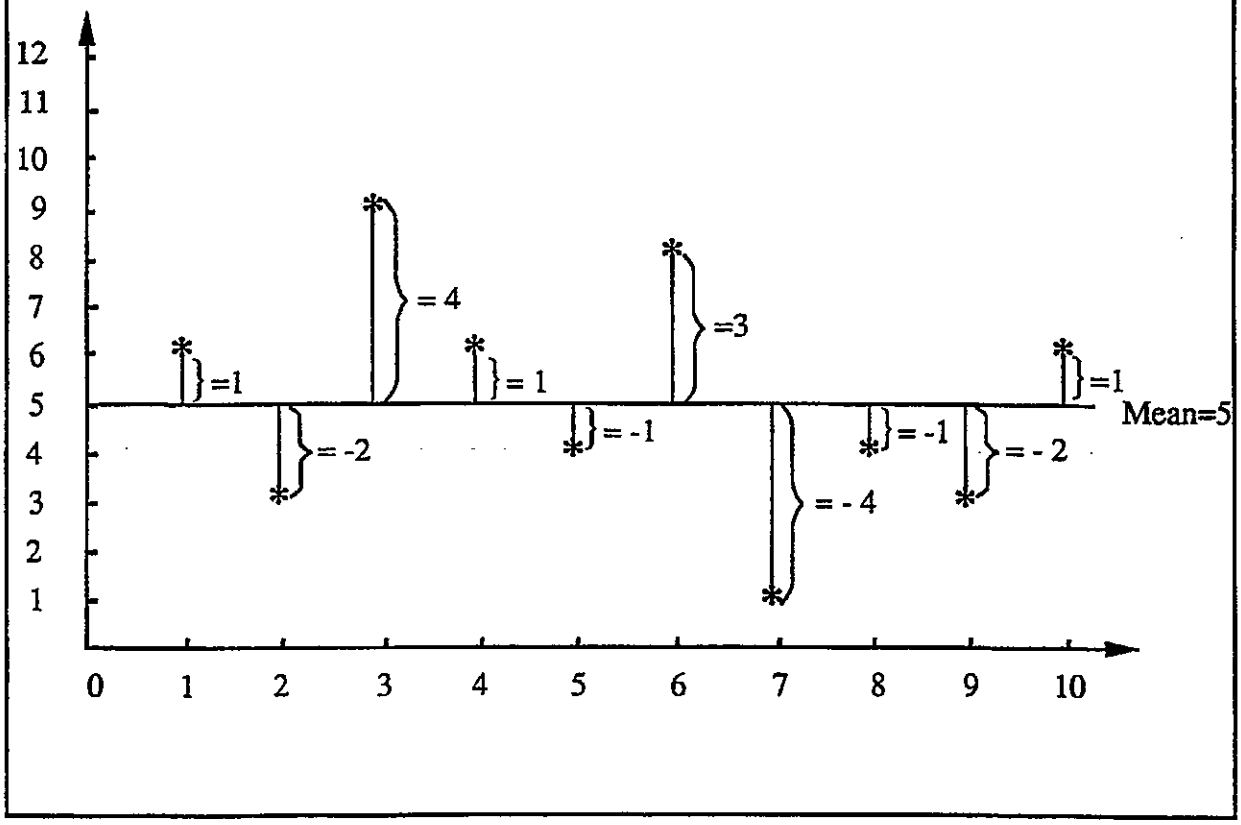

Figure 3.5 : The Chebyshev Proportion as a Function of k.

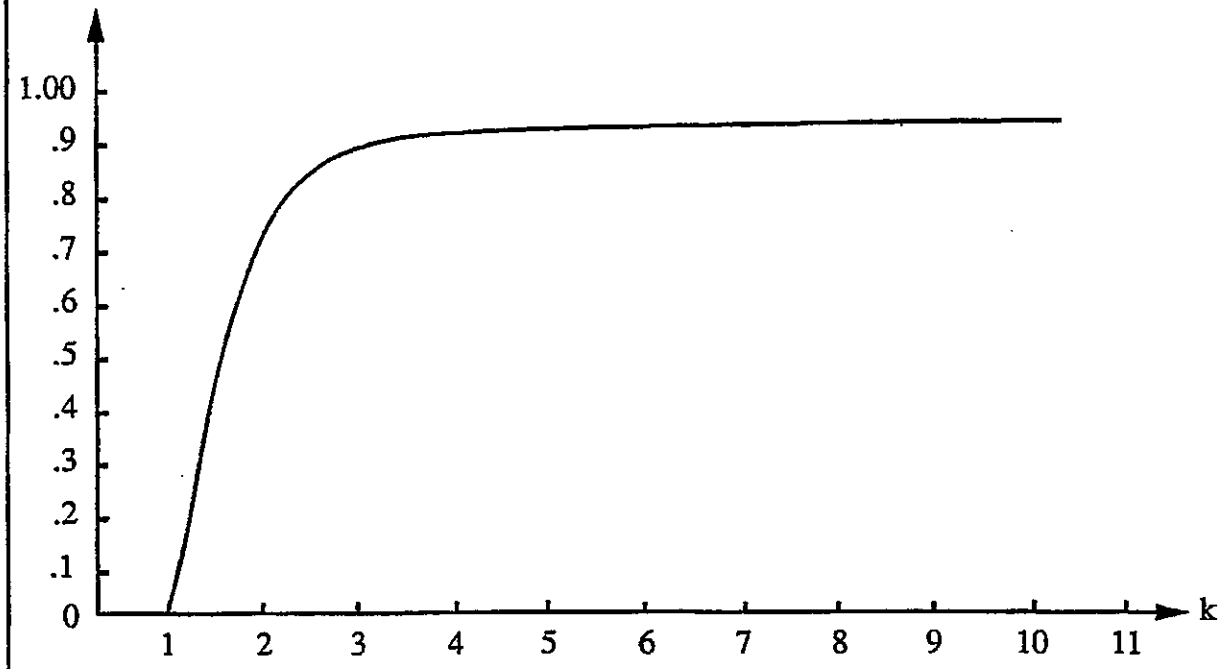

Figure 3.6 : A Bell-shaped Curve.