2 NUMBERS WORTH KNOWING

Quantitative information forms the core of what businesses must know to operate effectively. The current emphasis on business metrics, Key Performance Indicators (KPIs), and Balanced Scorecards demonstrates the importance of numbers in business. The messages contained in numbers are communicated most effectively when you understand the fundamental characteristics and meaning of the numbers that are commonly used in business, as well as the fundamental principles of effective communication that apply specifically to quantitative information.

- **Quantitative relationships**
  - **Relationships within categories**
  - **Relationships between quantities**
- **Numbers that summarize**
  - **Measures of average**
  - **Measures of distribution**
  - **Measures of correlation**
  - **Measures of ratio**
- **Measures of money**

Numbers are not intrinsically boring. Neither are they intrinsically interesting. The fact that they are quantitative has no bearing on their inherent appeal. They simply belong to the class of information that communicates the quantity of something. The impact and appeal of information, quantitative or not, flows naturally from the significance and relevance of the message it contains. As a communicator, it is up to you to give a clear and unobstructed voice to that information and its message, using language that is easily understood by your audience.

You may be anxious to jump right into the design of tables and graphs. After all, that's the fun stuff. I must admit, I was tempted to get right to it, but because numbers are the content and thus the substance of tables and graphs, it's important to begin our journey by getting acquainted with a few numbers that are particularly useful in quantitative communication.

Quantitative Relationships

When you design the display of quantitative information, whether you use a table or graph, the specific type of table or graph you use depends primarily on your message. What about the message? Quantitative messages are always about *relationships*. Numbers, in and of themselves, are of no use unless they measure something that is important to you. Here are some common examples of relationships that define the essential nature of quantitative messages:

Quantitative Information	*Relationship*
Units of a product sold per geographical region	Sales related to geography
Revenue by quarter	Revenue related to time
Expenses by department and month	Expenses related to organizational structure and time
A company's market share compared to that of its competitors	Market share related to companies
The number of employees who received each of the five possible performance ratings (1–5) during the last annual performance review	Employee counts related to performance ratings

In each of these examples, there is a simple relationship between some measure of quantity and one or more associated categories of interest to the business (geography, time, etc.). Quantitative information consists of two types of data: *quantitative* and *categorical*. Quantitative values measure things; categories subdivide the things that they measure into useful groups, such as geographical areas (e.g., north, east, south, and west) in the category called sales regions, or individual months in the category called time. This distinction between quantitative and categorical data is fundamental to tables and graphs. These two types of data play different roles in tables and graphs and are often structured and displayed in distinct ways.

Quantitative values are also expressed in units of measure. For instance, the quantitative value *$200* is made up of the quantity—200—and a relevant unit of measure—dollars.

Sometimes the relationships we display are simple associations between quantitative values and categorical subdivisions, such as those in the previous examples. Sometimes the relationships display direct associations between multiple sets of quantitative values, such as in the examples below:

Quantitative Information	*Relationship*
The effect of a mass-mailing marketing campaign on order volume	The number of letters sent related to the number of orders received
Units sold and the resulting revenue in correlation to pricing	Product price related to the associated number of units sold

This distinction between simple relationships that associate quantitative values and categorical subdivisions and somewhat more complex relationships that associate multiple sets of quantitative values is also fundamental to our use of tables and graphs. Different types of relationships require different types of displays.

So far we've only examined a few examples of quantitative relationships, but the list is endless. Think for a minute or two about the quantitative information that is communicated at your place of business. Can you think of any that doesn't involve relationships?

Thus far we've learned the following about quantitative information:

- Quantitative information consists of two types of data:
  - Quantitative
  - Categorical
- Quantitative information always describes relationships.
- These relationships involve either
  - Simple associations between quantitative values and categorical subdivisions or
  - More complex associations among multiple sets of quantitative values.

In addition to the two fundamental types of quantitative relationships that we've already noted, there are also a variety of ways in which categorical subdivisions or the quantitative values associated with them can relate to one another. Let's take a look at these ways.

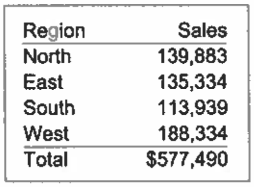
Relationships Within Categories

Categorical subdivisions can relate to one another in the following ways:

- Nominal
- Ordinal
- Interval
- Hierarchical

NOMINAL

A *nominal* relationship is one in which the individual subdivisions of a single category are discrete and have no intrinsic order. For instance, the four sales regions *East*, *West*, *North*, and *South*, in and of themselves, are not related in any particular order. These labels simply name the different sales regions, thus the term nominal, which means "in name only." Here's a simple example:



Region	Sales
North	139,883
East	135,334
South	113,939
West	188,334
Total	$577,490

FIGURE 2.1 This is an example of a nominal relationship.

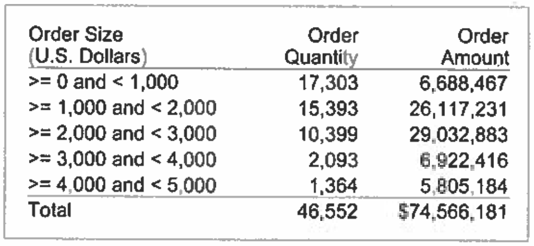
When you communicate a quantitative message that is nominal in nature, you simply divide up the quantitative values in association with separate categorical subdivisions, each bearing a different name, but your message does not relate those subdivisions to one another in any particular way.

ORDINAL

An *ordinal* relationship between categorical subdivisions is one in which the individual subdivisions have a prescribed *order*. Typical examples include "first, second, third . . ." and "small, medium, and large." To display them in any other order would rarely be meaningful.

INTERVAL

An *interval* relationship is one in which the categorical subdivisions consist of a series of individual, sequential numerical ranges that subdivide a full set of quantitative values into smaller ranges. These individual numerical ranges, called intervals, can be arranged in order from smallest to largest (ascending order) or largest to smallest (descending order). Interval relationships are used when you wish to look at how something is distributed across a broad range of quantitative values by subdividing the range into a set of smaller, more manageable ranges. Here's a common example:

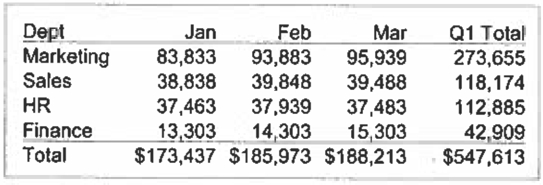


Order Size (U.S. Dollars)	Order Quantity	Order Amount
>= 0 and < 1,000	17,303	6,688,467
>= 1,000 and < 2,000	15,393	26,117,231
>= 2,000 and < 3,000	10,399	29,032,883
>= 3,000 and < 4,000	2,093	6,922,416
>= 4,000 and < 5,000	1,364	5,805,184
Total	46,552	$74,566,181

FIGURE 2.2 This is an example of an interval relationship. Notice that the intervals are equal in size. This is especially important when you intend to graphically display the distribution of a set of values across a range, called a *frequency distribution.*

In this example, to see how the orders were distributed across the entire range of order sizes, it wouldn't make sense to count the number of orders and sum their totals for each individual order amount, because that would involve an unmanageably large set of order sizes. The solution involves subdividing the full range of order sizes into a series of contiguous ranges.

Take a moment to test what you've learned so far. Look at the example below and determine which of the three relationships—nominal, ordinal, or interval—best describes its categorical subdivisions of time (months in this case).



Dept	Jan	Feb	Mar	Q1 Total
Marketing	83,833	93,883	95,939	273,655
Sales	38,838	39,848	39,488	118,174
HR	37,463	37,939	37,483	112,885
Finance	13,303	14,303	15,303	42,909
Total	$173,437	$185,973	$188,213	$547,613

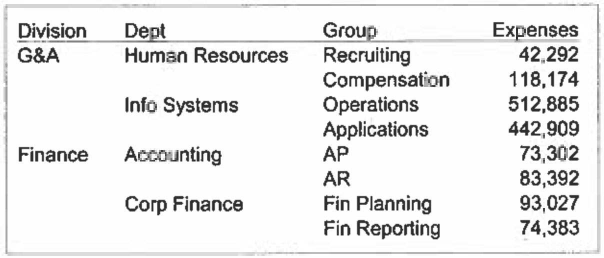
FIGURE 2.3 This is an example of time-series relationship.

Your initial inclination was probably to conclude that categorical subdivisions of time are ordinal, for they certainly make sense only when arranged chronologically. This begs the further question, however, "Do these subdivisions of time represent intervals along a quantitative scale?" The answer is "Yes, they do." Time is a quantitative scale that measures duration. Even though different months do not all represent the same exact number of days and are therefore not precisely equal intervals, for reporting purposes we treat them as equal.

So far the categorical relationships that we've examined involve relationships between members of the same categorical set. The remaining relationship discussed below does not.

HIERARCHICAL

A *hierarchical* relationship involves multiple categories that are closely related to each other as separate levels in a ranked arrangement. Starting from the top of the hierarchy and progressing down, each subdivision at each level is associated with only one subdivision at the level above it. Each subdivision at every level, except the bottom level, can have one or more subdivisions associated with it in the next level down. This is much easier to show than to describe with words. Here's a typical example viewed from left to right:



Division	Dept	Group	Expenses
G&A	Human Resources	Recruiting	42,292
		Compensation	118,174
	Info Systems	Operations	512,885
		Applications	442,909
Finance	Accounting	AP	73,302
		AR	83,392
	Corp Finance	Fin Planning	93,027
		Fin Reporting	74,383

FIGURE 2.4 This is an example of a hierarchical relationship. The *G&A* division is composed of two departments: *Human Resources* and *Info Systems*. The *Recruiting* and *Compensation* groups belong to the *Human Resources* department, and the *Operations* and *Applications* groups belong to the *Info Systems* department.

Hierarchical relationships between categories are commonly used in tables, and, to a lesser degree, in graphs, to organize quantitative information.

Relationships Between Quantities

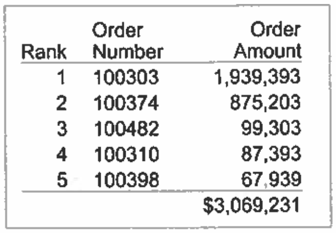
Categorical subdivisions can also relate to one another by virtue of the quantitative values associated with them. The quantitative values can be used to display the following relationships:

- Ranking
- Ratio
- Correlation

RANKING

When the order in which the categorical subdivisions are displayed is based on the associated quantitative values, either in ascending order or descending order, the relationship is called a *ranking*. If you need to construct a list of your company's top five sales orders for the current quarter based on revenue, the message would be enhanced if you arranged them by size, in this case from the largest to the smallest of the five, as you see in the following figure:

Technically, the term *ordinal* could be used to describe a ranking relationship as well, but I'm using distinct terms to highlight the difference between a sequence based on categorical subdivisions and one based on quantitative values.

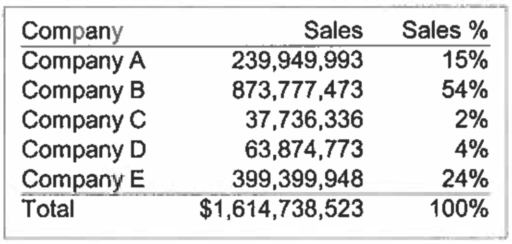


Rank	Order Number	Order Amount
1	100303	1,939,393
2	100374	875,203
3	100482	99,303
4	100310	87,393
5	100398	67,939
		$3,069,231

FIGURE 2.5 This is an example of a ranking relationship.

RATIO

A *ratio* is a relationship in which two quantitative values are compared by dividing one by the other. This produces a number that expresses their relative quantities. A common example is the relationship of the quantitative value for a single categorical subdivision compared to the sum of the entire set of subdivisions in the category (e.g., the sales of one region compared to the total sales of all regions). The ratio of a part to its whole is generally expressed as a percentage where the whole equals 100%, and the part equals some lesser percentage. Here's an example of a part-to-whole ratio in tabular form, which displays market share information for five companies, both in actual dollar sales and in percent-of-total sales:

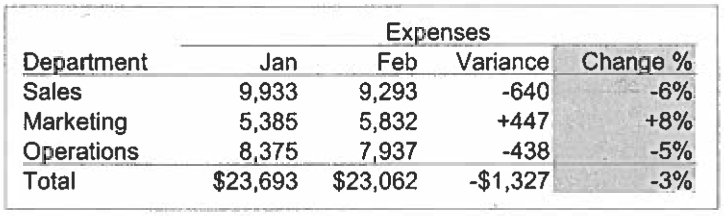


Company	Sales	Sales %
Company A	239,949,993	15%
Company B	873,777,473	54%
Company C	37,736,336	2%
Company D	63,874,773	4%
Company E	399,399,948	24%
Total	$1,614,738,523	100%

FIGURE 2.6: This is an example of a part-to-whole ratio.

When you want to compare the size of one part to another or to the whole, it is easier, more to the point, and certainly more efficient for your audience to interpret a table or graph that contains values expressed as percentages. This is true because percentages provide a common denominator, a common frame of reference—not just any common denominator but one with the nice round value of 100, which makes comparisons very easy to understand.

Another common use of ratios in business involves measures of change. When the value of something is tracked through time, it is often useful to note how it changes from one point in time to the next. Here's a common example of a ratio used to express change, in this case change in expenses from one month to the next:



		Expenses		
Department	Jan	Feb	Variance	Change %
Sales	9,933	9,293	-640	-6%
Marketing	5,385	5,832	+447	+8%
Operations	8,375	7,937	-438	-5%
Total	$23,693	$23,062	-$1,327	-3%

FIGURE 2.7 This is an example of a ratio used to compare the expenses from one month to the next.

CORRELATION

A *correlation* is a relationship in which the values of two paired sets of quantities are compared to determine whether increases in one set correspond to either increases or decreases in the other set. For instance, is there a correlation between the number of years employees have been doing particular jobs and their productivity in those jobs? Does productivity increase along with tenure, does it decrease, or is there no significant correlation in either direction?

Thus far we've learned the following about quantitative information:

- Quantitative information consists of two types of data:
  - Quantitative
  - Categorical
- Quantitative information always describes relationships.
- These relationships involve either
  - Simple associations between quantitative values and categorical subdivisions or
  - More complex associations among multiple sets of quantitative values.
- There are four types of relationships within categories:
  - Nominal
  - Ordinal
  - Interval
  - Hierarchical
- There are three types of relationships between quantitative values:
  - Ranking
  - Ratio
  - Correlation

We have not covered a comprehensive list of possible quantitative relationships. Rather, we've homed in on those that are most relevant to the common uses of numbers in business. If you're wondering why these different types of quantitative relationships are important enough to cover in this chapter, hold on for a while. When we get to the later chapters on tables and graphs, the significance of these quantitative relationships and your ability to identify them will become clear. You'll discover that there are many distinct table and graph design methods and principles that connect directly to these different quantitative relationships.

Numbers that Summarize

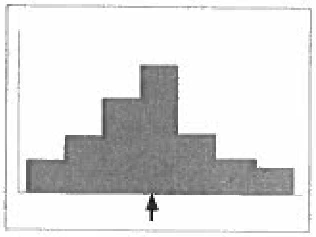
Now for some basic *business statistics*. This is liable to be one of the briefest and simplest looks at statistics that you'll ever encounter. The truth is, most of us who communicate quantitative business information can get by with a limited statistical vocabulary.

Statistics provide several methods for data reduction—in other words, *summarization*, or what we sometimes call *aggregation*. Often, your quantitative message is best communicated by reducing large sets of numbers to a few numbers, allowing your readers to easily and efficiently comprehend and assimilate the message the numbers convey. If an executive asks you how sales are doing this quarter, you wouldn't give her a report that listed each individual sales order; you would give her the information in summary form. Relevant data might include such aggregates as the *sum* of sales orders in U.S. dollars, the *count* of sales orders, and perhaps even the *average* sales order size in U.S. dollars.

We have several ways to summarize numbers, some of which are visual in nature and apply only to graphs, which we'll thoroughly explore later, and some of which are purely statistical in nature, which is our focus in this chapter. Summing and counting sets of numbers are the most common means of aggregation used in business-related quantitative communication. Because I assume that you already understand counts and sums, we'll skip them and proceed directly to the other more complex data reduction methods that are particularly useful in business.

Measures of Average

Let's begin with a question. Take a moment to finish this sentence: "An average represents . . ."



• • • • • • • •

It's interesting how many terms we carry around in our heads and use without ever really knowing how to define them. Ever had a child ask you what something quite familiar means and found yourself struggling for adequate words? If the concept of an average is one of those terms for you, here's a definition:

> An average is a single number that represents the middle of an entire set of numbers.

There are actually four distinct entities that are used in statistics to measure the middle of a set of quantitative values, and all of them are called averages:

- Mean
- Median
- Mode
- Midrange

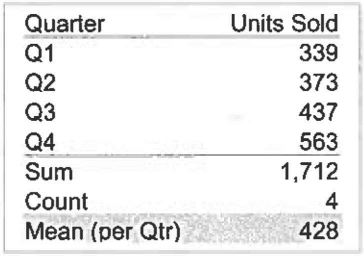
It's useful to understand how these four differ, for they are each designed to work best in particular circumstances. Selecting the wrong type of average for your message could result in misleading information.

MEAN

Normally, when most of us think of an average, we think of what is more precisely called the *arithmetic mean* or simply the *mean*. In fact, many software products label the function that calculates the mean as "Average" (or sometimes "AVG" for short). Statisticians must cringe when they see this. Statistical software wouldn't make this mistake. Means are calculated as follows:

> Sum all the values; then, divide the result by the number of values.

Here's an example:

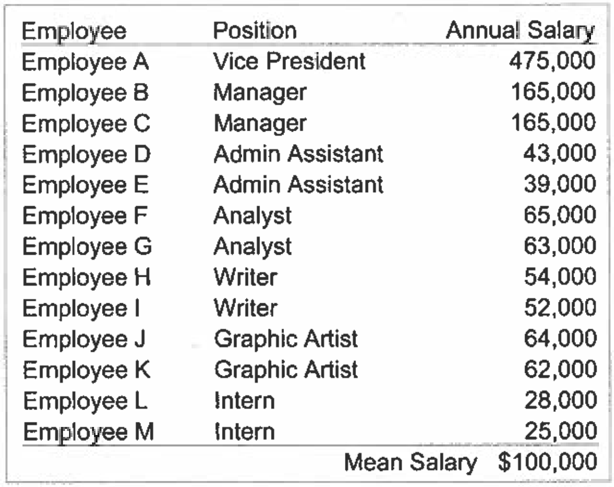


Quarter	Units Sold
Q1	339
Q2	373
Q3	437
Q4	563
Sum	1,712
Count	4
Mean (per Qtr)	428

FIGURE 2.8 This is an example of a *mean*, calculated as 1,712 (the sum) divided by 4 (the count), equaling 428.

A mean is the simplest type of average to calculate (excluding the *midrange*, which is seldom used), and the type most commonly supported by software. However, the mean isn't always the best choice for your message.

Means provide a measure of the middle in a manner that takes every value into account, no matter how extreme. Sometimes this is exactly what you need, but sometimes not. Take a look at the following example, and see if you can determine why use of the mean would be a misleading summary of employee salaries in the marketing department if your intention is to express the *typical* salary.



Employee	Position	Annual Salary
Employee A	Vice President	475,000
Employee B	Manager	165,000
Employee C	Manager	165,000
Employee D	Admin Assistant	43,000
Employee E	Admin Assistant	39,000
Employee F	Analyst	65,000
Employee G	Analyst	63,000
Employee H	Writer	54,000
Employee I	Writer	52,000
Employee J	Graphic Artist	64,000
Employee K	Graphic Artist	62,000
Employee L	Intern	28,000
Employee M	Intern	25,000
	Mean Salary	$100,000

FIGURE 2.9 This is an example of the use of a statistical mean in circumstances for which it is not well suited.

Why doesn't the mean work well for this purpose? The mean in this case is skewed heavily toward the higher salaries, giving the impression that employees are typically better compensated than they actually are. What you're seeing here is the fact that the mean is very sensitive to extremes. The Vice President's salary is definitely an extreme, a value that falls far outside the norm. When you need a measure that represents what is typical of a set of values, you would want to use an average that is not so sensitive to extremes.

Statisticians refer to extreme values in a data set (i.e., those that are located far away from most of the values) as *outliers*. The Vice President's salary in *Figure 2.9* is an outlier.

MEDIAN

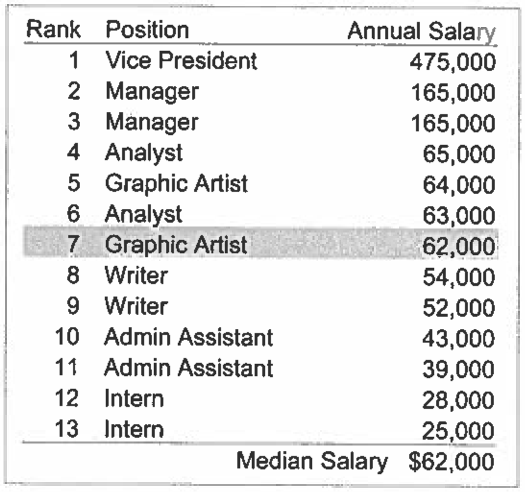
The statistical *median* is the average that comes in handy when you need to communicate quantitative messages such as the one in the above example because the median is not at all sensitive to extreme values.

Medians are calculated as follows:

> Sort the values in order (either high to low or low to high); then, find the value that falls in the middle of the set.

If you are using software or a calculator that supports the calculation of the median, you won't need to sort the set of numbers and manually select the middle value.

Here are the same salaries, but this time we'll determine the median:



Rank	Position	Annual Salary
1	Vice President	475,000
2	Manager	165,000
3	Manager	165,000
4	Analyst	65,000
5	Graphic Artist	64,000
6	Analyst	63,000
7	Graphic Artist	62,000
8	Writer	54,000
9	Writer	52,000
10	Admin Assistant	43,000
11	Admin Assistant	39,000
12	Intern	28,000
13	Intern	25,000
	Median Salary	$62,000

FIGURE 2.10 This is an example of the use of the statistical median.

This data set contains 13 values, so the value that resides precisely in the middle is the seventh, which is $62,000. If you want to communicate the typical marketing department salary, $62,000 would do a better job than $100,000. If your purpose is to summarize the salaries of each department in the company to show their comparative impact on expenses, however, which type of average would work better: the median or the mean? In this case the mean would be the better choice because you want a number that fully takes all values into account, including the extremes. To ignore them through use of the median would undervalue the financial impact.

The median is actually an example of a special kind of value called a *percentile*. A percentile expresses the percentage of values that fall below a particular value. The median is another name for the 50th percentile; that is, it expresses the value below which 50% of the values in the set fall.

You may have noticed while considering how to determine the median above that I ignored a potential complication in the process. What do you do if your data set contains an even number of values, rather than an odd number like the 13 employee salaries above? You simply take the *two* values that fall in the middle of the set (e.g., the fifth and sixth values in a set of ten); then, determine the value halfway in between the two. In fact, you can use the same method that you use for calculating the mean to find the value halfway between the two middle values: sum the two middle values then divide the result by two. If you're using software or a calculator to determine the median, this process is handled for you automatically.

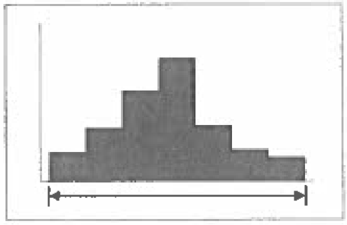
MODE AND MIDRANGE

The two remaining types of averages, modes and midranges, are rarely useful in business, but let's take a moment to understand what they are.

The *mode* is simply the value that appears most often in a set of values. In the set of marketing department salaries that we examined previously, the mode is $165,000 because this is the only value that appears more than once in the set. As you can see, the mode wouldn't be a useful means of expressing the middle of marketing department salaries. The most common value in a data set isn't necessarily anywhere near the middle. If no value appears more than once, the set doesn't even have a mode. If two values appear twice in the set, the set is *bimodal*. If more than two values appear more than once with the same degree of frequency, the set is *multimodal*. Modes are rarely useful for business purposes.

The last method for summarizing the middle of a set of values is the simplest to calculate, but you get what you pay for. It's called the *midrange*. The midrange is the value midway between the highest and lowest values in a set of values. To calculate the midrange, you find the highest and lowest values in the set, add them together, and then divide the result by two. This method is an extremely fast way to calculate an average. If you're on the spot for a quick estimate, you can use the midrange. Be careful, though, for unless the values in the set are distributed evenly across the range, the midrange is far too sensitive to the extremes of the highest and lowest values. You're always better off using the mean or the median.

Measures of Distribution



At times you need to communicate more than the center of a set of values. For example, sometimes you need to communicate the degree to which values vary—the range across which the values are distributed. Two sets of values can have exactly the same average value, but one set could be spread across a broad range while the other is tightly grouped around its average. In some cases this difference is significant. Values that vary widely are volatile. Perhaps they shouldn't be, so you're helping the business by pointing it out. For example, if salaries for the same position vary greatly across your company, this may be a problem worth noting and correcting. It may be useful to recognize and communicate to senior management that sales in January for the past 10 years were always only 4% of annual sales, varying no more than half a percent either way from year to year. Such a pattern, with no significant variation, despite expensive marketing campaigns, may indicate that you should save your marketing budget for later in the year. Values that fall far outside the normal range may indicate underlying problems or even extraordinary successes that should be investigated. A salesperson with an unusually high order-return ratio may be selling his customers products they don't need. A department with exceptionally low expenses per employee may have something useful to share with the rest of the company.

Another term for measures of distribution is *variation*.

The distribution of a set of values can be expressed succinctly through the use of a single number, but there are multiple methods for expressing distributions. We will examine the two that are most useful for business purposes:

- Range
- Standard Deviation

Like averages, these two measures of distribution each work best in specific circumstances. Let's use an example consisting of two sets of values to illustrate these circumstances. Imagine that you work for a manufacturer that uses two warehouses to handle the storage of inventory and the shipping of orders. You've been receiving complaints from customers about the shipments of orders from Warehouse B. To simplify the example, let's say that you've gathered information from each warehouse about shipments of 12 orders of the same product during the same period of time. Ordinarily you would gather shipment information for a much larger number of orders to ensure a statistically significant sample of data, but we'll stick with a small data set to keep the example simple. Here are the relevant values, which in this case are the number of days it took for each of the 12 orders to be processed, from the time each order was received to the time it was shipped:

	Days to Ship from	
Order	Warehouse A	Warehouse B
1	3	1
2	3	1
3	3	1
4	4	3
5	4	3
6	4	4
7	5	5
8	5	5
9	5	5
10	5	6
11	5	7
12	5	10

FIGURE 2.11 This table shows the days it took to ship two sets of 12 orders, one set from Warehouse A and one from Warehouse B.

Because the use of sums and averages is such a common way of analyzing and summarizing quantitative information, you could begin by performing these calculations, resulting in the following:

Warehouse	Sum	Mean	Median
A	51	4.25	4.5
B	51	4.25	4.5

FIGURE 2.12 This table contains various numbers that summarize the number of days it took the two warehouses to each ship a set of 12 orders.

If you were locked into this one way of summarizing and comparing sets of numbers, however, you might conclude and consequently communicate that the service provided by Warehouse B is equal to that of Warehouse A. If you did, you would be wrong.

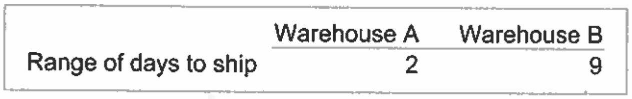
The significant difference in performance between the two warehouses jumps out at you when you focus on the variation. Warehouse A provides a consistent level of service, always shipping orders in three to five days from the date they're received from the customer. Warehouse B, in contrast, is all over the map. Sometimes it fulfills orders much faster than Warehouse A, and at others times its performance is much slower. It's likely that the complaints came from customers who received their orders after waiting longer than five days and perhaps also from regular customers who, like most, value consistency in service, and find it annoying to receive their orders anywhere from one to 10 days after placing them. Given this message about the inconsistent performance of Warehouse B, let's take a look at the two available ways to measure and communicate this variation.

RANGE

The simpler of the two methods is called the *range*. You can calculate the range as follows:

> Subtract the lowest value from the highest value.

That's it. This is a simple measure of distribution that everyone can understand, which is its strength. To summarize the variation in the performance of Warehouse A versus Warehouse B, you could do so as simply as follows:



	Warehouse A	Warehouse B
Range of days to ship	2	9

FIGURE 2.13 This table shows the ranges of days it took the warehouses to ship the two sets of orders.

Similar to the midrange averaging method, the range method of measuring the distribution of values suffers from its dependence on too few data (only the highest and lowest values), which robs it of the greater accuracy and usefulness of the standard deviation method, which we'll examine next. If Warehouse B had shipped seven orders in five days, including one order in one day and one order in 10 days, that would be a different story from the one contained in our data, but the range would be the same. Nevertheless, everyone can understand a range, which makes it a useful way to communicate distributions to audiences that haven't learned to interpret more complicated measures, such as standard deviations.

STANDARD DEVIATION

The measure of variation that is generally the most useful is the *standard deviation*. Here's a definition:

> The standard deviation of a set of values measures
> their distribution relative to the mean.

The bigger the standard deviation, the greater the range of distribution relative to the mean. This becomes a little clearer when you visualize it. First, take a look at the number of days it took Warehouse B to ship each order compared to the mean value of 4.25 days:

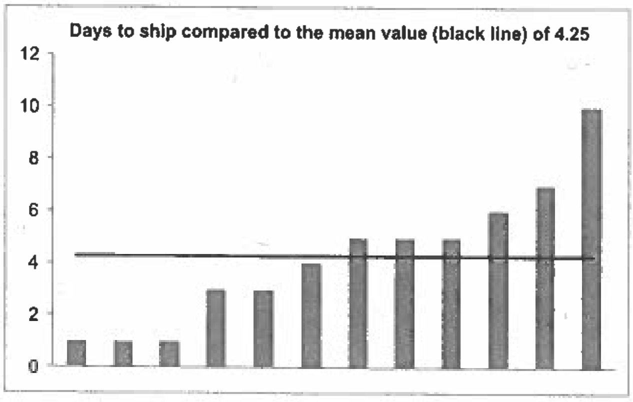


FIGURE 2.14 This graph shows a simple way to visualize the days it took Warehouse B to ship each of the 12 orders compared to the mean value of 4.25 days.

Or, better yet, because our purpose here is to examine the degree to which the shipments of the individual orders varied about the mean, this graph makes it a little easier to visualize:

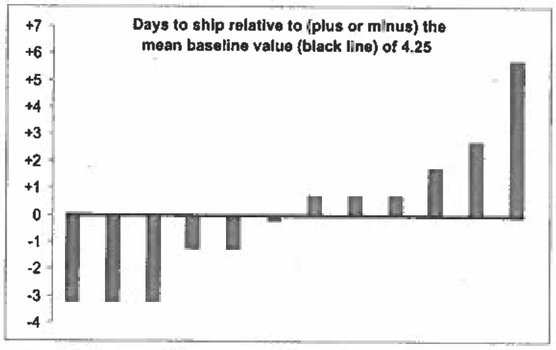


FIGURE 2.15 This graph displays the days it took to ship the individual orders relative to the mean.

So far we haven't displayed the standard deviation. We're still leading up to that. The standard deviation will provide a single value that summarizes the degree to which the 12 shipments as a whole were distributed about the mean (i.e., an average degree of distribution). The standard deviation can be calculated as follows:

1. Calculate the mean of the set of values.
2. Subtract each individual value in the set from the mean, resulting in a list of values that represent the differences of the individual values from the mean.
3. Square each of the values calculated in step 2.
4. Sum the values calculated in the step 3.
5. Divide the value calculated in step 4 by the number of values.
6. Calculate the square root of the value calculated in step 5.[1]

Technically, there are two formulas for calculating a standard deviation, one for the standard deviation of an entire population of values, and one for the standard deviation of a sample set of values. The steps above are used for an entire population of values. If the value set only includes a sample of the entire population of values, step 5 differs in that you divide by the number of values minus 1, rather than simply by the number of values. It is handy to know how to calculate a value like a standard deviation, but you may never need to do the math yourself. Most software products that produce tables and graphs provide a simple means to calculate the standard deviation.

Because the set of values representing the number of days it took for Warehouse B to ship orders is only a sample set of values (i.e., 12 orders that shipped on a particular day), we'll use the form of the calculation used for sample sets, which produces a standard deviation of 2.58602 days. We can round this figure off to 2.59. This compares to a standard deviation for Warehouse A's shipments of 0.83 days. The difference between a standard deviation of 2.59 and one of 0.83 succinctly indicates a much higher degree of variation in Warehouse B's shipping performance when compared to Warehouse A's. Standard deviations are a concise measure that can be used to compare the relative distribution among multiple sets of values.

[1] These steps were derived from Mario F. Triola (2001) *Elementary Statistics*, Eighth Edition. New York: Addison Wesley Longman Inc.

In addition to its use for comparisons, a single standard deviation can tell you something about the degree to which the values are distributed. However, to be able to simply look at a standard deviation and interpret the range of variation that it represents requires that you learn a little more about standard deviations.

In general, when individual instances of almost any type of event are measured, and those measurements are arranged by value from lowest to highest, most values tend to fall somewhere near the center (i.e., near some measure of the middle, such as the mean). The farther you get from the center, the fewer the instances you will find. If you display this in the form of a graph called a *frequency polygon*, which uses a line to trace the frequency of instances that occur for each value from lowest to highest, you have something that looks like a *bell-shaped curve*, formally called a *normal distribution* in statistics.

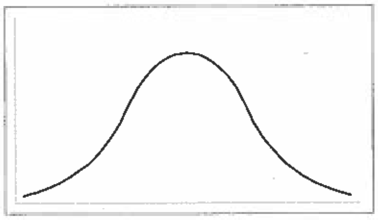


FIGURE 2.16 This curved line represents a *normal distribution.* It displays the frequency of values as they occur from the lowest value at the left to the highest value at the right. Most instances have values near the midpoint of the range of values, which represents the mean. In a perfect normal distribution, the frequency of instances decreases at the same rate to the left and to the right of the mean, resulting a curve (i.e., the black line) that is symmetrical.

The more closely the number of values that you include approaches the entire population of values, the more closely the curve resembles a bell. So what is the significance of a normal distribution to our examination of standard deviations? When you have a normal distribution, the standard deviation describes the distribution of the values as percentages of the whole. The following figure overlays the normal distribution displayed in the figure above with useful information that the standard deviation reveals.

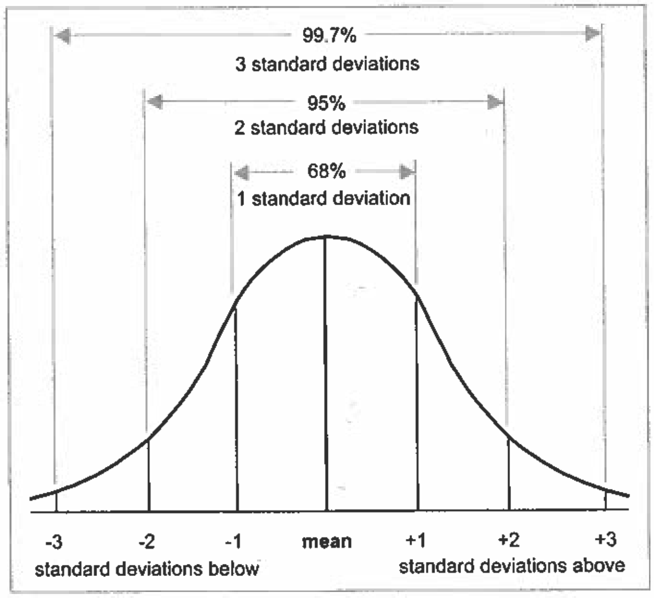


FIGURE 2.17 This figure shows a normal distribution of values in relation to the standard deviations of those values. The percentages of values that fall within one, two, and three standard deviations from the mean can be predicted with a normal distribution, and consequently can be predicted to a fair degree with anything that is close to a normal distribution. This is called the *empirical rule.*

With normal distributions, 68% of the values fall within one standard deviation above and below the mean, 95% fall within two standard deviations, and 99.7% fall within three. Stated differently, if you are dealing with a distribution of values that is close to normal, you automatically know that one standard deviation from the mean represents approximately 68% of the values, two represent approximately 95%, and so on. Given this knowledge, the standard deviation of a set of values has meaning in and of itself, not just as a tool for comparing the degree of distribution between two or more sets of values. The bigger the standard deviation, the broader the range of values, and thus the greater difference in distribution between them.

How does this relate to your world and the types of business phenomena that you study and communicate? Take a couple of minutes to list a few examples that are good candidates for measures of distribution. In what situations does distribution indicate something important to your business?

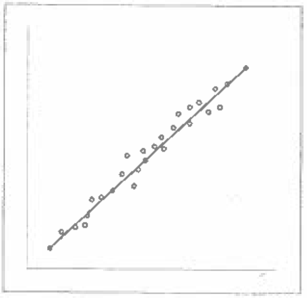
* * * * * * *

Here are a few examples that I've encountered:

- Distribution of the *selling price* of specific products or services. Is the distribution greater for some parts of the world or for some sales representatives than for others? Do the differences in distribution correspond to increased or decreased profits?
- Different distributions in measures of *performance*, such as the time it takes to manufacture products, answer phone calls, or resolve technical problems. Do instances of greater distribution indicate problems in training, employee morale, process design, or systems? What does a greater degree of distribution today compared to the past signify?
- Distribution of employee *compensation*. Why is there such a discrepancy in compensation for the same job in different departments? Does the broad distribution of salaries have an effect on employee morale or performance?
- Distribution of the *cost of goods* purchased by various buyers from various vendors. Why is the distribution of costs associated with some buyers so much greater than the distribution associated with others for the same goods?
- Distribution of the departmental *expenses*. How is it that some departments are managing to keep their expenses so much lower than other departments are?

I could go on, but I suspect the point is clear. Measures of distribution tell important stories, so familiarity with the available methods for summarizing and concisely communicating these messages is indeed useful.

Measures of Correlation



Earlier in this chapter, I described *correlation* as a particular type of quantitative relationship where two paired sets of values are compared to one another to see whether they correspond in some manner. For instance, does tenure on the job relate to productivity? In this section we are going to look at a particular way to measure correlation and express it as a single value. This single value is called the *linear correlation coefficient*. It answers each of the following questions about the correlation of two paired sets of quantitative values:

- Does a correlation exist?
- If so, is it strong or weak?
- If so, is it positive or negative?

Here's a concise definition:

> The linear correlation coefficient measures the direction (positive or negative) and degree (strong or weak) of the linear relationship between two paired sets of values.

By *two paired sets of values* I mean the two sets of values that are involved when you examine the relationship of one thing to another, such as an employee's tenure (e.g., number of years on the job) to his productivity on the job (e.g., number of products manufactured per hour). In this case, the two measures for each employee constitute a paired set of values. By *linear correlation* I mean a consistent relationship between two things; for instance, if you measure the correlation between employee tenure and productivity, and find that as tenure increases productivity increases, or that as tenure increases productivity decreases. However, a linear correlation cannot represent a relationship that varies, for example, if productivity increases along with tenure to a point but after that point an increase in tenure results in a productivity decrease. This is clearly a relationship, but it is *nonlinear*. The *direction* of a correlation is either positive or negative. With positive correlations between two sets of values (A and B), as the value of A increases, the value of B likewise increases and as the value A decreases, so does B. With negative correlations, as the value of A increases, the value of B decreases, and vice versa.

If you had to calculate the linear correlation coefficient manually, you would have to work through several steps. Very few of us who work with business numbers need to do so because we have software or calculators to do this for us. What really matters is that we know how to interpret the resulting value, so let's focus on the number itself and what it means.

Despite its intimidating name, the linear correlation coefficient is actually quite simple to interpret. Here are a few guidelines:

- All linear correlation coefficient values fall somewhere between +1 and -1.
- A value of 0 indicates that there is no correlation.
- A value of +1 indicates that there is a perfect positive correlation.
- A value of -1 indicates that there is a perfect negative correlation.
- The greater the value, in either the positive or the negative direction, the stronger the correlation.

Pretty simple, but it will still help to look at this visually. To do so, we're going to use a graph called a *scatter plot*, which is designed specifically to display the correlation of two paired sets of quantitative values. Perhaps you've seen this type of graph listed as one that is available in software that you use but have never used it, and perhaps have only a vague idea how it works. With a little exposure, you'll find that scatter plots are quite easy to use and interpret as well as quite useful for revealing and communicating quantitative relationships.

Here's a series of scatter plots that will help you visualize the types of relationships that a linear correlation coefficient is designed to reveal. Each graph displays the relationship between two paired sets of values, one horizontally along the X axis and one vertically along the Y axis. When you read a scatter plot, you should look for what happens to the value along the X axis in relation to the value along the Y axis. As X goes up, what happens to Y? As X goes down, what happens to Y? Is the relationship strong (i.e., close to a straight line) or is it weak (i.e., bounces around)? Is it positive (i.e., moves upward from left to right) or is it negative (i.e., moves downward from left to right)? Each of the following graphs displays a different relationship between the variable plotted along the X axis (horizontal) and the variable plotted along the Y axis (vertical), with the linear correlation coefficient in parentheses to help you understand its meaning.

A *variable* is simply something with values that can vary, such as employee productivity.

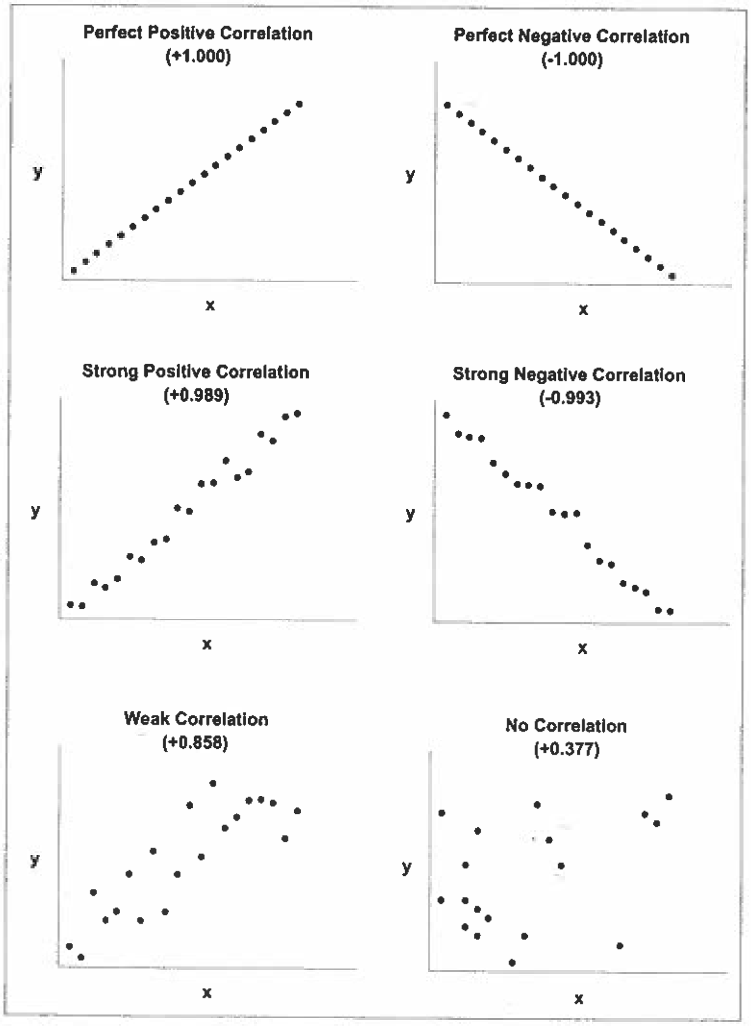


FIGURE 2.18 This is a series of scatter plots, each of which displays a different relationship between two sets of paired values (e.g., employee tenure and productivity).

Bear in mind that these scatter plots simply provide examples of correlations. If the linear correlation coefficient in the left-middle scatter plot was +0.970 rather than +0.989, it would still represent a strong positive relationship.

One way of looking at correlations as displayed in scatter plots is to imagine a straight line that passes through the center of the dots; then, determine the strength of the correlation based on the degree to which the dots are tightly grouped around that line: the tighter the grouping, the stronger the relationship. Here are examples of how scatter plots would look if you actually drew the lines:

Drawing a straight line of *best fit* through the center of a series of points on a scatter plot is a common technique for highlighting the relationships between two sets of values. It's called a *trend line, line of best fit,* or, more formally, a *regression line.*

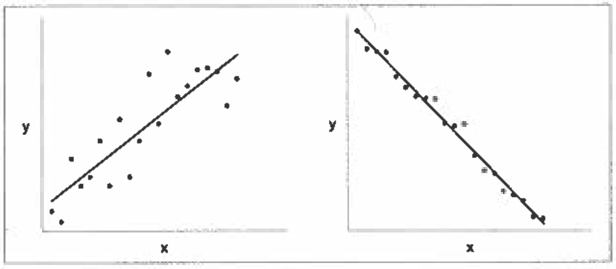


FIGURE 2.19 These are scatter plots with lines of best fit through the center of the dots to clearly delineate the nature of the relationship.

Based on what you've learned about scatter plots, how would you describe each of the relationships displayed above?

• • • • • • •

In the scatter plot on the left, the characteristics that you must consider are:

- The direction of the line, which in this case is upward from left to right
- The closeness of the grouping of dots around the line, which in this case is not terribly tight

Given these two observations, we can say that the scatter plot on the left depicts a correlation that is positive (i.e., upward from left to right) but not extremely strong (i.e., not tightly grouped around the line). Using this same method of interpretation, the scatter plot on the right depicts a correlation that is negative and very strong but not perfectly so.

At this point, you may be wondering: "At what value of a linear correlation coefficient does a correlation cease to be strong and begin to become weak or cease to be weak or even a correlation at all?" There is no precise answer to this question. It depends to some degree on the number of paired values included in your data sets; the more values you have, the greater confidence you can have in the validity of the linear correlation coefficient. Because our purpose here is not to delve too deeply into the realm of statistics, let's be content with the knowledge that values close to 1 in positive correlations and close to -1 in negative correlations indicate strong relationships and that the closer they are to 1 or -1, the stronger the relationship.

For an excellent introduction to statistics, including much more information than I've provided about correlations, I recommend the textbook by Mario F. Triola (2001) *Elementary Statistics*, Eighth Edition. New York: Addison Wesley Longman Inc.

Remember, linear correlation coefficients can only describe relationships that are linear—that is, ones that move in one direction or another—but not relationships that are positive under some circumstances and negative under others. Here's such an example:

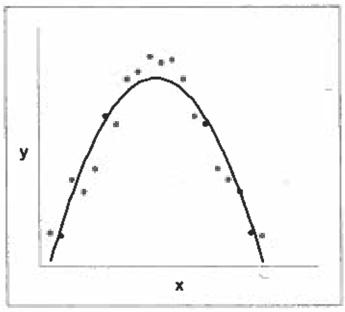


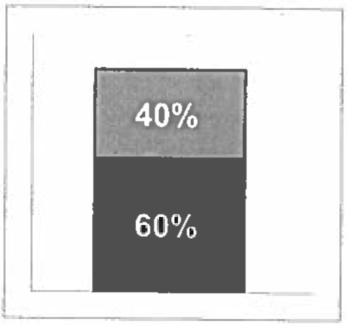
FIGURE 2.20 This is an example of a nonlinear correlation.

What you see here is definitely a correlation, but it certainly isn't linear. If this scatter plot represents the relationship between employee tenure (i.e., years on the job) on the X axis and employee productivity on the Y axis, how would you interpret this relationship, and how might you explain what is happening to productivity after employees reach a certain point in their tenure?

After studying this scatter plot and double-checking the data, you would likely suggest that something be done as employees reach the halfway point along their tenure timelines, such as offering new incentives to keep them motivated or retraining for new positions that they might find more interesting.

Measures of Ratio

In contrast to correlations, which measure the relationship between multiple paired sets of values, a ratio measures the relationship between a single pair of values. A typical example that we encounter in business is the book-to-bill rate, which is a comparison between the value associated with sales orders that have been booked (i.e., placed by the customer and accepted as viable orders) and the value associated with actual billings that have been generated in response to orders.



Ratios can be expressed in four ways:

- As a *sentence*, such as "Two out of every five customers who access our web site place an order."
- As a *fraction*, such as 2/5 (i.e., 2 divided by 5)
- As a *rate*, such as 0.4 (i.e., the result of the division expressed by the fraction above)
- As a *percentage*, such as 40% (i.e., the rate above multiplied by 100, followed by a percent sign)

Each of these expressions is useful in different contexts, but rates and percentages are the most concise and therefore the most useful for tables and graphs. Many measures of ratio have conventional forms of expression, such as the book-to-bill rate mentioned above, which is typically expressed as a rate (e.g., 1.25, which indicates that for every five orders that have booked, only four have been billed, or 5/4 = 1.25), or the profit margin, which is normally expressed as a percentage (e.g., 25%, which indicates that for every $100 of revenue, $75 goes toward expenses, leaving a profit of $25, or $25/$100 = 0.25 * 100 = 25%).

Take a moment to think about and list a few of the ways that ratios are used, or could be used, to communicate quantitative information related to your own work.

Ratios are simple shorthand for expressing the direct relationship between two values. One especially handy use of ratios is to compare several individual values to a particular value to show how they differ. In this case, your purpose is not to compare the actual values but to show the degree to which they differ. In such circumstances, you can simplify the message by setting the main value to which you are comparing all the others to a baseline of 1 (expressed as a rate) or 100% (expressed as a percentage); then, express the other values as ratios that fall above or below that baseline. Here's an example expressed in percentages:

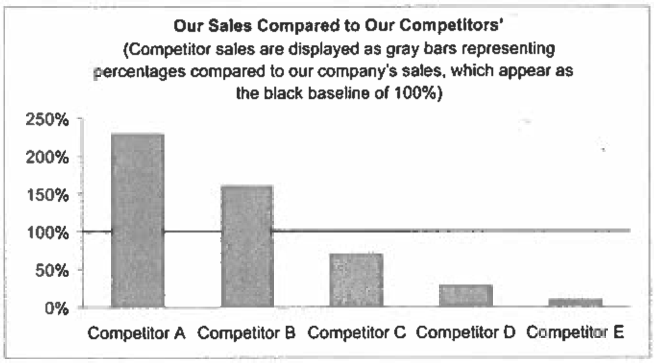


FIGURE 2.21 This graph includes a baseline of 100% for the primary set of values, making it easy to see how the other values, also expressed as percentages, differ.

By using a baseline, it is easy to see that the main competitor does about 250% of your company's sales, which is 2.5 times as much when expressed as a rate. Expressing the comparison in this manner eliminates the need for readers to do calculations in their heads when they want to think in terms of relative differences.

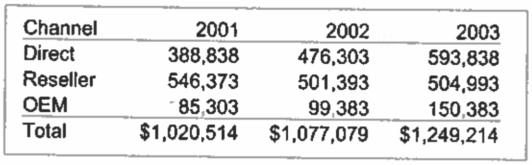
Measures of Money

Most quantitative information that we encounter in business involves some currency of exchange—in other words, money. Be it U.S. or Canadian dollars, Japanese yen, British pounds, Swiss francs, or the newer Euro, money is at the center of most business analysis and reporting. Unlike most other units of measure, currency has a characteristic that we must keep in mind when communicating information that spans time: the value of money is not static; it changes with time. The value of a U.S. dollar in November of 2001 was not the same as its value in November of 2002. If you've been asked to prepare a report that exhibits the trend of sales in U.S. dollars for the past three years, would you be justified in asserting that sales have increased by 20% during that time if three years ago annual sales were $100 million and today they total $120 million? Only if the value of a dollar today is the same as it was three years ago, which it isn't.

When the value of a dollar decreases over time, we refer to this as *inflation*. When comparing money across time, an accurate comparison can only be made when you adjust for inflation. I've found, however, that in business reporting this is rarely done. Despite the validity of the argument in favor for adjusting for inflation, doing so isn't always practical, so I won't attempt to force on you a practice that you may very well ignore. For those of you who can take extra time required to correct for skewed results due to inflation, I've included Appendix C, *Adjusting for Inflation*, in the back of the book. It isn't difficult, and the practice will improve the quality of your financial reporting.

Business today, especially in large companies, is often international and entails multiple currencies. This is a problem when we must produce reports that combine data across multiple currencies, such as sales in the Americas, Europe, and Asia. You can't just throw the numbers together because 100,000 U.S. dollars does not equal 100,000 British pounds or 100,000 Japanese yen. To combine them or to compare them, you must convert them all into a single currency. Fortunately, most operational software systems that we use to run our businesses today are designed to do this work for us, converting money based on tables of exchange rates, so we can easily see transactions both in their original currency and in some common currency used for international reporting, such as U.S. dollars. Because software typically does this work for us, my intention here is simply to caution you to avoid mixing currencies without converting them to a common currency. If you're not careful, you could inadvertently report results that are in error by a large order of magnitude.

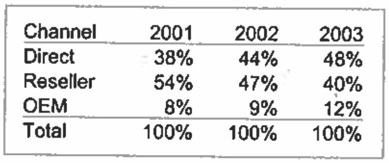
When your purpose is to compare monetary values, such as those associated with different categories (geographical area, departments, etc.), you can often avoid the challenge of mixing multiple currencies by expressing the numbers in the form of rates or percentages. For instance, if you want to compare the annual sales through your various sales channels (direct sales force, distributors, etc.) for the past three years by expressing the sales of each sales channel as a percentage of the whole rather than as currency, you not only avoid all problems associated with inflation and multiple currencies, you also present the numbers in a manner that speaks the message of comparative sales directly and clearly. If you wanted to see such a comparison, would you prefer this table . . .



Channel	2001	2002	2003
Direct	388,838	476,303	593,838
Reseller	546,373	501,393	504,993
OEM	85,303	99,383	150,383
Total	$1,020,514	$1,077,079	$1,249,214

FIGURE 2.22 This table displays a comparison of sales by sales channel, expressed as dollars.

. . . or this one?



Channel	2001	2002	2003
Direct	38%	44%	48%
Reseller	54%	47%	40%
OEM	8%	9%	12%
Total	100%	100%	100%

FIGURE 2.23 This table displays a comparison of sales by sales channel, expressed as percentages of total sales.

As you can see, the second approach completely eliminated the need to account for inflation and multiple currencies.

Understanding the relationships we've examined in this chapter lays the foundation that will help you design tables and graphs to effectively communicate quantitative information. In the next chapter, we'll look at the basics of tables and graphs and begin to see how they can effectively present the kinds of relationships we've just discussed in Chapter 2.

Summary at a Glance

Quantitative Relationships

- Quantitative information consists of two types of data:
  - Quantitative
  - Categorical
- Quantitative information always describes relationships.
- These relationships involve either
  - Simple associations between quantitative values and categorical subdivisions or
  - More complex associations among multiple sets of quantitative values.
- There are four types of relationships within categories:
  - Nominal
  - Ordinal
  - Interval
  - Hierarchical
- There are three types of relationships between quantitative values:
  - Ranking
  - Ratio
  - Correlation

Numbers that Summarize

Type of Summary	*Method*	*Note*
Average	Mean	Measures the center of a set of values in a manner that is equally sensitive to all values, including extremes
	Median	Measures the center of a set of values in a manner that is insensitive to extreme values
Distribution	Range	Simple to calculate, relying entirely on the highest and lowest values, but only roughly defines a ranges of values
	Standard Deviation	Provides a rich expression of the distribution of a set of values across its entire range
Correlation	Linear Correlation Coefficient	Indicates whether a correlation exists between two paired sets of values, and if so, its direction (positive or negative) and its strength (strong or weak)
Ratio	Rate or Percentage	Measures the direct relationship between two quantitative values

Measures of Money

- When comparisons of monetary value are expressed across time, adjusting the value to account for inflation produces the most accurate results.
- When reporting monetary values that combine multiple currencies, you must first convert them all into a common currency.