2 SIMPLE STATISTICS TO GET YOU STARTED

Quantitative information forms the core of what organizations must know to operate effectively. The current emphasis on metrics, Key Performance Indicators (KPIs), Balanced Scorecards, and performance dashboards demonstrates the importance of numbers to organizations today. Stories contained in numbers can be communicated most effectively when we understand the fundamental characteristics and meanings of simple statistics that are routinely used to make sense of numbers, as well as the fundamental principles of effective communication that apply specifically to quantitative information.

Numbers are neither intrinsically boring nor interesting. The fact that they are quantitative in nature 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, flow naturally from the significance and relevance of the message the information contains. As a communicator, it is up to you to give a clear and unhindered voice to that information and its story, using language that is easily understood by your audience.

You might 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 substance of tables and graphs, it's important to begin our journey by getting acquainted with a few numbers that are particularly useful.

Quantitative Relationships

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

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 (geography, time, etc.). Quantitative stories feature two types of data: *quantitative* and *categorical.* Quantitative values measure things; categories divide information into useful groups, and the items that make up each category identify the things that are measured. For example, geographical areas (e.g., north, east, south, and west) are items in a category that might be called sales regions and months are items in a category called time. This distinction between quantitative values and categorical items is fundamental to tables and graphs. Quantitative values and categorical items serve different but complementary purposes and are often structured and displayed in distinct ways.

Quantitative values are 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 quantitative relationships we display are simple associations between quantitative values and the categorical items that label them, such as those in the previous examples. Sometimes the relationships display direct associations between different sets of quantitative values, such as the number of marketing emails that were sent in relation to the resulting number of orders received, or the percentage of times that doctors forget to wash their hands in hospitals in relation to the percentage of infections. This distinction between simple relationships that associate quantitative values and categorical items 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, but the list of potential quantitative relationships is endless. Think for a minute or two about the quantitative information that's important to your organization. Can you think of any information that doesn't involve relationships?

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

- Quantitative stories include two types of data:
  - Quantitative
  - Categorical
- Quantitative stories always feature relationships.
- These relationships involve either:
  - Simple associations between quantitative values and corresponding categorical items 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 items or the quantitative values associated with them can relate to one another. Let's take a look.

Relationships Within Categories

Categorical items that we use in tables and graphs to label corresponding measures can relate to one another in the following ways:

- Nominal
- Ordinal
- Interval
- Hierarchical

NOMINAL

A *nominal* relationship is one in which the values in a single category are discrete and have no intrinsic order. For instance, the four sales regions East, West, North, and South have no particular proper order in and of themselves. These labels simply name the different sales regions, thus the term "nominal," which means "in name only." Here's an 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 tell a quantitative story that is nominal in nature, you associate the quantitative values with the corresponding categorical labels, but your story does not relate the categorical items to one another in any particular way.

ORDINAL

In an *ordinal* relationship, the categorical items have a prescribed *order.* Typical examples include "first, second, third..."; "small, medium, and large"; and "best salesperson, second best salesperson...". To display them in any other order, except in reverse, would rarely be meaningful.

INTERVAL

An *interval* relationship is one in which the categorical items consist of a sequential series of numerical ranges that subdivide a larger range of quantitative values into smaller ranges. These smaller numerical ranges, called intervals, are arranged in order from smallest to largest. Here's a typical example:

Order Size (U.S. $)	Order Quantity	Order Revenue
>= 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	$204,515,383

FIGURE 2.2 This is an example of an interval relationship. Notice that Order Size consists of a sequential series of numerical ranges that subdivide a larger range of quantitative values.

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 sequential, equally sized intervals.

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 items of time (months in this case).

Department	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 call this an ordinal relationship, for months usually make sense only when arranged chronologically. However, items that make up an interval scale also have a proper order, which invites the question: "Do these units of time represent intervals along a quantitative scale? The answer is "Yes, they do." Time is a quantitative scale that measures duration. Even though the months in the example above do not all represent the same exact number of days and are therefore not equally sized intervals, for reporting purposes we treat them as equal.

So far the categorical relationships that we've examined involve relationships between items in the same category. The remaining relationship, discussed next, is different.

HIERARCHICAL

A *hierarchical* relationship involves multiple categories that are closely associated with one another as separate levels in a series of "parent-to-child" connections. If we start from the top of the hierarchy and progress downward, each item at each level is associated with only one item at the level above it. Each item at every level, except the bottom level, however, can have one or more items 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 used routinely in tables to organize quantitative information.

Relationships Between Quantities

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

- Ranking
- Ratio
- Correlation

RANKING

When the order in which the categorical items are displayed is based on the associated quantitative values, either in ascending 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 story would be enhanced if you arranged them by size, in this case from the largest to the smallest, 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 items 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 that compares two quantitative values 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 item compared to the sum of the entire set of values in the category (e.g., the sales of one region compared to 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 relationship that 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 that have been expressed as percentages. This is true because percentages provide a common denominator and common frame of reference, and not just any common denominator but one with the nice round value of 100, which makes comparisons easy to understand.

Another common use of ratios 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 an example of a ratio that expresses the degree of change, in this case change in expenses from one month to the next:

	Expenses			
Department	Jan	Feb	Change	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	-$631	-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* compares two paired sets of quantitative values to determine whether increases in one correspond to either increases or decreases in the other. 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? Correlations are important to understand, in part because they make it possible to predict what will happen to values in one variable (e.g., sales revenue) by knowing or perhaps even controlling values in another variable (e.g., marketing emails).

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

- Quantitative stories include two types of data:
  - Quantitative
  - Categorical
- Quantitative stories always feature relationships.
- These relationships involve either:
  - Simple associations between quantitative values and corresponding categorical items or
  - More complex associations among multiple sets of quantitative values
- Categorical items exhibit four types of relationships:
  - Nominal
  - Ordinal
  - Interval
  - Hierarchical

- Quantitative values exhibit three types of relationships:
  - Ranking
  - Ratio
  - Correlation

We have not covered a comprehensive list of possible quantitative relationships. Rather, we've considered only those that are most relevant when presenting quantitative information in typical ways. If you're wondering why these different quantitative relationships are important enough to cover in this chapter, hold on for a while. When we get to later chapters on tables and graphs, the importance of these relationships and your ability to identify them will become clear. You'll discover that there are many specific ways to design tables and graphs that tie directly to these specific quantitative relationships.

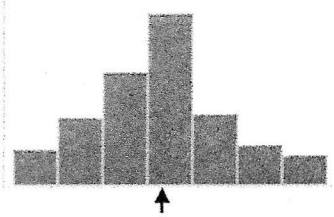
Numbers that Summarize

Statistics provide several ways reduce or summarize data. Summarization is also referred to as *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 story. 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 that are visual in nature and apply only to graphs, which we'll explore more thoroughly later, and some that are purely statistical in nature, which we'll examine now. Summing and counting sets of numbers are the most common aggregations used in quantitative communication. I assume that you already understand counts and sums, so we'll skip them and proceed directly to other less familiar data reduction methods that are also useful.

Measures of Average

Let's begin by examining what we already know about averages. 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 really knowing how to define them. Ever had a child ask you to explain something quite familiar and found yourself struggling for 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 center of an entire set of numbers.

There are actually four distinct ways in statistics to measure the center of a set of numbers, and all of them are called averages:

- Mean
- Median
- Mode
- Midrange

It's useful to understand how these four differ. Selecting the wrong type of average for your message could mislead your audience.

MEAN

Normally, when most of us think of an average, what we have in mind 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"). Statisticians must cringe when they see this. Statistical software wouldn't make this mistake. Means are calculated as follows:

> Sum all the values and 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 a simple form of average to calculate, but isn't always the best choice for your story.

Means measure the center of a set of numbers in a way 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 using the mean would produce a misleading summary of employees' salaries in the marketing department if your objective 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? In this case the mean is much higher than most salaries, giving the impression that employees are better compensated than they typically 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 far above the norm. When you need a measure of center that represents what is typical of a set of values, you should use an average that isn't 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 *median* is an expression of average that comes in handy when you need to tell quantitative stories such as the one in the previous example because the median is not at all sensitive to extremes and therefore does a better job of expressing what's typical.

Medians are calculated as follows:

> Sort the values in order (either high to low or low to high) and 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 sorted from high to low so we can easily 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 clearly do a better job than $100,000. However, if your purpose is to summarize the salaries of each department in the company to show their comparative impact on expenses, 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 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 just another name for the 50th percentile, for it expresses the value on or below which 50% of the values fall.

You might 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? In this case, 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) and then determine the value that's halfway 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.

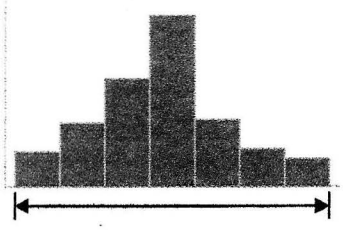
MODE AND MIDRANGE

The two remaining types of averages—modes and midranges—are rarely useful to non-statisticians, but let's take a moment to understand them anyway.

The *mode* is simply the specific 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 center of the marketing department's salaries. The most common value in a data set, especially a small one, isn't necessarily anywhere near the center. If no value appears more than once, the set doesn't even have a mode. If two values appear twice in the set and no other values appear more than once, the set is bimodal. If more than two values appear more than once with the same high degree of frequency, the set is *multimodal*. Modes are rarely useful for most data presentation purposes.

The final method for expressing the center 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, but be careful. Unless the values are uniformly distributed across the range, the midrange is far too sensitive to the extremes of the highest and lowest values. You're better off using the mean or the median.

Measures of Variation



It is often useful to present more than the center of a set of values. For example, sometimes you need to communicate the degree to which values vary, such as the full range across which the values are distributed. Two sets of values can have exactly the same average, but one set could be distributed 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 your organization by pointing this out. For example, if salaries for the same basic position vary greatly within your organization, this may be a problem worth noting and correcting. It might also 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, might indicate that the marketing budget should be saved for later in the year. Values that fall far outside the normal range might indicate underlying problems or even extraordinary successes that should be investigated. A salesperson with an unusually high order-return ratio might be selling products to his customers that they don't need. A department with exceptionally low expenses per employee might have something useful to share with the rest of the company.

Variation in a set of values can be expressed succinctly through the use of a single number, but there are multiple methods. We will examine the two that are typically most useful:

- Spread
- Standard Deviation

Like averages, these two measures of variation 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 shipments 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 set of orders to ensure a statistically significant sample, 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 day each order was received to the day it was shipped:

	Days to Ship	
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 values 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, 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. Shipments from Warehouse B, however, are 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 ten 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.

SPREAD

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

> Subtract the lowest value from the highest value.

That's it. This is a measure of variation that everyone can understand, which is its strength. To summarize variation in the performance of Warehouse A versus Warehouse B, you could do so 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 spread method of describing variation suffers from its reliance on too little information (only the highest and lowest values), which robs it of the greater accuracy and usefulness of the standard deviation method that we'll examine next. The spread also suffers from the fact that it is very affected by extreme values. If Warehouse B had shipped seven orders in 5 days, one order in 1 day, and one order in 10 days, that would be a different story from the one contained in the data, but the spread would be the same. Despite its limitations, spread is a characteristic of variation that's useful to know.

STANDARD DEVIATION

The single measure of variation that reveals more than others is the *standard deviation.* Here's a definition:

> The standard deviation measures variation in a set of values relative to the mean.

The bigger the standard deviation, the greater the range of variation relative to the mean. This becomes a little clearer when you visualize it. Look at the

number of days it took Warehouse B to ship each order compared to the mean value of 4.25 days:

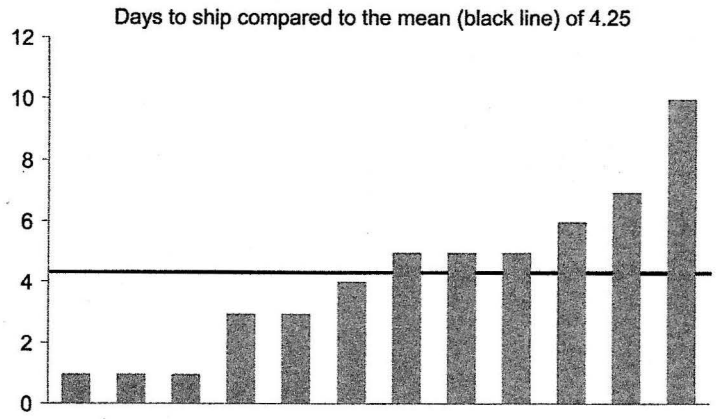


FIGURE 2.14 This graph shows a simple way to visualize the days it took Warehouse B to ship each of the 12 orders compared to the mean value of 4.25 days.

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

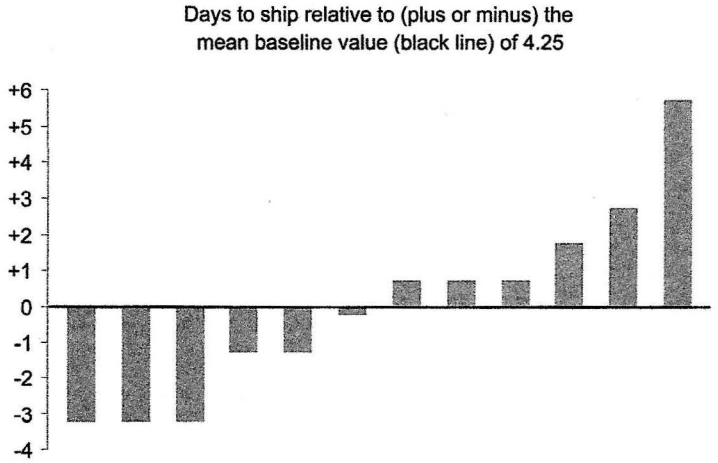


FIGURE 2.15 This graph displays the days it took Warehouse B to ship the individual orders relative to the mean.

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 varied in relation to the mean (i.e., average degree of variation). Standard deviation can be determined 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 total number of values.
6. Calculate the square root of the result from step 5.

Looks like a lot of work, doesn't it? To complicate matters further, there are technically 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 set consists only of a sample of the entire population of values, step 5 above would differ in that you would divide by the number of values minus 1, rather than simply by the number of values. Fortunately, most software products that produce tables and graphs include a simple way to calculate the standard deviation, so you don't have to perform the calculations yourself.

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

In addition to its use for comparisons, a single standard deviation can also tell you something about the degree to which the values vary. However, the ability to look at a standard deviation and interpret the range of variation that it represents requires a little more knowledge.

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 (e.g., the mean). The farther you get from the center, the fewer 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, called a *normal distribution* in statistics.

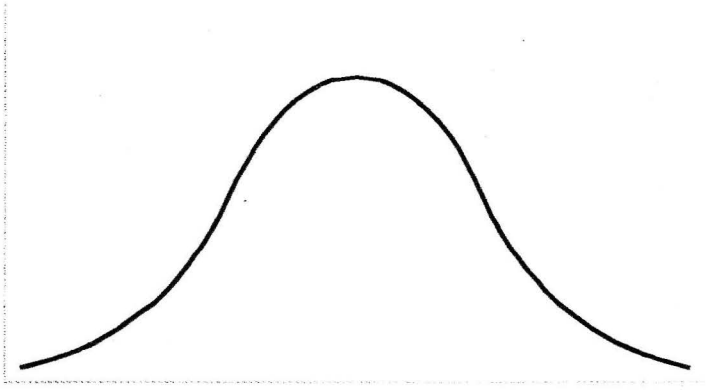


FIGURE 2.16 This curved line represents a normal distribution. It displays the frequency of values as they occur from the lowest value at the left to the highest value at the right. Most instances have values near the midpoint of the set of values, which represents the mean. In a perfect normal distribution, 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.

So how do normal distributions relate to our examination of standard deviations? When you have a normal distribution, the standard deviation describes

variation as percentages of the whole. The following figure overlays the normal distribution displayed in the previous figure with useful information that the standard deviation reveals.

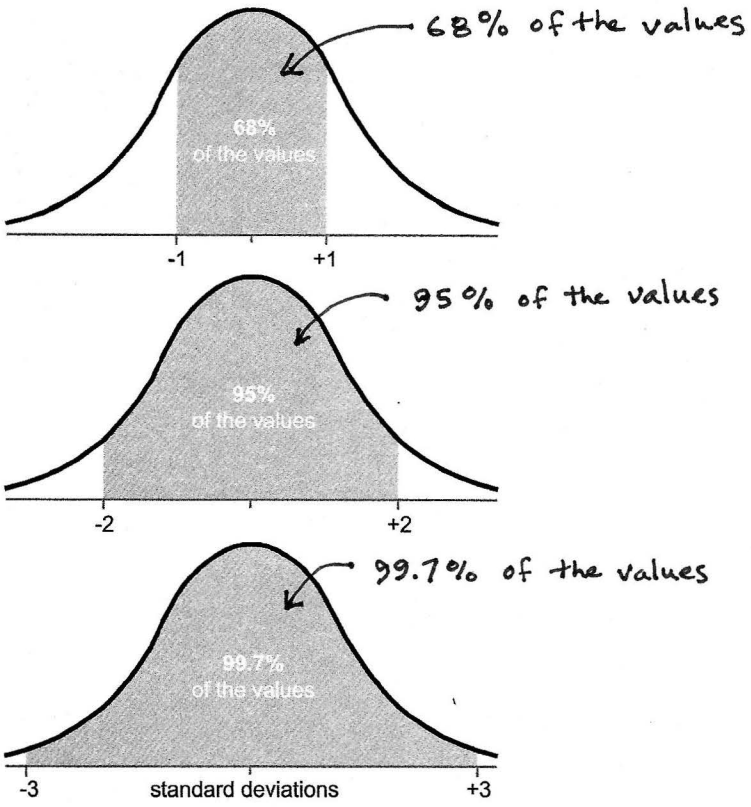


FIGURE 2.17 This figure shows a normal distribution of values in relation to the standard deviations of those values. The percentages of values that fall within one, two, and three standard deviations from the mean can be predicted with a normal distribution, and consequently can be predicted to a fair 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're dealing with a distribution that's close to normal, you automatically know that one standard deviation from the mean represents approximately 68% of the values, 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 variation among two or more sets of values. The bigger the standard deviation, the broader the range of values, and thus the greater difference in variation between them.

How does this relate to your world and the types of phenomena that you examine and present? Take a couple of minutes to list a few examples that are good candidates for measures of variation. In what situations would specific degrees of variation indicate something important to your organization?

Here are a few examples that I've encountered:

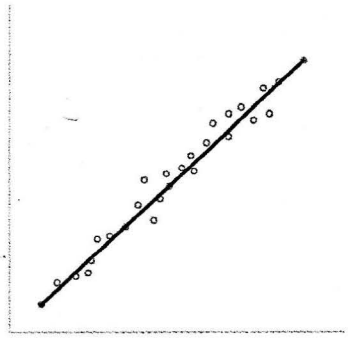
- Variation in the selling price of specific products or services. Is variation greater in some parts of the world or for some sales representatives? Do differences in variation correspond to increased or decreased profits?

- Different degrees of variation in measures of performance, such as the time it takes to manufacture products, answer phone calls, or resolve technical problems. Do instances of greater variation indicate problems in training, employee morale, process design, or systems? What does a greater degree of variation today compared to the past signify?
- Variation in employee compensation. Why is there such a discrepancy in compensation for the same job in different departments? Does this broad variation in salaries have an effect on employee morale or performance?
- Variation in the cost of goods purchased from different vendors. Why is variation in costs associated with some vendors so much more than others for the same goods?
- Variation in departmental expenses. How is it that some departments manage to keep their expenses so much lower than other departments?

I could go on, but I think the point is clear. Measures of variation tell important stories, so knowing ways to summarize and concisely communicate these stories 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 quantitative values are compared to one another to see if they correspond (i.e., co-relate) in some manner. For instance, does tenure on the job relate to productivity? In this section we're 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 the following questions:



A variable is something that can have multiple values, such as employee productivity.

- 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 quantitative values.

By "two paired sets of quantitative values" I mean the two sets of values that are considered when you examine the relationship of one measurable thing (a.k.a., a variable) 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 items manufactured per hour). In this case, the number of years on the job and items manufactured per hour for all employees 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 also increases, or that as tenure increases, productivity actually decreases. A linear correlation is limited, however, in that it cannot describe a relationship that is inconsistent, for example, if productivity increases along with tenure to a point but after that point it decreases as tenure continues to increase. This is still a relationship, but it's 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 the value of 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'd have to work through several steps. Very few of us need to do this because software or calculators 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 values fall between +1 and -1.
- A value of 0 indicates that there is no linear correlation.
- A value of +1 indicates that there is a perfect positive linear correlation.
- A value of -1 indicates that there is a perfect negative linear correlation.
- The greater the value, either positive or negative, the stronger the linear correlation.

It's getting clearer, 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 specifically designed to display the correlation of two paired sets of quantitative values. Perhaps you've seen this type of graph listed as one that's available in software but have never used it, or perhaps you 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 and 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 Y axis in relation to what happens to the value along the X axis. As X goes up, what happens to Y? As X goes down, what happens to Y? Is the relationship strong (i.e., it's close to a straight line) or is it weak (i.e., it bounces around)? Is it positive (i.e., it moves upward from left to right) or is it negative (i.e., it 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.

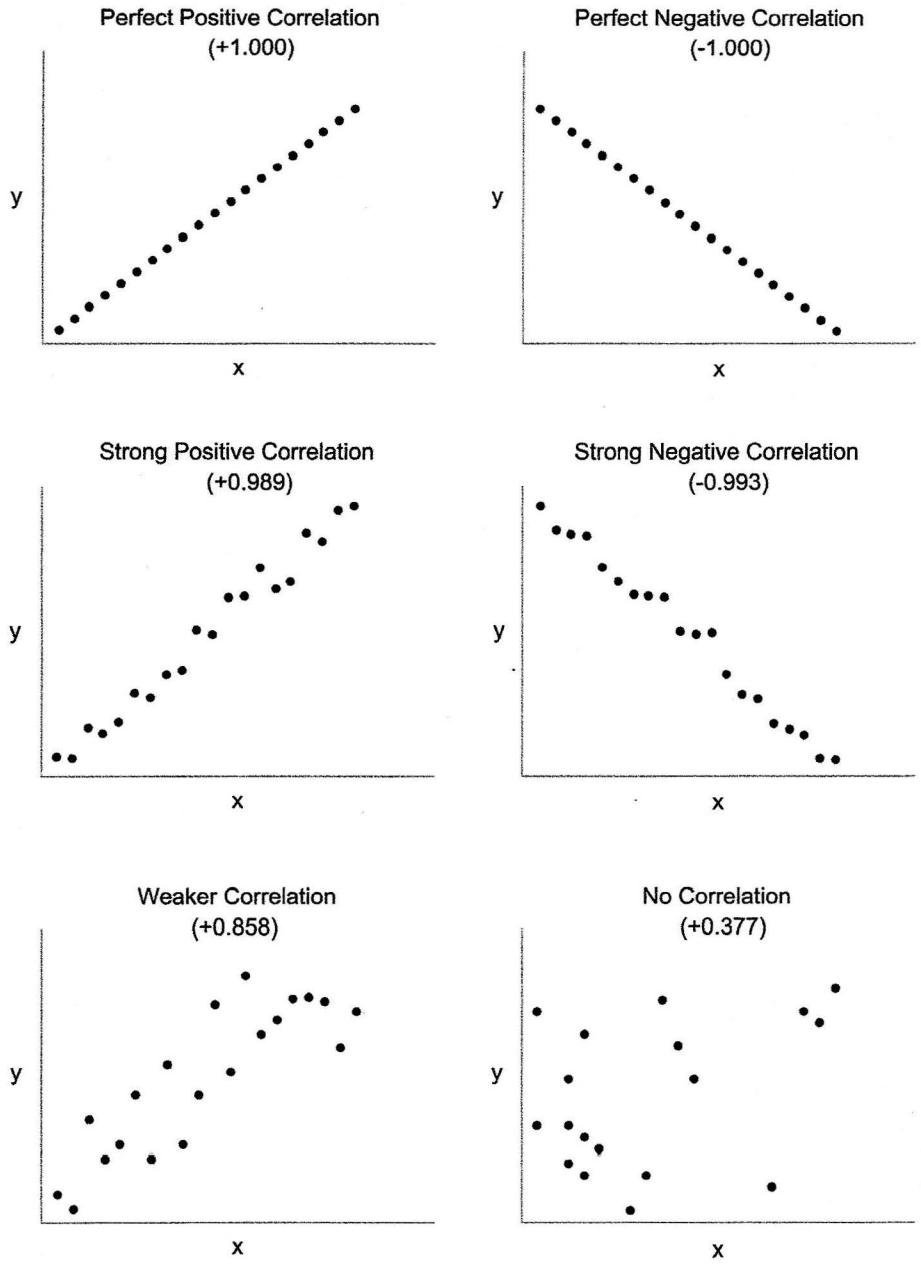


FIGURE 2.18 This is a series of scatter plots, each displaying a different relationship between two sets of paired values (e.g., employee tenure and productivity).

One way of looking at linear 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:

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

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

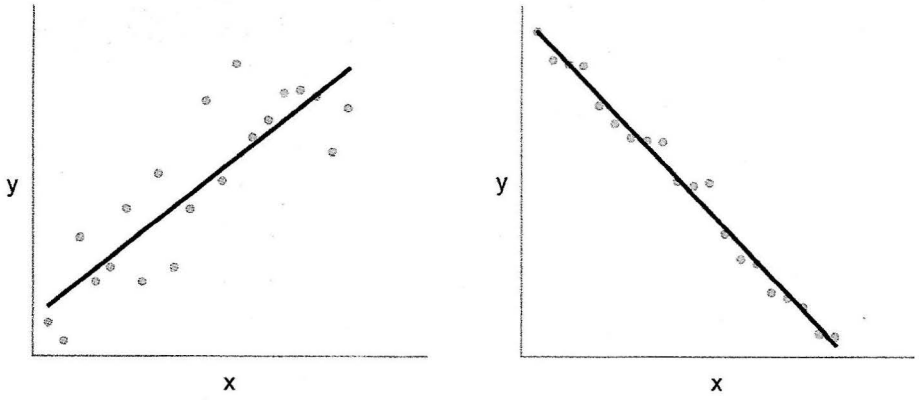


FIGURE 2.19 These are scatter plots with lines of best fit through the center of the dots to clearly delineate the nature of the relationship.

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 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 particularly 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 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 (2009) *Elementary Statistics*, Eleventh Edition. 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 roughly in a straight line—but not relationships that are positive under some circumstances and negative under others. Here's such an example:

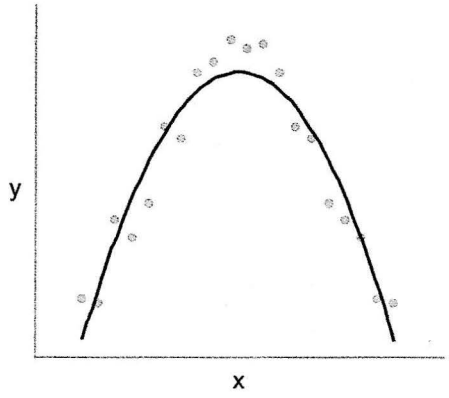


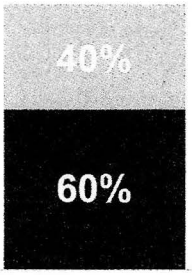
FIGURE 2.20 This is an example of a nonlinear correlation.

What you see here is definitely a correlation of sorts, but it certainly isn't linear (can't be described by a straight line). 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 them 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 customers 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:

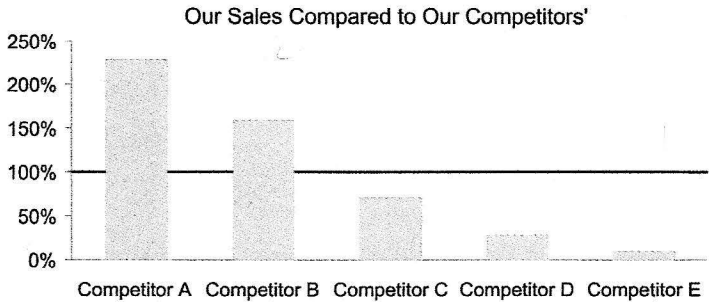
- As a sentence, such as "Two out of every five customers who access our website place an order."
- As a fraction, such as 2/5 (i.e., 2 divided by 5)
- As a rate, such as 0.4 (i.e., the result of the division expressed by the fraction above)
- As a percentage, such as 40% (i.e., the rate above multiplied 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 present quantitative information related to your own work.

Ratios are simple shorthand for expressing the direct relationship between two quantitative 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 1 (expressed as a rate) or 100% (expressed as a percentage); then, express the other values as ratios that fall above or below that point of reference. Here's an example expressed in percentages:



(Competitor sales are displayed as gray bars representing percentages compared to our company's sales, which appear as the gray line at 100%.)

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

Using 100% as a consistent point of reference, it is easy to see that the main competitor's sales are about 230% of your company's, or 2.3 times greater when expressed as a rate. Expressing comparisons in this manner eliminates the need for readers to do calculations in their heads when they care about relative differences.

Measures of Money

Much of the quantitative information that's important in business involves some currency of exchange—in other words, money. Be it U.S., Canadian, or Australian dollars, Japanese yen, British pounds, Swiss francs, or Euros, 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 2010. If you've been asked to prepare a report that exhibits the trend of sales in U.S. dollars for the past 10 years, would you be

justified in asserting that sales have increased by 100% during that time if 10 years ago annual sales were $100 million and today they total $200 million? That assertion would be true only if the value of a dollar today is the same as it was 10 years ago, which it isn't.

When the value of a dollar decreases over time, we refer to this as *inflation*. We can accurately compare money over time only if we adjust for inflation. I've noticed, however, that this is rarely done. Despite the validity of the case for adjusting for inflation, doing so isn't always practical, so I won't try to force on you a practice that you might very well ignore. For those of you who can take extra time required to correct for results skewed by inflation, I've included Appendix C, *Adjusting for Inflation*, in the back of the book. Adjusting for inflation isn't difficult, and doing so will improve the quality of your financial reporting.

Business today, especially in large companies, is often international and involves multiple currencies. This is a problem when we must produce reports that combine data in 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 or compare them, you must convert them into a single currency. Fortunately, most software systems that we use 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.

Understanding the relationships we've examined in this chapter lays a 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.

Summary at a Glance

Quantitative Relationships

- Quantitative stories include two types of values:
  - Quantitative
  - Categorical
- Quantitative stories always feature relationships.
- These relationships involve either:
  - Simple associations between quantitative values and categorical items or
  - More complex associations among multiple sets of quantitative values.

- There are four types of relationships between categorical items:
  - 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
Variation	Spread	Simple to calculate, relying entirely on the highest and lowest values, but only roughly defines a range 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 linear 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.