SHOW ME THE NUMBERS

Designing Tables and Graphs to Enlighten

SECOND EDITION

STEPHEN FEW

Analytics Press
BURLINGAME, CALIFORNIA

Analytics Press

PO Box 1545
Burlingame, CA 94011
SAN 253-5602
www.analyticspress.com
Email: info@analyticspress.com

PUBLISHER: Jonathan G. Koomey

COPY EDITOR: Nan Wishner
COMPOSITION: Bryan Pierce
COVER DESIGN: Stephen Few
PHOTOGRAPHY: John Fernez
PRINTER AND BINDER: C&C Offset Printing Company

ISBN-10: 0-9706019-7-2
ISBN-13: 978-0-9706019-7-1

This book was printed on acid-free paper in China.

10 9 8 7 6 5

CONTENTS

PREFACE TO SECOND EDITION

In 1914, ninety years before the first edition of this book was published, Willard C. Brinton wrote what was perhaps the first book about graphical data presentation, entitled *Graphic Methods for Presenting Facts*. The relatively few books on the topic that have been published since Brinton's have mostly appeared during the last 20 years. If you read Brinton's trailblazing book, you will be surprised by how little has changed since he wrote it. The problems that he tried to solve were not so different from those that we face today. Brinton began his book with the following words:

> *After a person has collected data and studied a proposition with great care so that his own mind is made up as to the best solution for the problem, he is apt to feel that his work is about completed. Usually, however, when his own mind is made up, his task is only half done. The larger and more difficult part of the work is to convince the minds of others that the proposed solution is the best one—that all the recommendations are really necessary. Time after time it happens that some ignorant or presumptuous member of a committee or a board of directors will upset the carefully-thought-out plan of a man who knows the facts readily enough to overcome the opposition. It is often with impotent exasperation that a person having the knowledge sees some fallacious conclusion accepted, or some wrong policy adopted, just because known facts cannot be marshaled and presented in such manner as to be effective.*
>
> *Millions of dollars yearly are spent in the collection of data, with the fond expectation that the data will automatically cause the correction of the conditions studied. Though accurate data and real facts are valuable, when it comes to getting results the manner of presentation is ordinarily more important than the facts themselves. The foundation of an edifice is of vast importance. Still, it is not the foundation but the structure built upon the foundation which gives the result for which the whole work was planned. As the cathedral is to its foundation so is an effective presentation of facts to the data.*[1]

1. Willard C. Brinton (1914) *Graphic Methods For Presenting Facts*. The Engineering Magazine Company, pages 1 and 2.

Accumulating information in and of itself is not useful. Information can't possibly serve a purpose until we first identify what's meaningful and then manage to make sense of it. Even once we understand the information, it remains inert until we actually do something with it. The true promise of the information age isn't tons of data but decisions and actions that are better because they're based on an understanding of what's really going on in the world. Any knowledge that you gain that could be used to make better decisions will amount to nothing if you can't communicate it to others in a way that makes sense to them. The ability to find what's useful in the mounds of data that surround us, to make sense of it, and to then present it clearly and accurately, forms the foundation on which the information age will finally fulfill its promise. Unless we give information a clear voice, its important stories will remain unheard, and ignorance will prevail.

As I write these words, more than 10 years into the 21st century, I still feel compelled to make the same essential case that Brinton made long ago when he wrote:

> *If an editor should print bad English he would lose his position. Many editors are using and printing bad methods of graphic presentation, but they hold their jobs just the same. The trouble at present is that there are no standards by which graphic presentations can be prepared in accordance with definite rules so that their interpretation by the reader may be both rapid and accurate. It is certain that there will evolve for methods of graphic presentation a few useful and definite rules which will correspond with the rules of grammar for the spoken and written language. The rules of grammar for the English language are numerous as well as complex, and there are about as many exceptions as there are rules. Yet we all try to follow the rules in spite of their intricacies. The principles for a grammar of graphic presentation are so simple that a remarkably small number of rules would be sufficient to give a universal language.*[2]

2. *ibid.*, page 3.

Even though no precedent for codifying the rules of graphic presentation existed in his day, Brinton made a bold and brilliant attempt to begin this work. Nearly a century has now passed, and in that time I and others have continued the work, resulting in the simple set of rules that Brinton hoped for, but these rules are known by relatively few who need them. In Brinton's time, the computer was still many years in the future. Graphics were produced by hand, usually by professional draftsmen. It is sad that since the advent of the computer—especially personal computing, which gave everyone the means to produce graphs and to do so efficiently—the quality of graphical communication has actually diminished. Having the means to create graphs with a computer doesn't guarantee that we'll do it effectively any more than having word processing software makes us great writers. It seems that we've lost sight of this distinction and assume that if we know how to use software that was designed to produce graphs, we have all that we need. Relying on software to do this for us results in failure. Software can do little to help us communicate graphically if we don't already possess the basic skills to do it ourselves.

You might be wondering why I'm writing a new edition of this book only eight years after the first edition was published. Has that much changed? Although relatively few changes have taken place in data visualization, I've learned a great deal more than I knew eight years ago. Most of the changes that I've made to this new edition are the results of my own professional growth.

Show Me the Numbers was the first of three books that I've now written. When I wrote it originally I had only recently begun focusing on data visualization despite having worked for nearly 20 years in information technology, mostly in *business intelligence*. I devoured the work of Edward Tufte and others, compiled the best of it, ran it through the filter of my own experience, then organized and expressed it to address the practical needs of people like you. Since writing the first edition of this book, I have taught data visualization courses internationally to thousands of people, written scores of articles, white papers, and two more books, and have worked with and advised many diverse organizations. As a result, my expertise has matured. This second edition of *Show Me the Numbers* is the result of this maturity.

In addition to my professional development, a few things have been happening in the world that affect graphical communication—some positive, which I describe in this edition, and some negative, which I warn against. On the positive side, two people in particular have shown that important stories involving numbers can be told in compelling ways. Even though Al Gore did not invent data visualization (or the Internet), the compelling nature of the graphical displays in his film *An Inconvenient Truth* began to change the tide of opinion about global warming. Those graphical displays moved people thanks to expert assistance from Nancy Duarte and Duarte Design. Another person who has used graphics to capture the imaginations of many in recent years is Hans Rosling of GapMinder.org. When this Swedish professor took the stage at the 2006 Technology, Entertainment, and Design (TED) Conference and told a story about the relationship between fertility and life expectancy throughout the world from 1962 to the present using an animated bubble chart, a new era of quantitative storytelling began.

On the negative side, the availability of bad graphs has increased since the first edition of this book. This expansion of bad graphical presentation has been made possible by the Web and has been fueled by uninformed so-called experts and self-serving software vendors. During the past few years, the number of people who claim expertise in data visualization has increased dramatically, but unfortunately many of them do not exhibit best practices. This is especially true of many graphic artists whose data-based visualizations are greeted with fanfare even though they don't actually inform, or do so poorly. Flashy visual displays are engaging, but unless they invite people to think about data in meaningful ways that lead to understanding, they fail in their purpose.

Nothing has been more disappointing to me personally than the lack of improvement in the charting capabilities of the product that is used more than any other to produce graphs: Microsoft Excel. The 2007 and 2010 releases of Excel have added superficial sizzle to the product's graphs without addressing any of the fundamental charting problems that have existed for years. Excel still encourages people to produce bad graphs—in some respects more than in the past because it now offers even more dysfunctional choices. Nevertheless, because practically everyone in the world who produces graphs has a copy of Excel, I've made a point of featuring graphs in this book that can be created effectively with Excel if you know what you're doing. Fortunately, the necessary skills are quite simple and easy to learn.

The purpose of this book is to help people present quantitative information in the most informative way possible, using simple skills and tools that are readily available. I invite you to enjoy the journey of learning these essential skills.

Steve

1 INTRODUCTION

The use of tables and graphs to communicate quantitative information is common practice in organizations today, yet few of us have learned the design practices that make them effective. This introductory chapter prepares the way for a journey of discovery that will enable you to become an exception to this unfortunate norm.

"Show me the numbers" is an expression that I've heard from time to time, especially in the workplace. Numbers are central to our understanding of performance. They enable us to make informed decisions. The way we determine success or failure is almost always based primarily on numbers. We derive great value from the stories that numbers tell, yet we rarely consider the significance of how we present them.

Contrary to popular wisdom, numbers cannot always speak for themselves. Inattention to the display of quantitative data results in large but hidden costs to most organizations. Time is wasted struggling to understand the meaning and significance of numbers—time that could be better spent doing something about them. What a shame, especially when this could be easily remedied. To provide a practical solution to this pervasive problem is the goal of this book.

Leveraging quantitative information—numbers—takes us out of the realm of assumption, feeling, guesswork, gut instinct, intuition, and bias, and into the realm of reliable fact based on measurable evidence. Too many decisions are based on perceptions that are fallible. You may wake up in the morning, step outside, feel the sunshine on your skin, and know deep down in your bones that it's warmer today than it was yesterday, only to glance at the thermometer and discover that it is in fact five degrees cooler. More to the point of this book, your gut may tell you that your organization is now doing better than ever, but a careful check of the numbers could reveal that performance has actually degraded during the past 12 months.

I respect the value of well-honed intuition. I personally trust a strong intuitive sense of what's going on, of what will work and what won't, that is rarely wrong when applied to my areas of expertise. Rarely wrong, but wrong often enough to keep me humble. By trusting my gut—overly confident that the extra step of actually checking the numbers would be wasted—I have at times been dragged kicking and screaming to the embarrassing admission that I was wrong. Numbers have an important story to tell, and it is up to us to help them tell it. Tables and graphs are usually the best means to communicate quantitative information. They are so common many of us assume that knowledge of their effective use is common as well. I assure you, it is not. Evidence of this fact in the form of countless poorly designed tables and graphs is visible all around us. According to Karen A. Schriver, an expert in document design, "Poor documents are so commonplace that deciphering bad writing and bad visual design have become part of the coping skills needed to navigate in the so-called information age."[1]

1. Karen A. Schriver (1997) *Dynamics in Document Design.* John Wiley & Sons, Inc., page xxiii.

The reason for this sad state of affairs is simple: few of us have ever been trained to design tables and graphs effectively. Some of us have struggled to do this work well but haven't found useful resources to assist us. Others of us haven't struggled at all, simply because we haven't seen enough examples of good design to recognize the inadequacy of our own efforts.

Before saying anything more, let me illustrate what I mean. Take a moment to examine the graph below.

FIGURE 1.1 This is a typical example of a poorly designed business graph. Notice the attempt at artistic flair in the use of color, 3D, and shading of the vertical bars.

This graph has visual impact—it's dramatic, it's colorful, it jumps off the page and demands attention—but to what end? What is the message?

Let's evaluate its effectiveness in light of the objective. Assume that its purpose is to inform a corporation's executives every Monday morning about the current state of quarterly sales, split into three geographical regions: Americas, Europe, and Asia. Given this intention, look at it again. What message could an executive get from this graph? Put yourself in the executive's position. Take a minute or two to examine the graph and interpret its message in this light.

. . . .

What did you get? Probably no more than the following:

- Sales for the Americas are better than sales for Europe, which in turn are better than sales for Asia. This much is clear.
- Sales for Asia don't amount to very much compared to the other regions.
- Sales for the Americas are more than 400,000, or thereabouts. Sales for Europe are somewhere around 200,000. Sales for Asia appear to be around 50,000, perhaps a little less.

That's not much information, and you certainly had to work for it, didn't you? Here are the actual values that were used to create this graph:

- Americas = $469,384
- Europe = $273,854
- Asia = $34,847

As an executive, you may not need to know precise sales amounts, but you would probably want greater accuracy than you could discern from this graph, and you would certainly want to get it faster and with less effort.

Several pieces of critical information aren't supplied by this graph:

- Given the fact that these sales are international, what is the unit of measure? Is it U.S. dollars, Euros, etc.?
- What is the date of the "quarter-to-date sales?" If you filed this report away for future reference and pulled it out again a year from now, you wouldn't know what day, what quarter, or even what year it represents.
- How do these sales figures compare to your plan for the quarter?
- How do these sales figures compare to how you did at this time last quarter or this same quarter last year?
- Are sales getting better or worse?

This graph lacks important contextual information and critical points of comparison. As a report of quarter-to-date sales across your major geographical regions, it doesn't show you very much. It uses a great deal of ink to say very little.

Given the intended message and the information an executive might find useful, the following table tells the story better:

2011 Q1-to-Date Regional Sales
As of March 15, 2011

	Current			Qtr End	
	Sales (U.S. $)	Percent of Total Sales	Percent of Qtr Plan	Projected Sales (U.S. $)	Projected Percent of Qtr Plan
Americas	469,384	60%	85%	586,730	107%
Europe	273,854	35%	91%	353,272	118%
Asia	34,847	5%	50%	43,210	62%
	$778,085	100%	85%	$983,212	107%

Note: To date, 82% of the quarter has elapsed.

FIGURE 1.2 This table contains all of the information that was contained in the graph in *Figure 1.1*, plus much more, in an easy-to-grasp format.

This table is simple and easy to read, yet it contains a great deal more information than the previous, colorful graph. No ink is wasted. As an executive, you might actually be able to use this report to make important decisions. It communicates.

Here's some information that I found on the Web that was featured by a PBS television program called NOW:

Favorable or Unfavorable View of the U.S.	
Brazil: % with somewhat or very favorable opinion of the U.S.:	52%
Brazil: % with somewhat or very unfavorable opinion of the U.S.:	32%
Mexico: % with somewhat or very favorable opinion of the U.S.:	64%
Mexico: % with somewhat or very unfavorable opinion of the U.S.:	25%
Britain: % with somewhat or very favorable opinion of the U.S.:	75%
Britain: % with somewhat or very unfavorable opinion of the U.S.:	16%
Germany: % with somewhat or very favorable opinion of the U.S.:	61%
Germany: % with somewhat or very unfavorable opinion of the U.S.:	35%
Russia: % with somewhat or very favorable opinion of the U.S.:	61%
Russia: % with somewhat or very unfavorable opinion of the U.S.:	33%
Poland: % with somewhat or very favorable opinion of the U.S.:	79 %
Poland: % with somewhat or very unfavorable opinion of the U.S.:	11%
South Africa: % with somewhat or very favorable opinion of the U.S.:	65%
South Africa: % with somewhat or very unfavorable opinion of the U.S.:	28%
Kenya: % with somewhat or very favorable opinion of the U.S.:	80%
Kenya: % with somewhat or very unfavorable opinion of the U.S.:	15%
India: % with somewhat or very favorable opinion of the U.S.:	54%
India: % with somewhat or very unfavorable opinion of the U.S.:	27%
Japan: % with somewhat or very favorable opinion of the U.S.:	72%
Japan: % with somewhat or very unfavorable opinion of the U.S.:	26%
South Korea: % with somewhat or very favorable opinion of the U.S.:	53%
South Korea: % with somewhat or very unfavorable opinion of the U.S.:	44%
Egypt: % with somewhat or very favorable opinion of the U.S.:	6%
Egypt: % with somewhat or very unfavorable opinion of the U.S.:	69%
Pakistan: % with somewhat or very favorable opinion of the U.S.:	10%
Pakistan: % with somewhat or very unfavorable opinion of the U.S.:	69%
Turkey: % with somewhat or very favorable opinion of the U.S.:	30%
Turkey: % with somewhat or very unfavorable opinion of the U.S.:	55%
Jordan: % with somewhat or very favorable opinion of the U.S.:	25%
Jordan: % with somewhat or very unfavorable opinion of the U.S.:	75%

FIGURE 1.3 This table reports information that was collected by the Pew Research Center, as presented by the PBS television program NOW.

When I first looked at this information not long after the horrific events of September 11, 2001, I was frustrated. I wanted to understand how these countries felt about America, but I couldn't get a clear picture from this table no matter how long I struggled with it. To solve the problem, I loaded the data into Excel and created the following series of graphs:

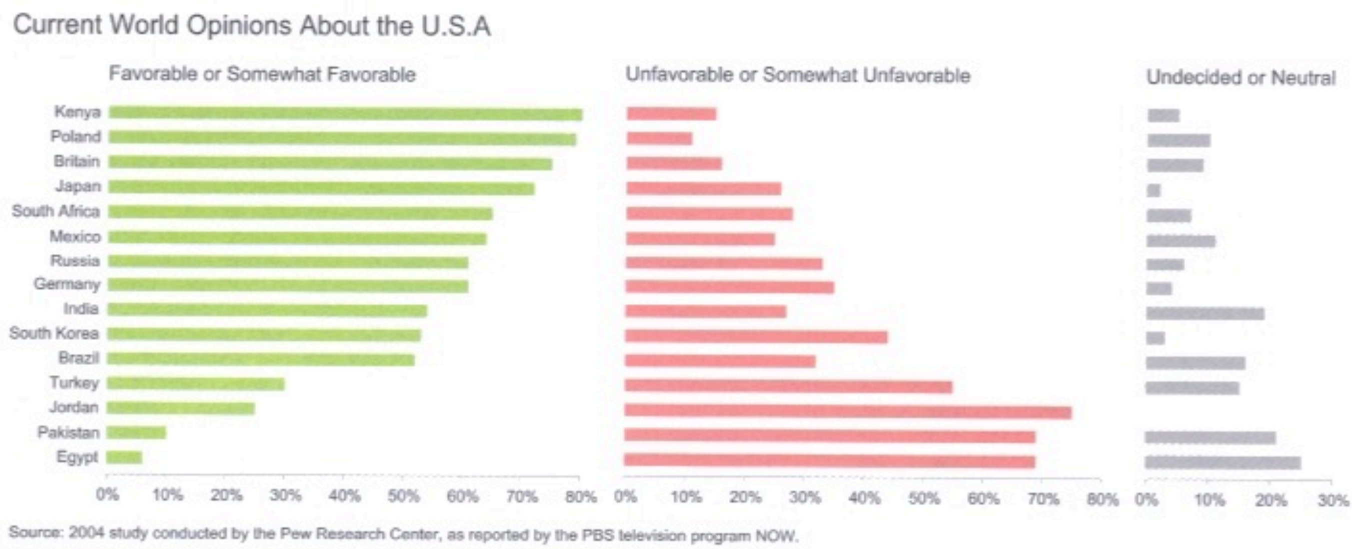


FIGURE 1.4 This is a graphical display of the information that appears in the form of a table in *Figure 1.3*.

With this graphical display of the same information, the story suddenly came alive. It gave me the overview that I needed, plus some details that I never expected. For example, take a look at the country of Jordan, which is third up from the bottom. In addition to the fact that relatively few people in Jordan held favorable opinions of the U.S.A. (the green bar) and a greater percent than in any other country held unfavorable opinions (the red bar), notice that nobody in Jordan who was polled at that time was without an opinion, which pops out as an empty space where a gray bar is expected.

Here's another fairly typical example of a business graph that suffers from severe design problems:

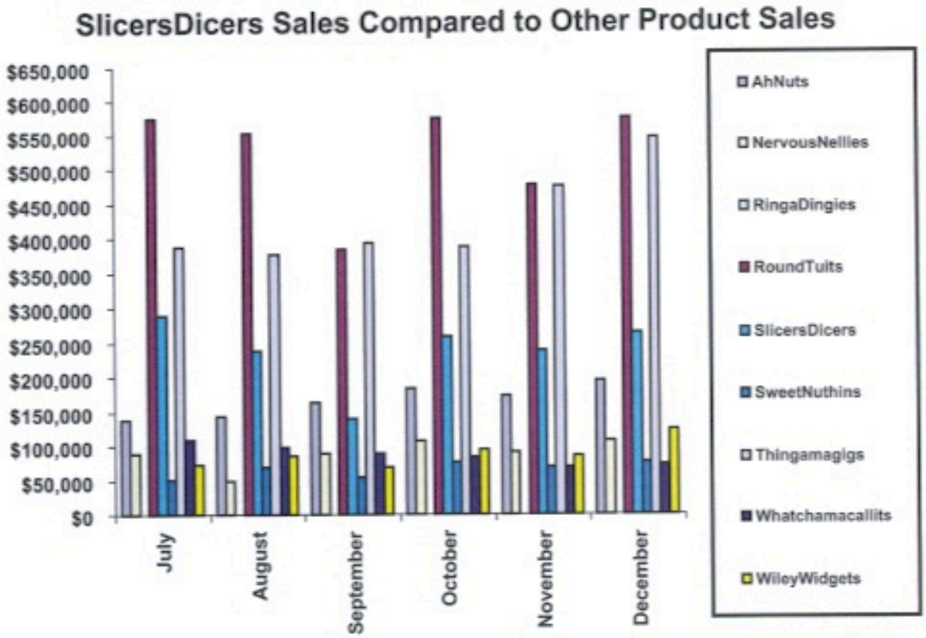


FIGURE 1.5 This is a typical example of a poorly designed vertical bar graph.

Without the graph's title, would you have any idea that its purpose is to compare the sales performance of the product named SlicersDicers to the performance of each of the other products? Designers speak of objects as having *affordances*—characteristics that reveal how they're supposed to be used and make them easy to use in those ways. A teapot has a handle. A door that you need to push has a push-plate. The design of an object should, in and of itself, suggest how the object should be used. This graph relies entirely on its title to declare its purpose. The design not only fails to suggest the graph's use, it actually subverts the user's ability to determine what the graph is for.

For a thorough examination of *affordances* in the broader context of design for usability, see Donald A. Norman's classic text, *The Design of Everyday Things* (1988) Basic Books.

Imagine for a few minutes that it is your job to assist the manager responsible for the products that appear in this graph. You've been asked to create a simple means to help her compare the performance of SlicersDicers to the performance of the other products in her portfolio. For her immediate purposes, she isn't particularly interested in actual dollars. She just wants to know how recent sales of SlicersDicers compare to sales of her other products. To begin your task, examine the graph again; then, make a list of the ways its design fails to support the product manager's needs. Take a minute to write your list next to the graph in the right margin of the book.

. . . .

Now, advance this exercise to the next level. Take a few more minutes to list the features you would incorporate into a new version of the graph to better serve the product manager's needs. Make use of your prior list to suggest areas that should be addressed.

. . . .

How did you do? Did you find yourself rising to the challenge? If so, you are already beginning to think critically and creatively about the design of tables and graphs. There is no single correct solution to the task at hand, but let me show one possible solution, which previews a design technique that may interest you.

FIGURE 1.6 This is a series of related graphs, each designed to compare the sales of the SlicersDicers product to sales of a different product.

This solution uses a divide-and-conquer approach to the problem. With eight small graphs—one for each product that needs to be compared to SlicersDicers—the complications that resulted from squeezing everything into a single graph have been reduced. The black reference line that represents SlicersDicers' sales in each of the graphs serves as a powerful affordance, strongly suggesting that this display should be used to compare the sales performance of each product to SlicersDicers.

Let's look at one last example, for now, of ineffective graph design. Everyone is familiar with the ubiquitous pie chart. This one is quite typical in its design:

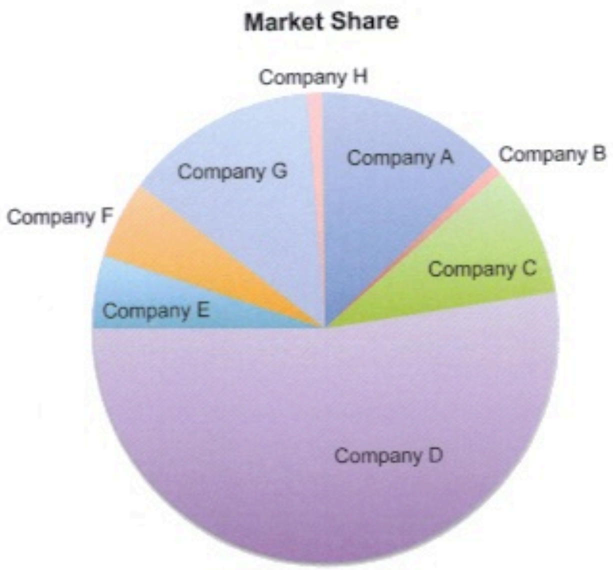


FIGURE 1.7 This is an example of an ineffective style of graph: a pie chart.

Here's the same graph below, but this time it's dressed up with the addition of a 3-D effect:

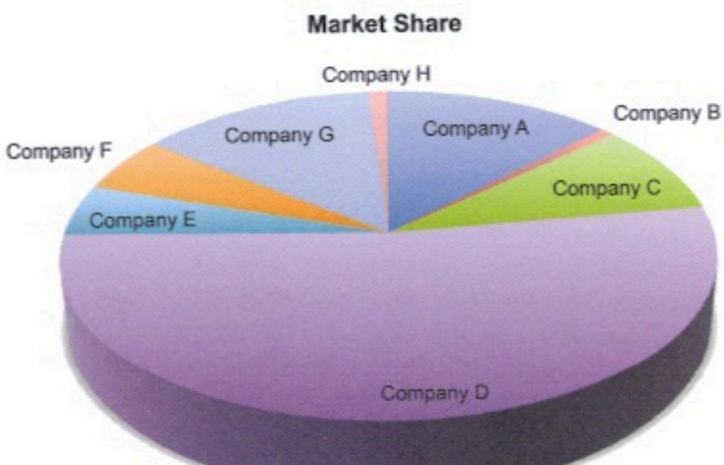


FIGURE 1.8 This is a slight variation of *Figure 1.7*, which adds 3-D perspective. This design is even less effective.

Did the use of 3D enhance the display? These two graphs are meant to display the same market share numbers, yet notice how the addition of 3-D perspective affects your perception of the data. The intention of these graphs is for

Company G to compare its market share to the shares of its competitors. Can you determine Company G's market share from either of these graphs? Can you determine its rank compared to its competitors? Which has the greater share: Company A or Company G? Because of fundamental limitations in visual perception, you really can't answer any of these questions accurately.

Now look at the exact same market share values displayed more effectively:

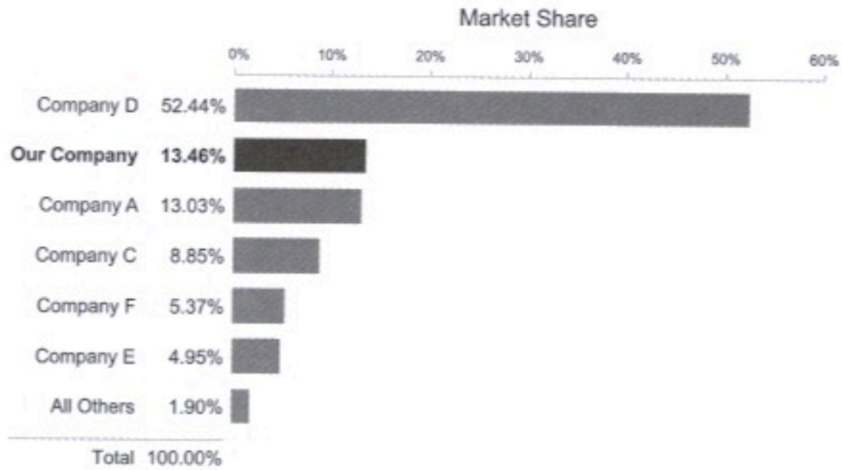


FIGURE 1.9 This horizontal bar graph displays the market share data in a way that is easy to interpret and compare.

Did you have any trouble interpreting this information? Did you struggle to find the most important information? It's obvious that Our Company ranks second, slightly better than Company A, and that our market share is precisely 13.46%. This display contains no distractions. It gives the numbers clear voices to tell their story.

Purpose

Although they are routinely used today, with few exceptions graphs have only been used to display quantitative information for a little more than two centuries. Even though tables have been around longer, it wasn't until the last quarter of the 20th century that the use of either became widespread. What caused this rise in their use? The personal computer.

Shortly after the advent of the PC, I began developing and teaching courses in the use of some of the earliest PC-based business software, including Lotus 1-2-3. Although Lotus wasn't the first electronic spreadsheet product, and it's no longer the most popular, it legitimized the PC as a viable tool for business. Prior to the advent of spreadsheet software, tables of quantitative information were generally produced using pencils, sheets of lined paper, a calculator, and hours of tedious labor. Graphs could only be produced using a pen (perhaps several, of different colors), a straight-edged device (e.g., a ruler or draftsman's triangle), a sheet of graph paper, and, once again, hours of meticulous labor. When chart-producing software hit the scene, many of us who would have never taken the time to draw a graph suddenly became Rembrandts of the X and Y axes , or so we thought. Like kids in a toy store, we went wild over all the available colors and cool effects, thrilled with the new means for techno-artistic expression. Through the *magic* of computers, making tables and graphs became easy—perhaps too easy.

Today, everyone can produce reports of quantitative information in the form of tables and graphs. Children in elementary school are taught the mechanics of using spreadsheet software. Something produced with a computer acquires an air of authenticity and quality that it doesn't necessarily deserve, however. In our excitement to produce what we could only do before with great effort, many of us have lost sight of the real purpose of quantitative displays: *to provide the reader with important, meaningful, and useful insight*. To communicate quantitative information effectively requires an understanding of the numbers and the ability to display their message for accurate and efficient interpretation by the reader.

The necessary knowledge and skills are well within your reach if you make a little effort. Once you've read this book and practiced a bit, you'll find that it is no more difficult and takes no longer to produce effective tables and graphs than it does to produce ineffective ones. By synthesizing the best practices of quantitative information design that have been learned through many years of research and real-world trial and error by trailblazers, I hope to make your effort relatively easy, and perhaps even enjoyable.

With few exceptions, the tables and graphs in this book were made using Microsoft Excel software, in part to demonstrate that good design can be achieved even if you have no software but Excel.

The purpose of quantitative tables and graphs is to communicate important information effectively. That's it. Not to entertain, not to indulge in self-expression, not to make numbers that you would otherwise find boring suddenly interesting through flash and dazzle. Edward Tufte, a well-known expert in this field, expresses this perspective quite simply: "The overwhelming fact of data graphics is that they stand or fall on their content, gracefully displayed . . . Above all else show the data." And to those who believe that they must dress up the numbers, he warns: "If the statistics are boring, then you've got the wrong numbers."[2]

2. Edward R. Tufte (2001) *The Visual Display of Quantitative Information*, Second Edition. Graphics Press. The first quotation appears on pages 121 and 92; the second on page 80.

As with any endeavor that's worthwhile, our goal in the design of quantitative information for the purpose of communication is *excellence*. Why would we aim for anything less when the means to attain excellence are readily available? Why anything less when doing good work is so satisfying, and simply getting by is such a drag? With the right knowledge and skills, we have a chance to make a difference. Even if you're not inspired by your organization's objectives, do it for yourself. Good work, or what Buddhists call "right livelihood," is one of fundamental delights of life. It is fulfilling in and of itself.

As presenters of quantitative information, it is our responsibility to do more than sift through the information and pass it on; we must help our readers gain the insight contained therein. We must design the message in a way that leads readers on a journey of discovery, making sure that what's important is clearly seen and understood. The right numbers have important stories to tell. They rely on us to find those stories, understand them, and then tell them to others in a way that is clear, accurate, and compelling.

Scope

One of the challenges in writing a book like this is to constrain its scope. The contents must constitute a coherent whole that addresses a specific set of needs for a particular audience. There is a risk of communicating too little by saying too much. This is indeed a challenge when your interests are far ranging, like mine. What to leave out? In light of this objective, I've limited this book to content that is 1) practical and applicable to typical organizational reporting, 2) focused on communication, and 3) focused on effective design.

Although some organizations employ professional statisticians to perform sophisticated quantitative analyses, most of us never touch advanced statistics and would be intimidated by anything much more complicated than a measure of average. Even though most organizations have a few specialized display techniques to fit their unique needs, display techniques that are not of general interest don't appear in this book. Rest assured that most or all of what you will find in this book will apply directly to your own work. You will not have to wade through information that might be interesting from an academic perspective but peripheral to the task at hand.

Tables and graphs of quantitative business information can be used for four purposes:

- Analyzing
- Communicating
- Monitoring
- Planning

When you use tables and graphs to discover the message in the data, you are performing analysis. When you use them to pass a message on to others, your purpose is communication. When you use them to track information about performance, such as the speed or quality of manufacturing, you are engaged in monitoring. When you use them to predict and prepare for the future, you are planning. All of these are important uses of tables and graphs, but the processes that you engage in and the design principles that you follow differ somewhat for each. My purpose in this book is to help you learn to design tables and graphs to communicate important information.

This book focuses on design. It is not about the mechanics of constructing tables and graphs using a particular software product (e.g., Microsoft Excel). It is not an encyclopedia of tables and graphs, listing and describing the countless variations that exist. It is not a book that will make you an expert in a particular type of graph (e.g., the nuances of histograms). It is not a book about making tables and graphs look pretty though there is certainly beauty in those that simply and elegantly hit their communication target spot on. It is a book about design, in particular about design practices that broadly apply to using tables and graphs for effective quantitative communication.

When do you use a table versus a graph? When do you use one type of graph rather than another? How do you highlight what's most important and make the message crystal clear? What should you avoid to ensure that you eliminate anything that might distract from your message? This book answers these questions and more, and the answers can be applied across the board to every table and graph you will ever need to create.

For instruction in the use of graphs for data exploration and analysis, see my book *Now You See It: Simple Visualization Techniques for Quantitative Analysis* (2009), and for instruction in the design of dashboards for performance monitoring, see my book *Information Dashboard Design: The Effective Visual Communication of Data* (2006).

For an excellent encyclopedia of charts, consider Robert L. Harris (1999) *Information Graphics: A Comprehensive Illustrative Reference*, Oxford University Press.

Intended Readers

Simply stated, this book is intended for anyone whose work involves the use of tables and graphs for presenting quantitative information. Those of us who hold this responsibility don't fit into a tidy collection of job titles. Our roles are spread across a broad spectrum. Some of us specialize in the production of analyses and reports, with job titles that often include the term analyst, for example: Financial Analyst, Business Analyst, Data Analyst, or Decision Support Analyst, to name a few. Some of us have managerial responsibilities and are occasionally required to prepare tables and graphs for other managers higher up in the ranks. Some of us are graphic artists, and someone decided that because the words graphs and graphic are related, responsibility for the production of quantitative graphs must belong to us. The rest of us are scattered over the organization. No matter what it says on your business card, if you are responsible for creating tables and graphs to communicate quantitative information, and you want to do it well, this book is for you.

Content Preview

Tables and graphs don't just display numbers; they present them in a manner that relates them to something, such as to time or to one another, to reveal a meaningful message in context. Tables and graphs are two members of a larger family of display methods known as *charts*. In addition to tables and graphs, there are other types of charts, such as diagrams, which illustrate a process or set of relationships, and maps, which depict information geospatially.

Although tables and graphs are both vehicles for presenting information visually, the role that visual perception plays in reading and interpreting the information presented differs significantly for tables versus graphs. Graphs are perceived almost entirely by our *visual* system, and, as such, employ a visual language of sorts. In a graph, lines, bars, and other objects positioned within a 2-D space formed by perpendicular axes are used to communicate visual patterns and relationships. To see patterns and relationships is a natural function of visual perception.

Tables, with their many columns and rows of text, interact primarily with our *verbal* system, which entails what we normally mean when we speak of language. As such, we process information contained in tables sequentially, reading down columns or across rows of numbers, comparing this number to that number, one pair at a time. This is different from the visual processing that occurs when we view graphs, which involves high-bandwidth, simultaneous input of more data, potentially enabling us to perceive a great deal of quantitative information in a burst of recognition. Neither method of quantitative display is inherently better than the other, but each is better than the other for particular communication tasks, and both play a vital role.

This book is designed to take you on a journey. We begin in Chapter 2, *Simple Statistics to Get You Started*, with an introduction to a few statistics that are easy to understand, incredibly handy, and particularly useful in tables and graphs.

We continue in Chapter 3, *Differing Roles of Tables and Graphs*, with an introduction to tables and graphs, including what's common to both, then proceed to which works best for particular purposes.

In Chapter 4, *Fundamental Variations of Tables*, we go on to break tables down into the types of information they can be used to display and into the basic ways they can be structured.

Before proceeding to a closer look at graphs, we take a brief detour in Chapter 5, *Visual Perception and Graphical Communication*, to learn how visual perception works, from the time when light enters our eyes to the time that the information that our brains have gleaned from that light is stored in memory for future reference or is immediately discarded. You may be tempted to skip this chapter. Don't give in. If you want to understand what works and what doesn't in the design of graphs so you can apply your knowledge to each new situation that arises, you will need to understand these basic concepts about visual perception.

In Chapter 6, *Fundamental Variations of Graphs*, we break graphs down into the types of information they can be used to display and explore the ways that data can be visually encoded to tell the story most effectively.

In Chapter 7, *General Design for Graphical Communication*, we apply what we've learned about visual design as general practices for both tables and graphs.

From there, we dive into the details of *Table Design* in Chapter 8. We examine the structural components of tables, how to combine them for optimal effect, and the all-too-common bad practices that you should avoid.

We then shift our focus in Chapter 9 to *General Graph Design*, beginning with design principles that apply to all types of graphs. In Chapter 10, *Component-Level Graph Design*, we look closely at each component of graphs to learn when and how to use them for effective communication. Chapter 11, *Displaying Many Variables at Once*, focuses on developing strategies for displaying complex messages. In Chapter 12, *Silly Graphs that Are Best Forsaken*, we learn to avoid graphs that don't present data effectively and what alternatives to use that do the job well.

In Chapter 13, *Telling Stories with Numbers*, we put the entire venture into perspective. Numbers are important because they have stories to tell that we must understand to make good decisions. We must learn to tell these stories in ways that are simple, clear, accurate, and compelling.

Finally, we end our journey in Chapter 14, *The Interplay of Standards and Innovation*, where I climb up on my soapbox and prophesy that you will suffer defeat if you don't establish and follow a practical set of standards for the design of tables and graphs.

Communication Style

At heart, I am a teacher. My mind works like a teacher's. When I tell you about something, I care that you get it. When you do, I feel good. When you don't, I feel that I've failed. Consequently, this book is designed as a learning experience, not simply to inform or entertain. It is not designed as a reference that you pull from the shelf occasionally. It is designed to get into your head in a way that is thorough and lasting.

As a result, this book is filled with examples that bring the material to life, as well as questions that invite you to think and perhaps even struggle a bit during the process. It contains exercises that provide an opportunity to practice what you're learning and to learn more thoroughly through that practice. It is laid out in a sequence that leads you through learning at a conceptual level, then allows you to apply your conceptual understanding to various real-world scenarios. I guide you on a journey of discovery, rather than presenting principles that you must memorize and follow based merely on my authority as an expert. I want you to learn these practices and make them a part of your work to the extent that you no longer need to think about them. If someone asks you why you design tables and graphs the way you do five years from now, I want you to still be able to explain it.

I love teaching, in part because I love learning. I try to approach each day as a student. Doing so enriches me and keeps life fresh and interesting. I invite you to approach this book with the curiosity of a student. If you do, your effort will be rewarded.

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.

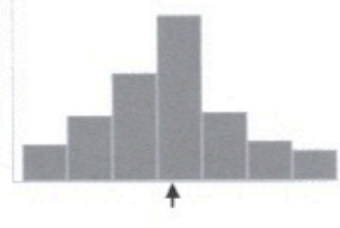
Numbers that Summarize

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 50^{th} 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.

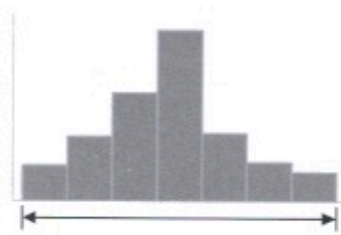
MODE AND MIDRANGE

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:

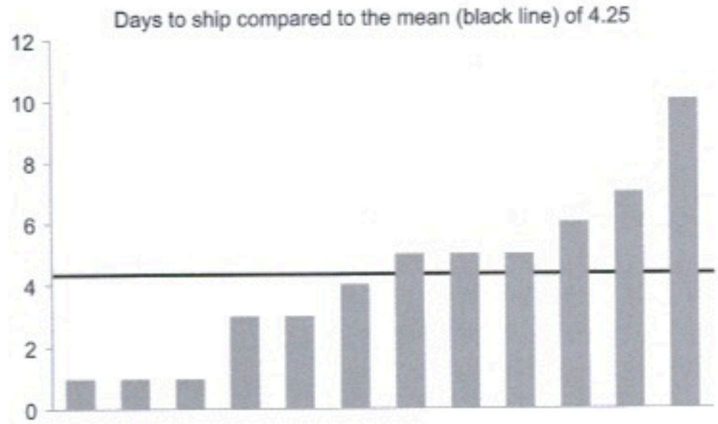


FIGURE 2.14 This graph shows a simple way to visualize the days it took Warehouse B to ship each of the 12 orders compared to the mean value of 4.25 days.

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:

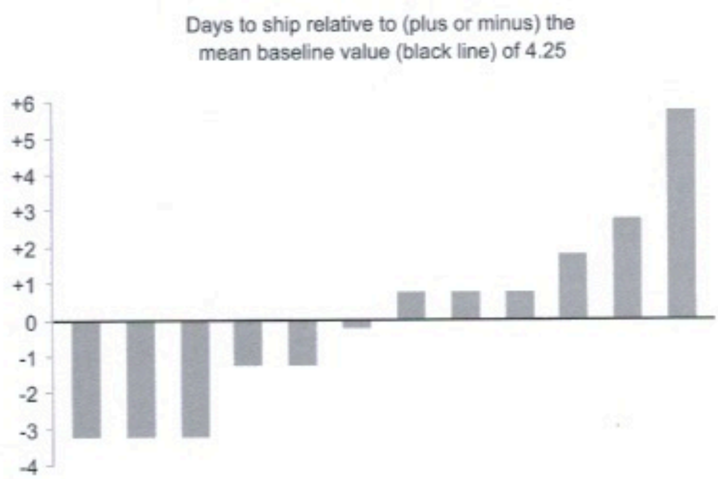


FIGURE 2.15 This graph displays the days it took Warehouse B to ship the individual orders relative to the mean.

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 2.70 days. We can compare this to the standard deviation for Warehouse A's shipments of 0.86 days. The difference between a standard deviation of 2.70 and one of 0.86 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.

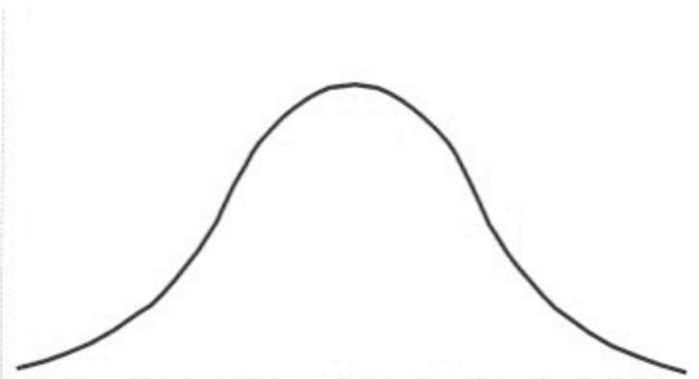


FIGURE 2.16 This curved line represents a normal distribution. It displays the frequency of values as they occur from the lowest value at the left to the highest value at the right. Most instances have values near the midpoint of the set of values, which represents the mean. In a perfect normal distribution,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

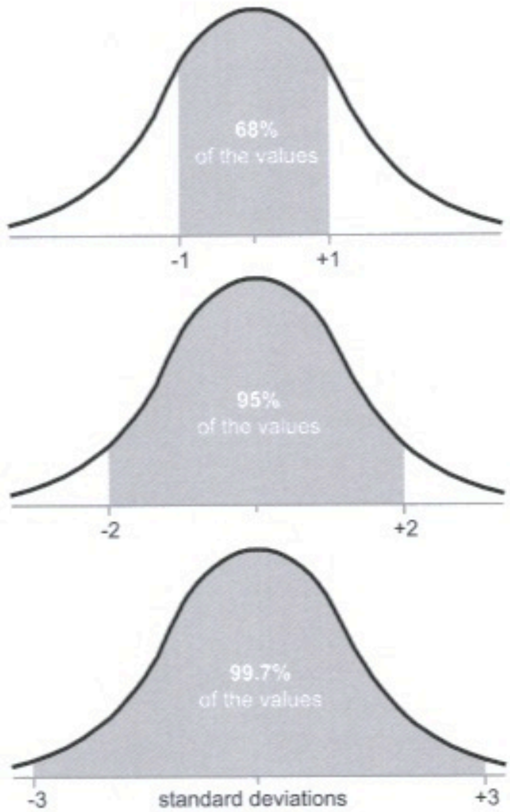


FIGURE 2.17 This figure shows a normal distribution of values in relation to the standard deviations of those values. The percentages of values that fall within one, two, and three standard deviations from the mean can be predicted with a normal distribution, and consequently can be predicted to a fair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

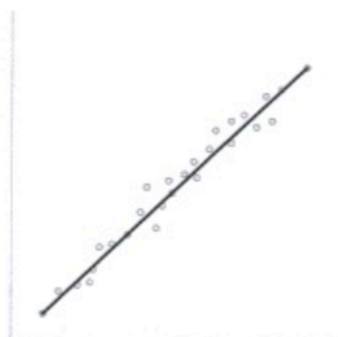
- Variation in the selling price of specific products or services. Is variation greater in some parts of the world or for some sales representatives? Do differences in variation correspond to increased or decreased profits?

- 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:



- 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

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

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.

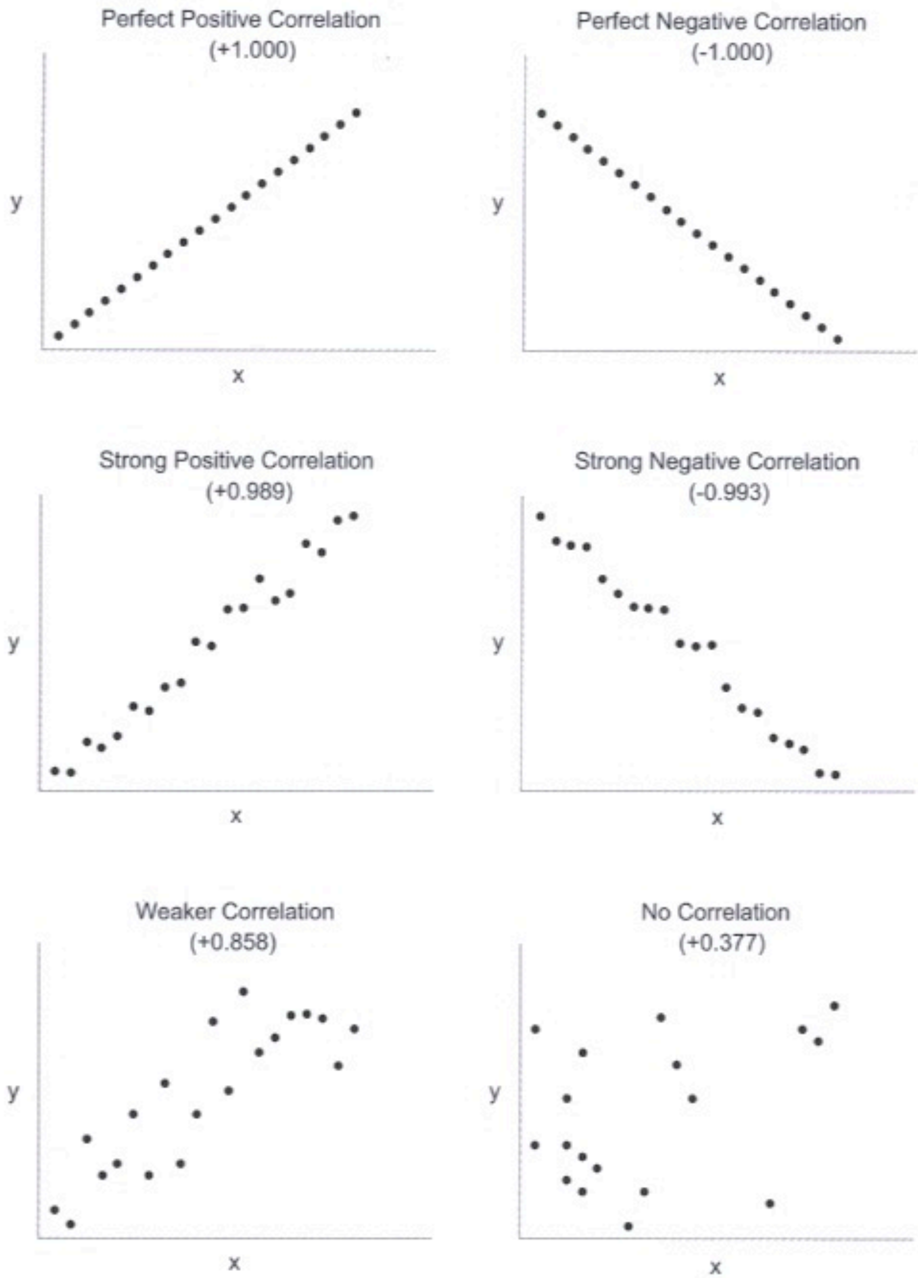


FIGURE 2.18 This is a series of scatter plots, each displaying a different relationship between two sets of paired values (e.g., employee tenure and productivity).

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.

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:

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*.

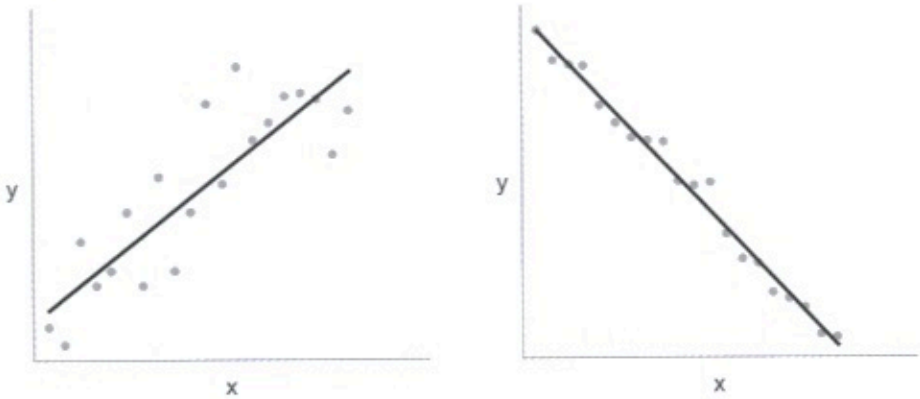


FIGURE 2.19 These are scatter plots with lines of best fit through the center of the dots to clearly delineate the nature of the relationship.

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:

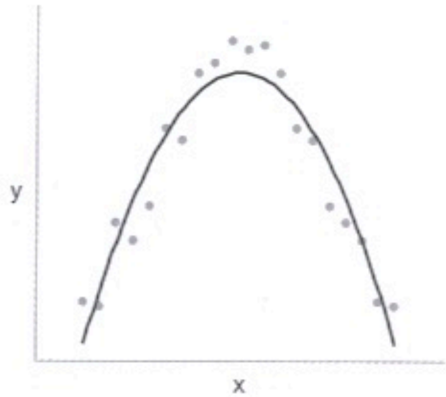


FIGURE 2.20 This is an example of a nonlinear correlation.

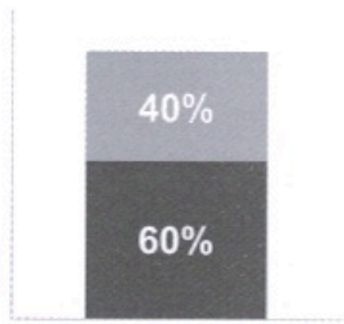
What you see here is definitely a correlation of sorts, but it certainly isn't linear (can't be described by a straight line). If this scatter plot represents the relationship between employee tenure (i.e., years on the job) on the X axis and employee productivity on the Y axis, how would you interpret 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:

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.

3 DIFFERING ROLES OF TABLES AND GRAPHS

Tables and graphs are the two fundamental vehicles for presenting quantitative information. They have developed over time to the point that we now thoroughly understand which works best for different circumstances and why. This chapter introduces tables and graphs and gives simple guidelines for selecting which to use for your particular purpose.

Tables and graphs are the two primary means to structure and communicate quantitative information. Both have been around for quite some time and have been researched extensively to hone their use to a fine edge of effectiveness. The best practices that have emerged are easy to learn, understand, and put to use in your everyday work with numbers.

Occasionally, the best way to display quantitative information is not in the form of a table or graph. When the quantitative information you want to convey consists only of a single number or two, written language is an effective means of communication; your message can be expressed simply as a sentence or highlighted as a bullet point. If your message is that last quarter's sales totaled $1,485,393 and exceeded the forecast by 16%, then it isn't necessary to structure the message as a table, and there is certainly no need to create a graph. You can simply say something like:

Q2 sales = $1,485,393, exceeding forecast by 16%

Alternatively, it wouldn't hurt to structure this message in simple tabular form such as this:

Q2 Sales	Compared to Forecast
$1,485,393	+16%

FIGURE 3.1 This table shows sales value information, arranged in columns.

or this:

Q2 Sales	$1,485,393
Compared to Forecast	+16%

FIGURE 3.2 This table shows sales value information, arranged in rows.

If you're like a lot of folks, however, you might be tempted while structuring this information as a table to jazz it up in a way that actually distracts from your simple, clear message—perhaps something like this:

FIGURE 3.3 This table shows sales values, designed to impress, or perhaps to entertain, but not primarily to communicate.

You might even be tempted to pad the report with an inch-thick stack of pages containing the details of every sales order received during the quarter, eager to demonstrate how hard you worked to produce those two sales numbers. However, as we'll observe many times in this book, there is *eloquence in simplicity*.

Quantities and Categories

Before we launch into an individual examination of tables and graphs, let's review an important fact, common to both, that quantitative messages are made up of two types of data:

- Quantitative
- Categorical

Quantitative values measure something (number of orders, amount of profit, rating of customer satisfaction, etc.). Categorical items (i.e., members of a category) identify what the quantitative values measure. These two types of data fulfill different roles in tables and graphs. In the following simple table, which displays exempt and non-exempt employee compensation by department, all the information consists either of quantities being measured (compensation) or items belonging to categories (sales, operations, and manufacturing are items in the department category) to which the quantities relate.

Department	Exempt	Non-Exempt
Sales	950,003	1,309,846
Operations	648,763	2,039,927
Manufacturing	568,543	2,367,303
Total	$2,167,309	$5,717,076

FIGURE 3.4 This is a table of employee compensation information that you can use to practice distinguishing quantitative and categorical data.

The labels that identify the various departments in the far left column, including Total, belong to the category Department, and the labels Exempt and Non-exempt belong to a separate category that we could call Employee Type. All of the other data (i.e., all the numbers in this table) are quantitative values.

Do numbers always represent quantitative values? No. Sometimes numbers are used to label things and have no quantitative meaning. Order numbers (e.g., 1003789), numbers that identify the year (e.g., 2011), and numbers that sequence information (item number 1, item number 2, item number 3, etc.), are examples of numbers that express categorical data. One simple test to determine this distinction is to ask the question, "Would it make sense to add these numbers up, or to perform any mathematical operation on them?" For instance, would it make sense to add up order numbers? No, it wouldn't. How about numbers that rank a list of sales people as number 1, number 2, etc.? Once again, the answer is no; therefore, these numbers represent categorical, rather than quantitative, data.

Let's take a look at one more example to practice making the distinction between categorical and quantitative information. Using the graph below, test your skills by identifying which components represent categorical items and which represent quantitative values.

FIGURE 3.5 This graph depicts sales information for practice in distinguishing quantitative and categorical data.

Let's examine one component at a time. For each of the items below, indicate whether the information is categorical or quantitative.

1. The values of time along the bottom (Q1, Q2, etc.)
2. The dollars along the left side
3. The legend, which encodes Direct and Indirect
4. The vertical bars in the body of the graph
5. The title in the top center

Once you've identified each, take a moment to compare your answers to those on the right.

ANSWERS

1. Categorical, labeling the quarters of the year
2. Quantitative, providing dollar values for interpreting the heights of the bars
3. Categorical, providing a distinction between direct and indirect sales
4. Both quantitative and categorical; the heights of the bars encode quantitative information about sales in dollars; the colors of the bars encode categorical data identifying which sales are direct vs. indirect
5. Categorical, identifying the year of the sales

. . . .

Did you catch the dual role of the bars, that they contain both quantitative values and categorical items? With a little practice, you will be able to easily deconstruct graphs into quantitative and categorical data. This ability will enable you to apply the differing design practices that pertain to each type of data.

Choosing the Best Medium of Communication

Choosing whether to display data in one or more tables, one or more graphs, or some combination of the two, is a fundamental challenge of data presentation. This decision should never be arbitrary. It is sad, however, that this choice is often made using the "eeney, meeney, miney, moe" method. Imagine that you're Joe, and you've just interviewed three candidates for a new position in your department. It's now time to report the results to your boss, who's responsible for making the hiring decision. During interviews with the candidates, you used your company's handy interview score sheet to evaluate each person's aptitudes on a zero to five point scale in six areas: experience, communication, friendliness, subject matter knowledge, presentation, and education. How do you display the results? Well, if you're like a lot of people who feel pressure to make an impression, you might decide to use something unusual, like one of those radar charts that are available in your spreadsheet software. As you hand your boss the following report, you hope that he's thinking: "Wow, that Joe is an exceptional employee. Perhaps he deserves a raise."

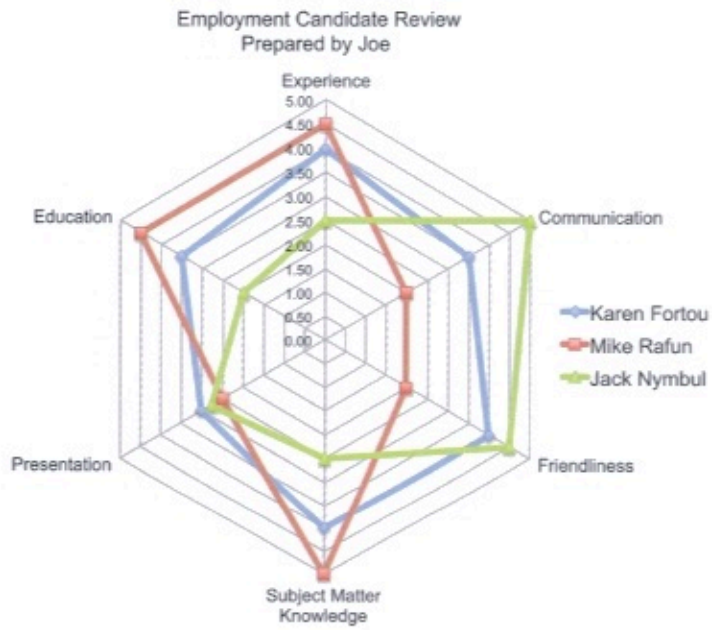


FIGURE 3.6 This radar chart presents aptitudes of job candidates in an overly complicated way.

What's sad is that the boss, upon first glance, might actually think, "That Joe certainly outdid himself," despite the fact that he hasn't a clue how to read this spider-web of a chart. He might actually blame himself for lacking the skill that's needed to read this chart, assuming he must have missed that day in math class when they covered this. Regardless of how the boss responds, Joe would have made his job easier if he'd prepared something like this instead:

Employment Candidate Review

	Candidates		
Rating Areas	Karen Fortou	Mike Rafun	Jack Nymbul
Experience	4.00	4.50	2.50
Communication	3.50	2.00	5.00
Friendliness	4.00	2.00	4.50
Subject matter knowledge	4.00	5.00	2.50
Presentation	3.00	1.50	2.75
Education	3.50	4.50	2.00
Average Rating	3.67	3.25	3.21

FIGURE 3.7 This table presents the same information as Figure 3.6 but in a way that is simple and clear.

This table presents the information in a way that is simple to understand and efficient to use. To select the appropriate medium of communication, we must understand the needs of our audience as well as the purposes for which various forms of display can be effectively used. Let's begin by considering the overall strengths and weaknesses of tables and graphs.

Tables Defined

A table is a structure for organizing and displaying information; a table exhibits the following characteristics:

- Information is arranged in columns and rows.
- Information is encoded as text (including words and numbers).

A single set of values, occupying a single column or row, is merely a list, not a table.

Although the column and row structure of tables is often visually reinforced by grid lines (i.e., horizontal and vertical lines outlining the columns and rows), it is the arrangement of the information that characterizes tables, not the presence of lines that visibly delineate the structure of the underlying grid. In fact, as we will see later in the chapter on table design, grid lines must be used with care to keep them from diminishing a table's usefulness.

Tables are not used exclusively to display quantitative information. Whenever you have more than one set of values, and a relationship exists between values in the separate sets, you may use a table to align the related values by placing them in the same row or column. For instance, tables are often used to display meeting agendas, with start times in one column, the names of the topics that will be covered in the next, and the names of the facilitators in the next, as in the following example:

Time	Topic	Facilitator
09:00 AM	Opening remarks	Scott Wiley
09:15 AM	Product demo	Sheila Prescott
10:00 AM	Discussion	Jerry Snyder
10:45 AM	Planning	Pamela Smart

FIGURE 3.8 This is an example of a table that does not contain quantitative data, in this case a meeting agenda.

Tables have been in use for almost two millennia, so they are readily understood by almost everyone who can read.

When to Use Tables

A handful of conditions should direct you to select a table, rather than a graph, as the appropriate means of display, but tables provide one primary benefit:

> Tables make it easy to look up individual values.

Tables excel as way to display simple relationships between quantitative values and the categorical items to which they're related so that these individual values can be easily located.

When deciding whether to use a table or graph to communicate your quantitative message, always ask yourself how the information will be used. Will you or others use it to look up one or more particular values? If so, it's a prime candidate for expression in a table. Or might the information be used to examine a set of quantitative values as a whole to discern patterns? If so, it's a prime candidate for expression in a graph, as we'll see soon.

Tables also make it easy to compare pairs of related values (e.g., sales in quarter 1 to sales in quarter 2). Here's a typical example:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Monthly Average
2000	138.1	138.6	139.3	139.5	139.7	140.2	140.5	140.9	141.3	141.8	142.0	141.9	**140.3**
2001	142.6	143.1	143.6	144.0	144.2	144.4	144.4	144.8	145.1	145.7	145.8	145.8	**144.5**
2002	146.2	146.7	147.2	147.4	147.5	148.0	148.4	149.0	149.4	149.5	149.7	149.7	**148.2**
2003	150.3	150.9	151.4	151.9	152.2	152.5	152.5	152.9	153.2	153.7	153.6	153.5	**152.4**
2004	154.4	154.9	155.7	156.3	156.6	156.7	157.0	157.3	157.8	158.3	158.6	158.6	**156.9**
2005	159.1	159.6	160.0	160.2	160.1	160.3	160.5	160.8	161.2	161.6	161.5	161.3	**160.5**
2006	161.6	161.9	162.2	162.5	162.8	163.0	163.2	163.4	163.6	164.0	164.0	163.9	**163.0**
2007	164.3	164.5	165.0	166.2	166.2	166.2	166.7	167.1	167.9	168.2	168.3	168.3	**166.6**
2008	168.8	169.8	171.2	171.3	171.5	172.4	172.8	172.8	173.7	174.0	174.1	174.0	**172.2**
2009	175.1	175.8	176.2	176.9	177.7	178.0	177.5	177.5	178.3	177.7	177.4	176.7	**177.1**
2010	177.1	177.8	178.8	179.8	179.8	179.9	180.1	180.7	181.0	181.3	181.3	180.9	**179.9**

FIGURE 3.9 This is an example of a simple table that can be used to look up several years of monthly rates.

Tables work well for look-up and one-to-one comparisons, in part because their structure is so simple, and in part because the quantitative values are encoded as text, which we can understand directly, without translation. Graphs, by contrast, are visually encoded, which requires translation of the information into the numbers it represents.

The textual encoding of tables also offers a level of precision that cannot be provided by graphs. It is easy to express a number with as much specificity as you wish using text (e.g., 27.387483), but the visual encoding of individual numbers in graphs doesn't lend itself to such precision.

Another strength of tables is that they can include multiple sets of quantitative values that are expressed in different units of measure. For instance, if you need to provide sales information that includes the number of units sold, the dollar amount, and a comparison to a forecast expressed as a percentage, doing so in a single graph would be difficult, because a graph usually contains a single quantitative scale with a single unit of measure.

And a final strength of tables is their ability to combine summary and detail information in a single display. For example, a table might include the amount of revenue earned per month (detail), with a total for the year (summary).

To summarize, tables are used to display simple relationships between quantitative values and corresponding categorical items, which makes tables ideal for looking up and comparing individual values. The entire set of reasons to use a table consists of the following; if one or more of these are true, you should probably display the data in a table:

1. The display will be used to look up individual values.
2. It will be used to compare individual values but not entire series of values to one another.
3. Precise values are required.
4. The quantitative information to be communicated involves more than one unit of measure.
5. Both summary and detail values are included.

Graphs Defined

Graphs exhibit the following characteristics:

- Values are displayed within an area delineated by one or more axes.
- Values are encoded as visual objects positioned in relation to the axes.
- Axes provide scales (quantitative and categorical) that are used to label and assign values to the visual objects.

Axes delineate the space that is used to display data in a graph.

Essentially, a graph is a *visual display of quantitative information*. Whereas tables encode quantitative values as text, graphs encode quantitative values visually. Consider this simple example:

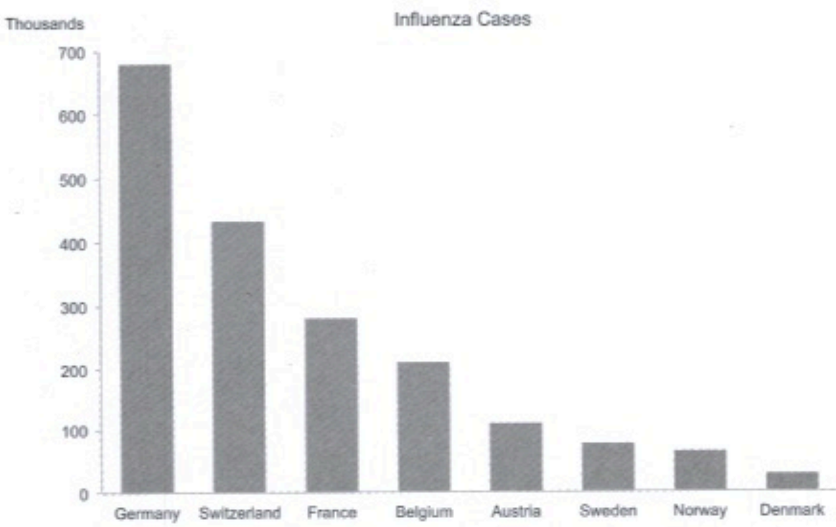


FIGURE 3.10 This is an example of a simple graph, which displays a count of influenza cases per country.

This graph has two axes: one that runs horizontally, called the *X axis*, and one that runs vertically, called the *Y axis*. In this graph, the categorical scale, which labels the countries, resides along the X axis, and the quantitative scale (i.e., counts of influenza cases) resides along the Y axis. The values themselves are encoded as rectangles, called *bars*. Bars are one of several visual objects that can be used to encode data in graphs. The number of influenza cases in each country is encoded as the height of its bar and the position of its top in relation to the scale on the Y axis. The horizontal position of each bar along the X axis is labeled to denote the specific categorical item to which the values are related (e.g., Belgium).

With a little practice, even someone who has never previously used a graph can learn to interpret the information contained in a simple one like this. Although the values of sales for each country cannot be interpreted to the exact number, this isn't the graph's purpose. Rather, the graph paints the picture that influenza has affected these countries to varying degrees, with the greatest effect in Germany and the least in Denmark, and significant differences in between. Information like this, which is intended to show patterns in data, is best presented in a graph rather than a table, as we'll see in the section *When to Use Graphs*. But first, a little history.

A Brief History of Graphs

Graphs of quantitative information have been in use for only a few hundred years, which is a relatively short time given the thousands of years that mathematics has existed. Despite how natural it may seem to see quantitative information displayed in graphs, the original notion that numbers could be displayed visually in relation to two perpendicular axes involved a leap of imagination. The launching pad for this leap had already been around for many centuries before quantitative graphs emerged. An earlier type of visual information display, also assisted by a scale of measurement along perpendicular axes, eventually suggested the possibility of graphs. Can you guess what it was? It is still in common use today for the purpose of navigation. It is a two-dimensional representation of the physical world that's used to measure distances between locations. I'm referring to a *map*. The earliest known map dates back about 4,300 years. It was drawn on a clay tablet and represented northern Mesopotamia. When a map depicts the entire world, or some large part of it, the standard set of grid lines that allow us to determine location and distance are longitudes and latitudes. These grid lines form a quantitative scale of sorts.

In the 17th century these grids were adapted for the representation of numbers alone. In his *La Géométrie* (1637), René Descartes introduced them as a means to encode numbers as coordinate positions in a two-dimensional (2-D) grid. This innovation provided the groundwork for an entire new field of mathematics that is based on graphs. To some extent even prior to Descartes, others had already experimented with graphical displays of numbers. Ron Rensink writes:

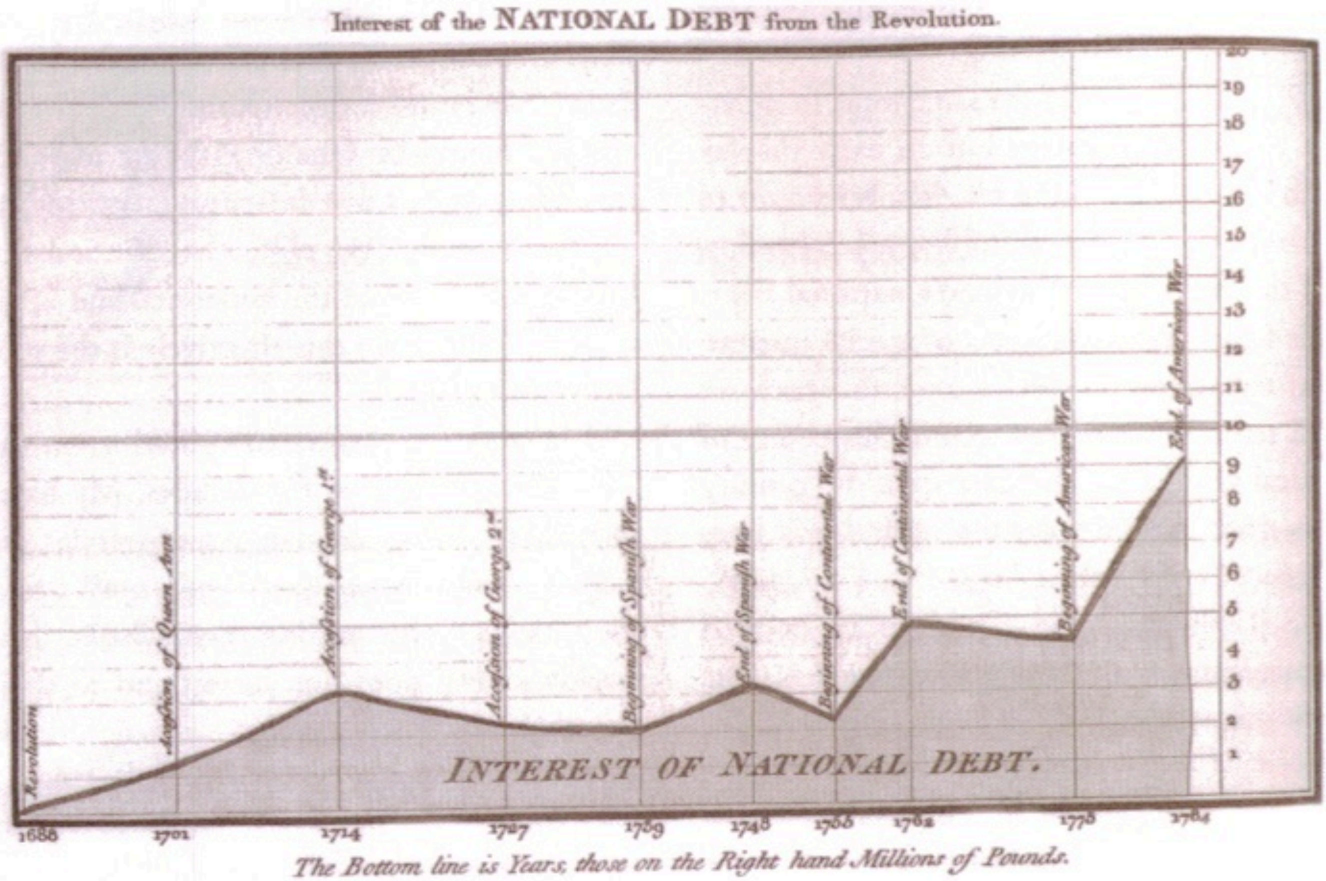
> *Although Descartes did contribute to the graphic display of quantitative data in the 17th century, other forms of quantitative graphs had already been used to represent things such as temperature and light intensity three centuries earlier. Indeed, as Manfredo Massironi discusses in his book (Massironi, M. (2002) The Psychology of Graphic Images: Seeing, Drawing, Communicating. Erlbaum, page 131), quantities such as displacement were graphed as a function of time as far back as the 11th century. But while these facts may be of interest in their own right, the more important point*

is that techniques in graphic representation have been developed over many centuries, and many of these techniques have been subsequently forgotten—perhaps fallen out of vogue, or never found wide use to begin with. But the reasons for their dismissal may not necessary apply in this day and age. Indeed, several techniques might lend themselves quite well to modern technology, and so might be worth resurrecting in one form or other. Books such as Massironi's are helpful in discovering such possibilities. [1]

1. Ron Rensink (2011) "Four Futures and a History," www.Interaction-Design.org.

Despite these earlier ventures, it wasn't until the late 18th century that the use of graphs to present numbers became popular. William Playfair, a Scottish social scientist, used his imagination and design acumen to invent many of the graphing techniques that we use today. He pioneered the use of graphs to reveal the shape of quantitative information, thus providing a way to communicate quantitative relationships that numbers expressed as text could never convey. He invented the bar graph, was perhaps the first person to use a line to show how values change through time, and on a day when he was probably under the weather, invented the pie chart as well, but we'll forgive his brief lapse of judgment.

Information about the early development of the graph, including its precursor, the map, may be found in Robert E. Horn (1998) *Visual Language*. MacroVU, Inc. Robert Horn provides an informative timeline, which cites the milestones in the historical development of visual information display.



The old saying, "A picture is worth a thousand words" applies quite literally to graphs. By presenting quantitative information in visual form, graphs efficiently communicate what might otherwise require a thousand or even a million words, and sometimes communicate what words could never convey.

FIGURE 3.11 This graph was included in William Playfair's book *The Commercial and Political Atlas* in 1786 to make a case against England's policy of financing colonial wars through national debt.

From the time of Playfair until today, many innovators have added to the inventory of graph designs available for the representation, exploration, analysis, and communication of quantitative information. During the past 50 years, none has contributed more to the field as an advocate of excellence in graphic design than Edward Tufte, who in 1983 published his landmark treatise on the subject, *The Visual Display of Quantitative Information*. With the publication of additional books and articles since, Tufte continues to be respected as a major authority in this field.

The work of William S. Cleveland, especially his book *The Elements of Graphing Data*, is also an outstanding resource.[2] Cleveland's work is particularly useful to those with statistical training who are interested in sophisticated graphs, such as those used in scientific research.

2. William S. Cleveland (1994) *The Elements of Graphing Data*. Hobart Press.

When to Use Graphs

Graphs reveal more than a collection of individual values. Because of their visual nature, graphs present the overall shape of the data. Text, displayed in tables, cannot do this. The patterns revealed by graphs enable readers to detect many points of interest in a single collection of information. Take a look at the next example, and try to identify some of the features that graphs can reveal. Approach this by first determining the messages in the data, then by identifying the visual cues—the shapes—that reveal each of these messages.

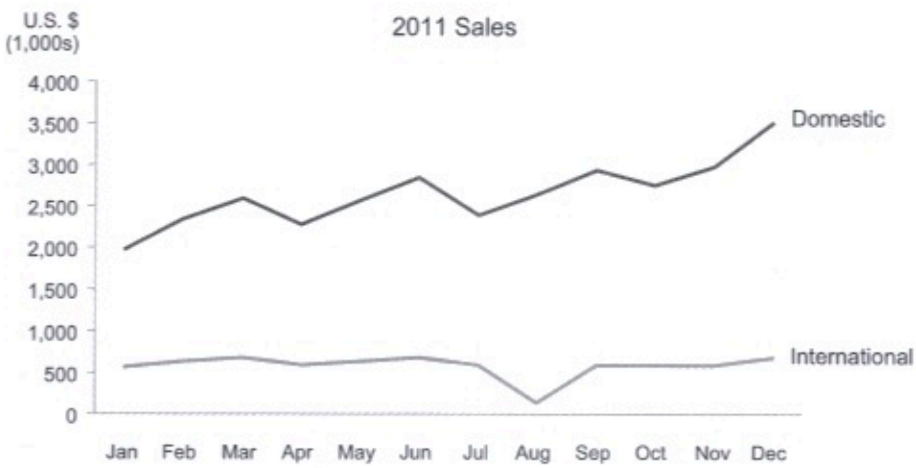


FIGURE 3.12 This fairly typical line graph is an example of the shapes and patterns in quantitative data that graphs make visible.

What insights are brought to your attention by the shape of this information? Take a moment to list them in the right margin.

. . . .

Let's walk through a few of these revelations together, beginning with domestic sales. What does the shape of the black line tell you about domestic sales during the year 2011? One message is that, during the year as a whole, domestic sales increased; these sales ended higher than they started, with a gradual increase through most of the year. One name for this type of pattern is a *trend*, which displays the overall nature of change during a particular period of time.

Was this upward trend steady? No, it exhibits a pattern that salespeople sometimes call the *hockey stick*. Sales go down in the first month of each quarter and then gradually go up to a peak in the last month of each quarter. If you examine the shape of the line from the last month of one quarter, such as March, to the last month of the next quarter, such as June, you'll recognize that these segments each look a little like a hockey stick. This pattern in sales usually occurs when salespeople are given bonuses based on reaching or exceeding a quarterly quota.

Now, if you look at sales in the first quarter of the year versus sales in the second quarter, then the third, and finally the fourth, you find that the graph makes it easy to see how these different periods compare. When information is given shape in a graph, it becomes possible to compare entire sets of data. In this case the comparison is between different quarters.

Now look for a moment at international sales. Once again you are able to easily see the trend of sales throughout the year, which in this case is relatively flat. Compared with domestic sales, international sales seem to exhibit less fluctuation through time and relatively little difference between the beginning and end of the year. Although international sales appear to be less affected by quarters and seasons, one point along the line stands out as quite different from the rest: the month of August. Sales in August took an uncharacteristic dip compared to the rest of the year. As an analyst, this abnormal sales value would make you want to dig for the cause. If you did, perhaps you would discover that most of your international customers were vacationing in August and therefore weren't around to place orders. Whatever the cause, the current point of interest to us is that graphs make exceptions to general patterns stand out clearly from the rest. This wouldn't be nearly so obvious in a table of numbers.

Finally, if you widen your perspective to all the quantitative information, you find that the graph makes it easy to see the similarities and differences between the two sets of values (domestic and international sales), both overall and at particular points in the graph. We could go on, adding more to the list of characteristics that graphs reveal, but we've hit the high points and are now ready to distill what we've detected to its essence:

> Graphs are used to display relationships among and between sets of quantitative values by giving them shape.

The visual nature of graphs endows them with their unique power to reveal patterns of various types, including changes, differences, similarities, and exceptions. Graphs can communicate quantitative relationships that are much more complex than the simple associations between individual quantitative values that tables can express.

Graphs can display large data sets in a way that can be readily perceived and understood. You could gather data regarding the relationship between employee productivity and the use of two competing software packages, involving thousands of records across several years, and, with the help of a graph, you would be able to immediately see the nature of the relationship. If you have ever tried to use a huge table of data for analysis, you would quickly fall in love with graphs like scatter plots, which can make the relationships among thousands of individual data points instantly intelligible.

Patterns in data are difficult to discern even in a small table of numbers. Despite our agreement on most matters, Tufte and I disagree slightly about this. According to Tufte:

> *Tables usually outperform graphics in reporting on small sets of 20 numbers or less. The special power of graphics comes in the display of large data sets.* [3]

3. Edward Tufte (2001) *The Visual Display of Quantitative Information*, Second Edition. Graphics Press, page 56.

Although it is true that one's chance of seeing a meaningful pattern in small table of numbers is greater than in one that is large, those patterns are much easier to see in a graph. Take a look at the following table.

Job Satisfaction By Income, Education, and Age

	College Degrees		No College Degrees	
Income	Under 50	50 & over	Under 50	50 & over
Up to $50,000	643	793	590	724
Over $50,000	735	928	863	662

FIGURE 3.13 This table displays measures of employee job satisfaction, divided into categories.

Do any unusual patterns of job satisfaction pop out? Can you see that one group of employees exhibits a different pattern of satisfaction from the others? Now take a look at the following graph.

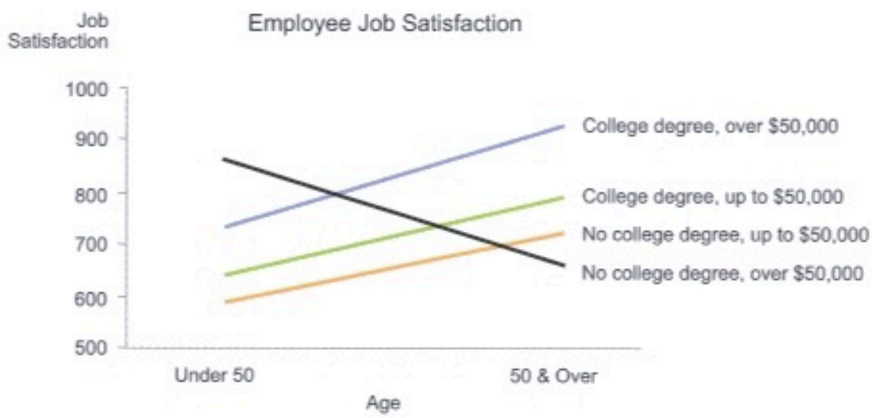


FIGURE 3.14 This graph causes the pattern represented by the black line to pop out.

The fact that employees with salaries over $50,000 but no college degrees experience a significant decrease in job satisfaction in their older years now jumps out. Even with only eight numbers, this graph did what the table couldn't do—it made this diverging pattern obvious. This results not only because the information is displayed graphically, but also because the graph was specifically designed to feature this particular pattern. Had I designed the graph in the following manner, that pattern would no longer be obvious.

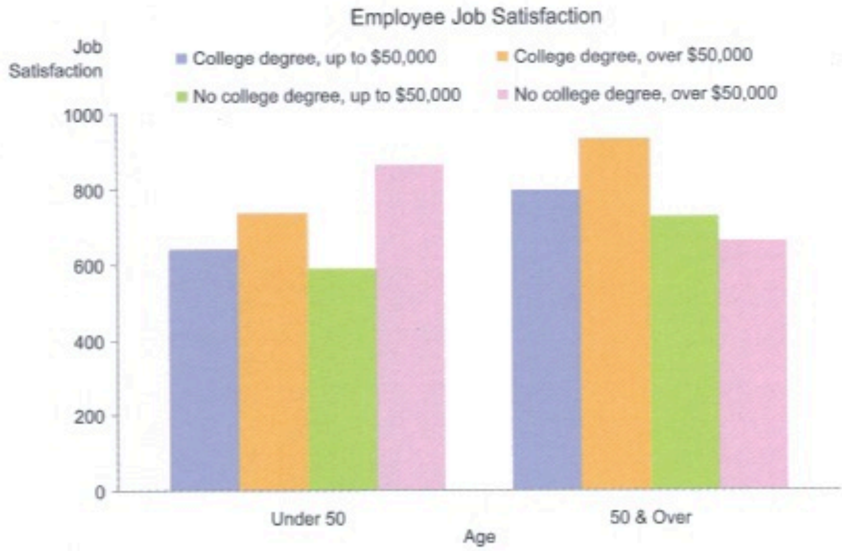


FIGURE 3.15 Unlike the graph in *Figure 3.14*, this graph does not make the pattern of decreasing job satisfaction among the one group of employees obvious.

This example is not based on real data, so if you make more than $50,000 a year but don't have a college degree and will soon turn 50, don't worry, you'll probably do just fine.

So displaying information in a graph rather than a table does not by itself make meaningful patterns visible. We must design the graph to feature evidence of the particular story that we're trying to tell. Many possible stories dwell in a set of data, even in a small one such as illustrated above. Later in this book, we'll learn to make the necessary graph design choices to tell particular stories.

Summary at a Glance

Use Tables When	Use Graphs When
• The display will be used to look up individual values. • It will be used to compare individual values. • Precise values are required. • The quantitative values include more than one unit of measure. • Both detail and summary values are included.	• The message is contained in the shape of the values (e.g., patterns, trends, and exceptions). • The display will be used to reveal relationships among whole sets of values.