BUSINESS ANALYTICS

PRINCIPLES, CONCEPTS, AND APPLICATIONS

WHAT, WHY, and HOW

MARC J. SCHNIEDERJANS • DARA G. SCHNIEDERJANS • CHRISTOPHER M. STARKEY

2. Why Are Business Analytics Important?

Chapter objectives:

• Explain why business analytics are important in solving business problems.

• Explain why business analytics are important in identifying new business initiatives.

• Describe the kinds of questions business analytics can help answer.

• Explain how business analytics can help an organization achieve a competitive advantage.

• Explain different types of competitive advantages and their relationship to business analytics.

• Explain the importance of business analytics for a business organization.

2.1. Introduction

Telecommunication and information systems are collecting data on every aspect of life with incredible rates of speed and comprehensiveness. In addition, businesses are running opinion surveys and collecting all forms of data for their operations. With information system clouds providing large amounts of data that are easily available and data warehousing systems capable of storing big data in large databases, there is presently a need to process information out of data to gain knowledge and justify data investment. As **Demirkan and Delen (2013)** have shown, placing large data into computer clouds can provide business analytics in a timely and agile way. Firms recognize the need for this information to be competitive, and business analytics is one strategy to gain the knowledge they seek.

The problem with big data or even small data files is that they can easily obscure the information desired. Sometimes a small alteration in a piece of data located in a file can change meanings. The 1960s television program *The Prisoner* used the catch phrase, "I want information." When this phrase is seen in print or spoken, it denotes that someone wants information. Yet when the term was used in *The Prisoner*, it referred to "in" and "formation." (That is, "I want in formation.") The phrase was used to make the prisoner do what he was told and act like the others. Note that a single space in this second phrase completely changes the meaning. Mining for relevant business information in big databases when small differences can alter meanings makes it a challenge to find relevant and useful information. Business analytics as a process is designed to meet this challenge.

2.2. Why BA Is Important: Providing Answers to Questions

It may seem overly virtuous, but BA is the next best thing to a crystal ball for answering important business questions. In each of the three steps of the BA process (from **Chapter 1**, "**What Are Business Analytics?**"), answers to a variety of questions can and should be answered as a logical outcome of the analysis. The answers become the basis of information and knowledge that makes BA a valued tool for decision-making and helps explain why it is important to learn and use.

As can be seen in **Table 2.1**, the sampling of the kinds of questions a typical BA analysis can render is re-

lated not only to each step in the BA process, but to the context of time. To better understand the value of the information BA analysis provides and understand why this subject is important to improved business performance, a simple illustrative case scenario is presented.

	Time Period		
Step in BA	**Past**	**Present**	**Future**
1. Descriptive	What happened in the past?	What is happening now based on the past?	What will appear to happen based on the past?
2. Predictive	How did it happen in the past? Why did it happen in the past?	What possible trends exist in the data that can predict or forecast what course of action should be taken now?	What is the range and likelihood of possible outcomes that can happen if the current trends or forecasts are allowed?
3. Prescriptive	How best can we leverage what we know from the trends and forecasts?	How can we optimally apply resources to maximize the business performance outcomes in the future?	How can we continuously apply BA in the future to optimize upcoming business performance outcomes?

Source: Adapted from Exhibit 9.1 from **Isson and Harriott (2013)**, p. 169.

Table 2.1 Questions Business Analytics Seeks to Answer*

In this illustrative case scenario a local credit union offers a series of packaged homeowner loans that are periodically marketed by running a promotional campaign in a variety of media (print ads, television commercials, radio spots). The idea is to bring in new customers to make home loans that fit one of the packaged deals. Halfway through the marketing program, the credit union does not know if the business generated is due to the promotional campaign or just a result of their normal business cycle. To clarify, the credit union undertakes a BA analysis. The resulting information from the BA analysis (based on the same questions as in **Table 2.1**) is presented in **Table 2.2**. Reading first the Descriptive step, Past, Present, and Future, and then sequentially following the same pattern with the Predictive and Prescriptive steps, the possible types of information gleaned from these BA questions and answers can be illustrated by this example.

	Time Period		
Step in BA	**Past**	**Present**	**Future**
1. Descriptive	Based on graphics results, past ad campaigns resulted in a moderate increase in new loans.	Based on sorting of loan activities, new homeowner loans are experiencing just a moderate increase in new loan applications.	Based on histogram of loans to date, there is no discernible pattern, just uniform new loan sales that are constant over time. No business cycle impact is observed to alter loan patterns.

2. Predictive	Statistically, correlations have revealed in the past that marketing promotions will increase new loans, but why they generate new loans depends on how the promotion campaign invests its funding.	Utilizing multiple regression, the model predicts that a greater allocation in funds for television commercials and print ads will be more effective in generating new loans than investing in radio	Utilizing variance statistics from the regression model, a confidence interval can estimate the number of new loans possible if a reallocation of promotion funds is implemented.
3. Prescriptive	Reallocating marketing budget funds from radio spots to television commercials and print ads is required to more effectively reach the target audience.	Given the constrained resource of funding, a linear programming model is used to optimally allocate the marketing budget in dollars to maximize the promotional outcome for new loans.	Continuous tracking over a period of time of the new loan applications caused by the promotion campaign needs to be monitored and mapped against the predicted outcomes suggested in the analysis.

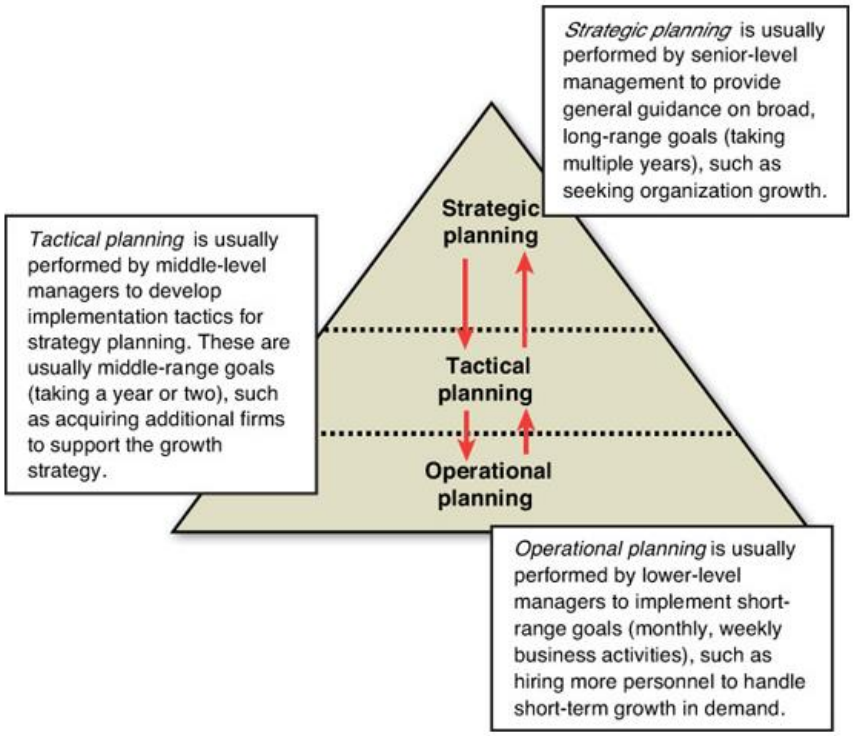
Table 2.2 Credit Union Example of BA Analysis Information

The answers to the questions raised in the credit union example are typical of any business organization problem-solving or opportunity-seeking quest. The answers were not obtained by just using statistics, computer search routines, or operations research methodologies, but rather were a result of a sequential BA process. The informational value of the answers in this scenario suggests a measurable and precise course of action for the management of the credit union to follow. By continuously applying BA as a decision support system, firms have come to see not only why they need BA, but also how BA can become a strategy to achieve competitive advantage. **Kiron et al. (2012)** reported in a survey on business through the year 2012 that firms applying business analytics permit the organization to have better access to data for decision-making and offer a competitive advantage.

2.3. Why BA Is Important: Strategy for Competitive Advantage

Companies that make plans that generate successful outcomes are winners in the marketplace. Companies that do not effectively plan tend to be losers in the marketplace. Planning is a critical part of running any business. If it is done right, it obtains the results that the planners desire.

Business organization planning is typically segmented into three types, presented in **Figure 2.1**. The planning process usually follows a sequence from strategy, down to tactical, and then down to operational, although **Figure 2.1** shows arrows of activities going up and down the depicted hierarchal structure of most business organizations. The upward flow in **Figure 2.1** represents the information passed from lower levels up, and the downward flow represents the orders that are passed from higher levels of management down to lower levels for implementation. It can be seen in the **Teece (2007)** study and more recently in **Rha (2013)** that the three steps in the BA process and strategic planning embody the same efforts and steps.



*Source: Adapted from Figure 1.2 in Schniederjans and LeGrand (2013), p.9.

Figure 2.1 Types of organization planning*

Effectively planning and passing down the right orders in hopes of being a business winner requires good information on which orders can be decided. Some information can become so valuable that it provides the firm a *competitive advantage* (the ability of one business to perform at a higher level, staying ahead of present competitors in the same industry or market). Business analytics can support all three types of planning with useful information that can give a firm a competitive advantage. Examples of the ways BA can help firms achieve a competitive advantage are presented in Table 2.3.

Type of Competitive Advantage	Description	Ways BA Can Help Achieve the Competitive Advantage
Price Leadership	From a marketing standpoint, offer products or services at the lowest cost to customers in the industry, while making acceptable profit for the company.	Identify main competitors; monitors, reports, and accurately forecasts competitive prices so firm can keep lowest cost profile while maintaining and measuring profit margins.
Sustainability	To ensure the firm's resource usage in a way that seeks balance to hurt neither the environment nor the bottom line of a firm's profitability.	Identify areas where resource reallocations are needed to avoid damaging the environment, suggest ways to reallocate the resources, and help allocate them optimally to achieve the best possible balance.
Operations Efficiency	Improve the internal business operations and activities over competitors, lessening the cost to the customer. That reduced cost, if passed on to customers, can provide a lower price advantage based on efficiency.	Identify operation areas needing correction or modification and suggest alternatives to improve efficiency. Also, this can be useful in selecting which alternative to use to maximize business performance.

Operations Efficiency	Improve the internal business operations and activities over competitors, lessening the cost to the customer. That reduced cost, if passed on to customers, can provide a lower price advantage based on efficiency.	Identify operation areas needing correction or modification and suggest alternatives to improve efficiency. Also, this can be useful in selecting which alternative to use to maximize business performance.
Service Effectiveness	Make customer transactions easier or more pleasurable than with other firms. This improves the service characteristics of the firm while lowering the time it takes to get services to the customer, thus enhancing customer value.	Obtain customer opinions on problem service areas needing fixing; explain why the fix is needed, suggest alternatives to the fix, improve the effectiveness of the service operations, and measure and report improvements.
Innovation	Introduce completely new or notably better products or services with the intention of disrupting competitors' businesses by obsoleting the current market entries with break-through product offerings.	Obtain and validate customer ideas and suggestions on new products or enhancements in current products. Monitor customer reactions and suggest refinements as new products are introduced to customers. Monitor performance on new products and report results.
Product Differentiation	Provide customers with a variety of products, services, or features that competitors are not yet offering or are unable to offer.	Identify new products not offered by competitors, suggest new services to offer, forecast potential of new products for profitability measurement.

Table 2.3 Ways BA Can Help Achieve a Competitive Advantage

2.4. Other Reasons Why BA Is Important

There is an almost endless list of potential applications of BA to provide information on which decisions can be made or improved.

2.4.1. Applied Reasons Why BA Is Important

Some potential applications for decision-making will be illustrated in later chapters. Several brief examples are described in **Table 2.4**.

Decision-Making Application	Description
Increasing Customer Profitability	BA can provide detailed information (current pricing and pricing trends) on competitor products. This information can be used to set prices that allow a firm to keep profit margins at a profit-maximizing level by balancing sales volume with lower prices and margins or increasing prices to increase margins, depending on competitor pricing.
Risk Reduction	With the types of information provided in all steps of the BA process, businesses do not have to guess but instead can be guided with some probabilistically computed likelihood of certainty on sales, budgets, and human and technology decisions. Having a probability estimate as a guide reduces the risk of poor judgments.

Merchandize Strategy Optimization	Quantitative tools used in the Prescriptive step of the BA process can be used to determine optimal layout designs for store merchandize, cost minimizing inventory levels, and even scheduling for sales staff to help achieve maximum merchandising results.
Human Resource Decisions	According to Fitx-enz (2013, pp. 223–245) analytics determined on *human resources* (HR) in the Predictive step of the BA process can be computed (workforce productivity, revenue per *full-time equivalent* (FTE), cost per FTE, and so on). In turn, this can answer questions like what should be done to improve the quality of HR hires, what types of training would be most effective, and how incentive pay can be used to stimulate performance.
Business Performance Tracking	In addition to the normal data collection in the Descriptive step of the BA process, specific business performance parameters can be continually collected, monitored, and measured. Then analytics can update those measures to provide an up-to-date performance achievement index useful for comparing performance over time. In addition, the Predictive step of the BA forecasts of expected performance can be used to set planning and performance goals to guide operations.

Table 2.4 Applications of BA to Enhance Decision-Making

2.4.2. The Importance of BA with New Sources of Data

As advances in new computer and telecommunication technologies take place, they provide new types of data. Therefore, new types of analytics need to be applied in BA analyses. *Digital analytics* is a term that describes any source of data that is conveyed using digital sources. Examples of these new sources of data-based analytics include text analytics and unstructured data analytics. *Text analytics* can be defined as a set of linguistic, statistical, and computer-based techniques that model and structure the information content from textual sources (**Basole, et al., 2013**). It is a search process in databases to find patterned text material that provides useful information. Also referred to as *text data mining*, text analytics uses data mining software to look into databases to find and validate the kinds of information on which predictions can be made.

Being able to search and quantify textual data using text analytics opens great opportunities to glean information about customers and markets based on technology-driven data collection technologies. One example of technology-driven data is social media data. *Social media* can be defined as interactions or communications among people or communities, usually performed on a technology platform, involving the sharing, creating, discussing, and modifying of communicated verbal or electronic content. Two global social platforms are *Twitter* and *Facebook*. The methodologies or technologies used in the purveyance of social media data can include any means of distribution of verbal or other types of communications, including, but not limited to, photographs or pictures, video, Internet forums, web logs, discussion forums, social blogs, wikis, social networks, and podcasts. These sources of data are the basis of *social media analytics*, on which the analytics information can aid in learning new types of social media behavior and information. They provide a great challenge for BA analysts because of the excessive volume and difficulty in quantifying the information in useful ways. They also provide a great opportunity to find information that might create a competitive advantage. An example of how social media analytics helped find auto defects was illustrated in a study by **Abrahams et al. (2012)**. By employing text mining on a social medium (online discussion forums) used by vehicle enthusiasts, a variety of quality defects were identi-

Another similar digital source of analytics is referred to as mobile analytics. *Mobile analytics* can be defined as any data secured from mobile devices, such as smartphones, iPhones, iPads, and Web browsers. These are all mobile technologies used to obtain digital data from the interaction of people (**Evans, 2010**). The fact that they are mobile and move from location to location with the user differentiates the type of information available to the analytics analysts. For example, the mobile technology allows analysts to not only track what a potential customer might talk about on the use of a product (such as in social media analytics), but track movements of where the customer makes decisions on products. That can help explain why those decisions are made. For example, mobile technology might reveal the location of a purchaser of hair spray to have been physically located near an area where billboards are used for hair spray advertising, thus helping to reveal the possible connection and effectiveness of a billboard promotion.

When data is placed in databases and can be logically filed, accessed, referenced, and used, it is known as *structured data*. When data or information, either digital or nondigital, cannot be put into a database or has no predefined structure, it is known as *unstructured data*. Examples of unstructured data include images, text, and other data that, for one reason or another, cannot be placed in a logically searchable database based on content. This data can be digitally stored in a file, but not in a way that can be usefully retrieved using any kind of logic model or sorting process. Much of the data contained in emails and on the Web is unstructured. Another way of looking at unstructured data is that it is what is left over and cannot be placed in a structured database. As time goes on, more effort in developing complex algorithms and other computer-based technologies will be applied to unstructured data, reducing the amount of data that falls into this category. Given the volume of graphics data or other unstructured data generated every day, the challenge to BA analysts will be an ever-growing effort to understand and structure the unstructured data that remains in an effort to gain its informational value. Part of the value and importance of BA is in accepting this challenge.

Summary

This chapter sought to explain why business analytics is an important subject for business organizations. It discussed how BA can answer important questions and how it can help a firm achieve a competitive advantage. In addition, it presented the role of BA in organization planning. Finally, it introduced other types of digital analytics to explain their beneficial role and challenges to BA.

We move in the next chapter to further explain why BA is important in the context of its required investment. Like any management task, the successful use of BA requires an investment in human and technology resources. **Chapter 3**, "**What Resource Considerations Are Important to Support Business Analytics?**", explores the allocation of resources to maximize BA performance and explains why the investment is needed.

Discussion Questions

1. Why does each step in the business analytics process have a past, present, and future dimension?

2. What is a competitive advantage, and how is it related to BA?

3. Why does having the ability to aid in decision-making make BA important?

4. How does BA help achieve sustainability?

Chapter 3. WHAT RESOURCE CONSIDERATIONS ARE IMPORTANT TO SUPPORT BUSINESS ANALYTICS?

Chapter objectives:

• Explain why personnel, data, and technology are needed in starting up a business analytics program.

• Explain what skills business analytics personnel should possess and why.

• Describe the job specialties that exist in business analytics.

• Describe database encyclopedia content.

• Explain the categorization of data in terms of sources.

• Describe internal and external sources of data.

• Describe an information technology infrastructure.

• Describe a database management system and how it supports business analytics.

3.1. Introduction

To fully understand why business analytics (BA) is necessary, one must understand the nature of the roles BA personnel perform. In addition, it is necessary to understand resource needs of a BA program to better comprehend the value of the information that BA provides. The need for BA resources varies by firm to meet particular decision support requirements. Some firms may choose to have a modest investment, whereas other firms may have BA teams or a department of BA specialists. Regardless of the level of resource investment, at minimum, a BA program requires resource investments in BA personnel, data, and technology.

3.2. Business Analytics Personnel

One way to identify personnel needed for BA staff is to examine what is required for certification in BA by organizations that provide BA services. *INFORMS* (**www.informs.org/Certification-Continuing-Ed/Analytics-Certification**), a major academic and professional organization, announced the startup of a *Certified Analytic Professional* (CAP) program in 2013. Another more established organization, *Cognizure* (**www.cognizure.com/index.aspx**), offers a variety of service products, including business analytic services. It offers a general certification *Business Analytics Professional* (BAP) exam that measures existing skill sets in BA staff and identifies areas needing improvement (**www.cognizure.com/cert/bap.aspx**). This is a tool to validate technical proficiency, expertise, and professional standards in BA. The certification consists of three exams covering the content areas listed in **Table 3.1**.

Exam	Topic	Specific Content Areas Covered	Examples
I	Statistical Methods	1. Visualizing and Exploring Data 2. Descriptive Statistics 3. Probability Distributions 4. Sampling and Estimation 5. Statistical Inference 6. Regression Analysis 7. Predictive Modeling and Analysis	1. Graphs and charts 2. Mean, median, mode 3. Normal distribution 4. Confidence intervals 5. Hypothesis testing 6. Multiple regression 7. Curve fitting of models and functions to raw data
II	Operations Research Methods	1. Linear Optimization 2. Integer Optimization 3. Nonlinear Optimization 4. Simulation 5. Decision Analysis 6. Forecasting	1. Linear programming 2. Integer programming 3. Quadratic programming 4. Monte Carlo method 5. Expected value analysis 6. Exponential smoothing
III	Case Studies	Practical knowledge of real-world situations	Application of the methods above to solve a real world problem

Table 3.1 Cognizure Organization Certification Exam Content Areas*

Most of the content areas in **Table 3.1** will be discussed and illustrated in subsequent chapters and appendixes. The three exams required in the Cognizure certification program can easily be understood in the context of the three steps of the BA process (descriptive, predictive, and prescriptive) discussed in previous chapters. The topics in **Figure 3.1** of the certification program are applicable to the three major steps in the BA process. The basic statistical tools apply to the descriptive analytics step, the more advanced statistical tools apply to the predictive analytics step, and the operations research tools apply to the prescriptive analytics step. Some of the tools can be applied to both the descriptive and the predictive steps. Likewise, tools like simulation can be applied to answer questions in both the predictive and the prescriptive steps, depending on how they're used. At the conjunction of all the tools is the reality of case studies. The use of case studies is designed to provide practical experience where all tools are employed to answer important questions or seek opportunities.

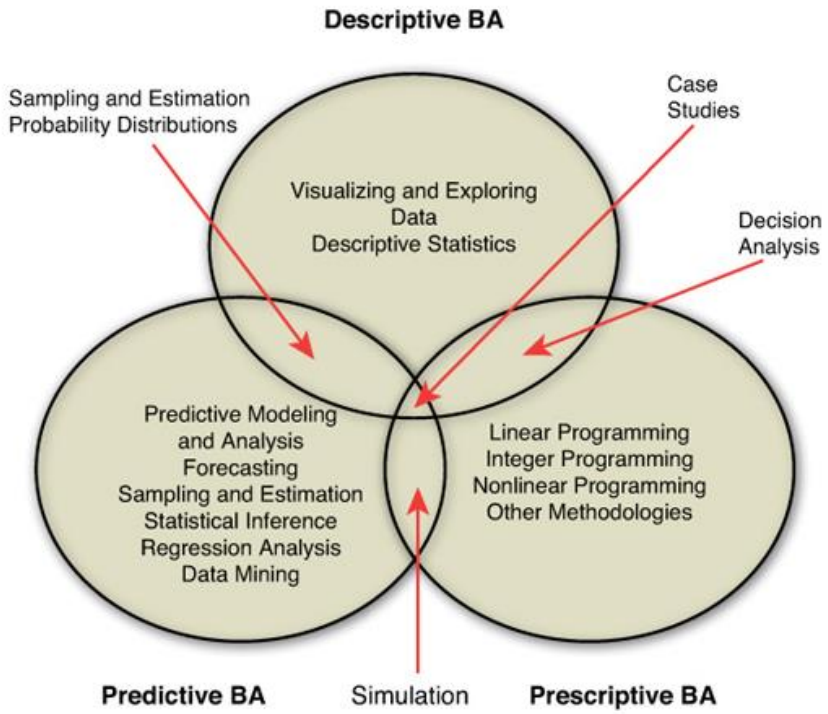


Figure 3.1 Certification content areas and their relationship to the steps in BA

Other organizations also offer specialized certification programs. These certifications include other areas of knowledge and skills beyond just analytic tools. IBM, for example, offers a variety of specialized BA certifications (**www-03.ibm.com/certify/certs/ba_index.shtml**). Although these include certifications in several dozen statistical, information systems, and analytic methodologies related to BA, they also include specialized skill sets related to BA personnel (administrators, designers, developers, solution experts, and specialists), as presented in **Table 3.2**.

BA Personnel Specialty	Description
Administrators	Within the context of the IBM BA and business intelligence (BI) software platforms, administrators manage servers (their load balancing, installation, and configurations). They manage reports from computer portals, manage dispatchers, and perform troubleshooting for technology. They are also in charge of user authorization and authentication for security.
Designers	As members of a team, designers are responsible for building reports using relational data models, as well as enhancing, customizing, and managing professional reports.
Developers	As members of a team, the skills required for developers are closely tied to the BA process and involve the application of analytics, data warehousing, model building, use of operations research and statistical methodologies, and real-time monitoring of data flow to users.

Solution Experts	As members of a team, solution experts analyze, plan, design, deploy, and operate BA applications using an appropriate methodology and development approach. This requires knowledge in many differing BA software applications, including statistical, information systems, and operations research methods.
Technical Specialists	As members of a team, they are responsible for the installation and configuration of BA and BI applications.

Source: Adapted from IBM website (www-03.ibm.com/certify/certs/ba_index.shtml).

Table 3.2 Types of BA Personnel*

With the variety of positions and roles participants play in the BA process, this leads to the question of what skill sets or competencies are needed to function in BA. In a general sense, BA positions require competencies in business, analytic, and information systems skills. As listed in Table 3.3, business skills involve basic management of people and processes. BA personnel must communicate with BA staffers within the organization (the BA team members) and the other functional areas within a firm (BA customers and users) to be useful. Because they serve a variety of functional areas within a firm, BA personnel need to possess customer service skills so they can interact with the firm's personnel and understand the nature of the problems they seek to solve. BA personnel also need to sell their services to users inside the firm. In addition, some must lead a BA team or department, which requires considerable interpersonal management leadership skills and abilities.

Type of Skill or Competency	Description of Possible Roles
Business	• Leadership • People-related management and communication skills • Manage BA projects (prioritize, schedule, and so on) • Manage BA processes (rules, procedures governing acceptance and use) • Determine project requirements • Train BA personnel to develop competencies
Analytic	• Know how to use statistical methodologies • Know how to use operations research methodologies • Know how to use data mining for quantitative data and text analytics for unstructured data
Information system	• Maintain and use computer portals • Identify and extract data • Maintain quality data

Table 3.3 Select Types of BA Personnel Skills or Competency Requirements

Fundamental to BA is an understanding of analytic methodologies listed in Table 3.1 and others not listed. In addition to any tool sets, there is a need to know how they are integrated into the BA process to leverage data (structured or unstructured) and obtain information that customers who will be guided by the analytics desire.

In summary, people who undertake a career in BA are expected to know how to interact with people and utilize the necessary analytic tools to leverage data into useful information that can be processed, stored, and shared in information systems in a way that guides a firm to higher levels of business performance.

3.3. Business Analytics Data

Structured and unstructured data (introduced in Chapter 2, "Why Are Business Analytics Important?") is needed to generate analytics. As a beginning for organizing data into an understandable framework, statisticians usually categorize data into meaning groups.

3.3.1. Categorizing Data

There are many ways to categorize business analytics data. Data is commonly categorized by either internal or external sources (Bartlett, 2013, pp. 238–239). Typical examples of internal data sources include those presented in Table 3.4. When firms try to solve internal production or service operations problems, internally sourced data may be all that is needed. Typical external sources of data (see Table 3.5) are numerous and provide great diversity and unique challenges for BA to process. Data can be measured quantitatively (for example, sales dollars) or qualitatively by *preference surveys* (for example, products compared based on consumers preferring one product over another) or by the amount of consumer discussion (chatter) on the Web regarding the pluses and minuses of competing products.

Type of Internal Data	Description
Billing and Reminder Systems	Billing systems and reminder systems print bills and monitor customer payment information on valued-based customer segments.
Business	Industry codes, accounting information, personnel information, and so on are routinely collected in the course of business.
Customer	Names, addresses, returns, special contracts, segmentations, and so on are obtained when customers sign for or pay for products or services.
Customer Relationship Management Systems	*Customer relationship management* (CRM) systems collect and provide data on customer history, behavior on matters like complaints, the end of a relationship with a firm, and so on.
Human Resources	Information about employees, salaries, competencies, and so on is recorded by routine efforts over the history of employment.
Information from Enterprise Resource Planning Systems	*Enterprise resource planning* (ERP) systems are used to communicate internal business transactions to provide a direct feed of information on management issues and concerns, as well as other operations activities required to produce and sell products.

Product	Information is collected from procurement through post sales to monitor profitability, durability, and quality.
Production	Information that can be used to optimize production, inventory control, and supply chain delivery of the product to the customers is collected during the production processes.
Questionnaires	Information on customer behavior is obtained by customer questionnaires to measure customer service and product quality, among other things.
Web Logs	Information is collected on the firm's Web site usage via cookies and other means to learn customer navigation behavior and product interests.

Table 3.4 Typical Internal Sources of Data on Which Business Analytics Can Be Based

Type of External Data	Measured By
Customer Satisfaction	• Revenue, profit • Market share, sales • Product/service survey data • Loyalty • Brand awareness • Average spend per customer
Customer Demographics	• Geographic location (distance from market) • Income level • Market size
Competition	• Market share • Competitor profitability • Advertising/promotion efforts • Preference surveys • Web chatter on products
Economic	• Population statistics • Income distribution statistics

Table 3.5 Typical External Sources of Data on Which Business Analytics Can Be Based

A major portion of the external data sources are found in the literature. For example, the *US Census* and the *International Monetary Fund* (IMF) are useful data sources at the macroeconomic level for model building. Likewise, audience and survey data sources might include *Nielsen* (**www.nielsen.com/us/en.html**), psychographic or demographic data sourced from *Claritas* (**www.claritas.com**), financial data from *Equifax* (**www.equifax.com**), Dun & Bradstreet (**www.dnb.com**), and so forth.

3.3.2. Data Issues

Regardless of the source of data, it has to be put into a structure that makes it usable by BA personnel. We will discuss data warehousing in the next section, but here we focus on a couple of data issues that are critical to the usability of any database or data file. Those issues are data quality and data privacy. *Data quality* can be defined as data that serves the purpose for which it is collected. It means different things for different applications, but there are some commonalities of high-quality data. These qualities usually include accurately representing reality, measuring what it is supposed to measure, being timeless, and having completeness. When data is of high quality, it helps ensure competitiveness, aids customer service, and improves profitability. When data is of poor quality, it can provide information that is contradictory, leading to misguided decision-making. For example, having missing data in files can prohibit some forms' statistical modeling, and incorrect coding of information can completely render databases useless. Data quality requires effort on the part of data managers to cleanse data of erroneous information and repair or replace missing data. We will discuss some of these quality data measures in later chapters.

Data privacy refers to the protection of shared data such that access is permitted only to those users for whom it is intended. It is a security issue that requires balancing the need to know with the risks of sharing too much. There are many risks in leaving unrestricted access to a company's database. For example, competitors can steal a firm's customers by accessing addresses. Data leaks on product quality failures can damage brand image, and customers can become distrustful of a firm that shares information given in confidence. To avoid these issues, a firm needs to abide by the current legislation regarding customer privacy and develop a program devoted to data privacy.

Collecting and retrieving data and computing analytics requires the use of computers and information technology. A large part of what BA personnel do is related to managing information systems to collect, process, store, and retrieve data from various sources.

3.4. Business Analytics Technology

Firms need an *information technology* (IT) *infrastructure* that supports personnel in the conduct of their daily business operations. The general requirements for such a system are stated in **Table 3.6**. These types of technology are elemental needs for business analytics operations.

Type of Technology	Description
Computer Hardware	This is physical equipment used for input, processing, and output activities in an information system. Hardware can include computers of various sizes, various input, output and storage devices; and telecommunications devices that link computers, including mobile handheld devices.
Computer Software	These are the preprogrammed instructions that control and coordinate the computer hardware components in the information system. They include system-wide software like ERP and smaller *apps* (computer software applications) for mobile devices.

Networking and Telecommunications Technology	Physical devices and software link the various pieces of hardware and transfer data from one physical location to another. They include the computers and communications equipment connected in networks for sharing voice, data, images, sound, and video. They also include the Internet, *intranets* (internal corporate networks based on Internet technology with limited access to employees within the firm), and *extranets* (private intranets extended to authorized users outside the organization).
Data Management Technology	Software governs the organization of data on physical storage media. It includes database management systems, data warehouses, data marts, and online analytical processing, as well as data, text, and Web mining technologies.

Table 3.6 General Information Technology (IT) Infrastructure

Of particular importance for BA is the data management technologies listed in **Table 3.6**. *Database management systems* (DBMS) is a data management technology software that permits firms to centralize data, manage it efficiently, and provide access to stored data by application programs. DBMS usually serves as an interface between application programs and the physical data files of structured data. DBMS makes the task of understanding where and how the data is actually stored more efficient. In addition, other DBMS systems can handle unstructured data. For example, *object-oriented DBMS systems* are able to store and retrieve unstructured data, like drawings, images, photographs, and voice data. These types of technology are necessary to handle the load of big data that most firms currently collect.

DBMS includes capabilities and tools for organizing, managing, and accessing data in databases. Four of the more important capabilities are its data definition language, data dictionary, database encyclopedia, and data manipulation language. DBMS has a *data definition* capability to specify the structure of content in a database. This is used to create database tables and characteristics used in fields to identify content. These tables and characteristics are critical success factors for search efforts as the database grows in size. These characteristics are documented in the *data dictionary* (an automated or manual file that stores the size, descriptions, format, and other properties needed to characterize data). The *database encyclopedia* is a table of contents listing a firm's current data inventory and what data files can be built or purchased. The typical content of the database encyclopedia is presented in **Table 3.7**. Of particular importance for BA is the *data manipulation language* tools included in DMBS. These tools are used to search databases for specific information. An example is *structure query language* (SQL), which allows users to find specific data through a session of queries and responses in a database.

Database Content Item	Description
Purpose	Why the database exists, including any additional reports or analyses used in leveraging the data.
Time	Window of time period when the data is collected or will be useful.
Source	Internal (auditing, accounting, and so on) and external (customers, and so on) sources.
Schematics	Diagrams illustrating the connections between tables and other data files.
Cost	Expense of collecting data, including purchasing prices.
Availability of Data	Window of time when the data may be available.
Collection Techniques	Methods of collection, including observation, data mining, census, and focus groups.
Collection Tools	Web, customer generated, e-survey, and so on.

Table 3.7 Database Encyclopedia Content

Data warehouses are databases that store current and historical data of potential interest to decision makers. What a data warehouse does is make data available to anyone who needs access to it. In a data warehouse, the data is prohibited from being altered. Data warehouses also provide a set of query tools, analytical tools, and graphical reporting facilities. Some firms use intranet portals to make data warehouse information widely available throughout a firm.

Data marts are focused subsets or smaller groupings within a data warehouse. Firms often build enterprise-wide data warehouses where a central data warehouse serves the entire organization and smaller, decentralized data warehouses (called data marts) are focused on a limited portion of the organization's data that is placed in a separate database for a specific population of users. For example, a firm might develop a smaller database on just product quality to focus efforts on quality customer and product issues. A data mart can be constructed more quickly and at lower cost than enterprise-wide data warehouses to concentrate effort in areas of greatest concern.

Once data has been captured and placed into database management systems, it is available for analysis with BA tools, including online analytical processing, as well as data, text, and Web mining technologies. *Online analytical processing* (OLAP) is software that allows users to view data in multiple dimensions. For example, employees can be viewed in terms of their age, sex, geographic location, and so on. OLAP would allow identification of the number of employees who are age 35, male, and in the western region of a country. OLAP allows users to obtain online answers to ad hoc questions quickly, even when the data is stored in very large databases.

Data mining is the application of a software, discovery-driven process that provides insights into business data by finding hidden patterns and relationships in big data or large databases and inferring rules from them to predict future behavior. The observed patterns and rules are used to guide decision-making. They can also act to forecast the impact of those decisions. It is an ideal predictive analytics tool used in the BA process mentioned in **Chapter 1**, "**What Are Business Analytics?**" The kinds of information obtained by data mining include those in **Table 3.8**.

Types of Information	Description	Example
Associations	Occurrences linked to a single event.	An ad in a newspaper is associated with greater sales.

Classification	Recognizes patterns that describe the group an item belongs to by examining previous classified existing items and by inferring a set of rules that guide the classification process.	Identify customers who are likely to need more customer service than those who need less.
Clustering	Similar to classification when no groups have yet been defined, helps to discover different groupings within data.	Identify groups that can be differentiated within a single, large group of customers. An example would be identifying tea drinkers who choose that beverage from others offered in flight from an airline.
Forecasting	Predicts values that can identify patterns in customer behavior.	Estimate the value of a future stream of dollar sales from a typical customer.
Sequence	Links events over time.	Identify a link between a person who buys a new house and subsequently will buy a new car within 90 days.

Table 3.8 Types of Information Obtainable with Data Mining Technology

Text mining (mentioned in **Chapter 2**) is a software application used to extract key elements from unstructured data sets, discover patterns and relationships in the text materials, and summarize the information. Given that the majority of the information stored in businesses is in the form of unstructured data (emails, pictures, memos, transcripts, survey responses, business receipts, and so on), the need to explore and find useful information will require increased use of text mining tools in the future.

Web mining seeks to find patterns, trends, and insights into customer behavior from users of the Web. Marketers, for example, use BA services like *Google Trends* (**www.google.com/trends/**) and *Google Insights for Search* (**http://google.about.com/od/i/g/google-insights-for-search.htm**) to track the popularity of various words and phrases to learn what consumers are interested in and what they are buying.

In addition to the general software applications discussed earlier, there are focused software applications used every day by BA analysts in conducting the three steps of the BA process (see **Chapter 1**). These include *Microsoft Excel*® spreadsheet applications, SAS applications, and SPSS applications. *Microsoft Excel* (**www.microsoft.com/**) spreadsheet systems have add-in applications specifically used for BA analysis. These add-in applications broaden the use of Excel into areas of BA. *Analysis ToolPak* is an Excel add-in that contains a variety of statistical tools (for example, graphics and multiple regression) for the descriptive and predictive BA process steps. Another Excel add-in, *Solver*, contains operations research optimization tools (for example, linear programming) used in the prescriptive step of the BA process.

SAS® Analytics Pro (**www.sas.com/**) software provides a desktop statistical toolset allowing users to access, manipulate, analyze, and present information in visual formats. It permits users to access data from nearly any source and transform it into meaningful, usable information presented in visuals that allow decision makers to gain quick understanding of critical issues within the data. It is designed for use by analysts, researchers, statisticians, engineers, and scientists who need to explore, examine, and present data in an easily understandable way and distribute findings in a variety of formats. It is a statistical package chiefly useful in the descriptive and predictive steps of the BA process.

IBM's *SPSS software* (**www-01.ibm.com/software/analytics/spss/**) offers users a wide range of statistical and decision-making tools. These tools include methodologies for data collection, statistical manipulation, modeling trends in structured and unstructured data, and optimizing analytics. Depending on the statistical packages acquired, the software can cover all three steps in the BA process.

Other software applications exist to cover the prescriptive step of the BA process. One that will be used in this book is LINGO® by Lindo Systems (**www.lindo.com**). LINGO is a comprehensive tool designed to make building and solving optimization models faster, easier, and more efficient. LINGO provides a completely integrated package that includes an understandable language for expressing optimization models, a full-featured environment for building and editing problems, and a set of built-in solvers to handle optimization modeling in linear, nonlinear, quadratic, stochastic, and integer programming models.

In summary, the technology needed to support a BA program in any organization will entail a general information system architecture, including database management systems and progress in greater specificity down to the software that BA analysts need to compute their unique contributions to the organization. Organizations with greater BA requirements will have substantially more technology to support BA efforts, but all firms that seek to use BA as a strategy for competitive advantage will need a substantial investment in technology, because BA is a technology-dependent undertaking.

Summary

Why BA is important is directly proportional to what it costs. In this chapter, we have explored costs, but also many of the benefits of BA as a means to justify why a BA program is necessary. This chapter discussed what resources a firm would need to support a BA program. From this, three primary areas of resources were identified: personnel, data, and technology. Having identified BA personnel and needed skill sets, a review of content in BA certification exams was presented. Types of personnel specialties also were discussed. BA data internal and external sources were presented as a means of data categorization. Finally, BA technology was covered in terms of general, organization-wide information systems support to individual analyst support software packages.

In this chapter, we focused on the investment in resources needed to have a viable business analytics operation. In **Chapter 4**, we begin **Part III**, "**How Can Business Analytics Be Applied?**" Specifically, in the next chapter we will focus on how the resources mentioned in this chapter are placed into an organization and managed to achieve goals.

Discussion Questions

1. How does using BA certification exam content explain skill sets for BA analysts? What skill sets are necessary for BA personnel?

2. Why is leadership an important skill set for individuals looking to make a career in BA?

3. Why is categorizing data from its sources important in BA?

4. What is data quality, and why is it important in BA?

5. What is the difference between a data warehouse and a datamart?

4. How Do We Align Resources to Support Business Analytics within an Organization?

Chapter objectives:

• Explain why a centralized business analytics (BA) organization structure has advantages over other structures.

• Describe the differences between BA programs, projects, and teams and how they are used to align BA resources in firms.

• Describe reasons why BA initiatives fail.

• Describe typical BA team roles and reasons for their failures.

• Explain why establishing an information policy is important.

• Explain the advantages and disadvantages of outsourcing BA.

• Describe how data can be scrubbed.

• Explain what change management involves and what its relationship is to BA.

4.1. Organization Structures Aligning Business Analytics

According to **Isson and Harriott** (**2013**, p. 124), to successfully implement business analytics (BA) within organizations, the BA in whatever organizational form it takes must be fully integrated throughout a firm. This requires BA resources to be aligned in a way that permits a view of customer information within and across all departments, access to customer information from multiple sources (internal and external to the organization), access to historical analytics from a central repository, and making technology resources align to be accountable for analytic success. The commonality of these requirements is the desire for an alignment that maximizes the flow of information into and through the BA operation, which in turn processes and shares information to desired users throughout the organization. Accomplishing this information flow objective requires consideration of differing organizational structures and managerial issues that help align BA resources to best serve an organization.

4.1.1. Organization Structures

As mentioned in **Chapter 2**, "**Why Are Business Analytics Important?**", most organizations are hierarchical, with senior managers making the strategic planning decisions, middle-level managers making tactical planning decisions, and lower-level managers making operational planning decisions. Within the hierarchy, other organizational structures exist to support the development and existence of groupings of resources like those needed for BA. These additional structures include programs, projects, and teams. A *program* in this context is the process that seeks to create an outcome and usually involves managing several related projects with the intention of improving organizational performance. A program can also be a large project. A *project* tends to deliver outcomes and can be defined as having temporary rather than

permanent social systems within or across organizations to accomplish particular and clearly defined tasks, usually under time constraints. Projects are often composed of teams. A *team* consists of a group of people with skills to achieve a common purpose. Teams are especially appropriate for conducting complex tasks that have many interdependent subtasks.

The relationship of programs, projects, and teams with a business hierarchy is presented in **Figure 4.1**. Within this hierarchy, the organization's senior managers establish a *BA program* initiative to mandate the creation of a BA grouping within the firm as a strategic goal. A BA program does not always have an end-time limit. Middle-level managers reorganize or break down the strategic BA program goals into doable *BA project* initiatives to be undertaken in a fixed period of time. Some firms have only one project (establish a BA grouping) and others, depending on the organization structure, have multiple BA projects requiring the creation of multiple BA groupings. Projects usually have an end-time date in which to judge the successfulness of the project. The projects in some cases are further reorganized into smaller assignments, called *BA team* initiatives, to operationalize the broader strategy of the BA program. BA teams may have a long-standing time limit (for example, to exist as the main source of analytics for an entire organization) or have a fixed period (for example, to work on a specific product quality problem and then end).

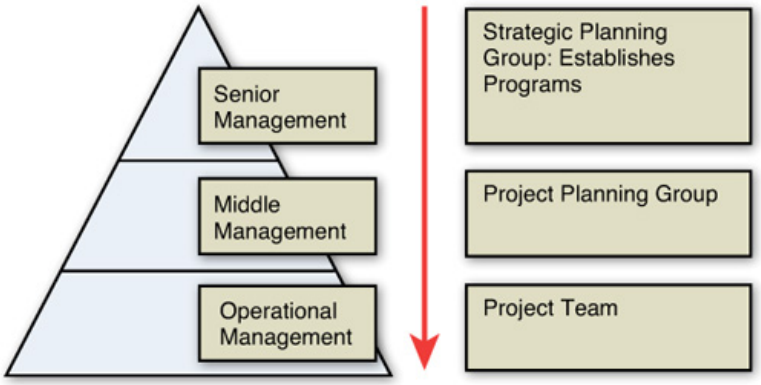


Figure 4.1 Hierarchal relationships program, project, and team planning

In summary, one way to look at the alignment of BA resources is to view it as a progression of assigned planning tasks from a BA program, to BA projects, and eventually to BA teams for implementation. As shown in **Figure 4.1**, this hierarchical relationship is a way to examine how firms align planning and decision-making workload to fit strategic needs and requirements.

BA organization structures usually begin with an initiative that recognizes the need to use and develop some kind of program in analytics. Fortunately, most firms today recognize this need. The question then becomes how to match the firm's needs within the organization to achieve its strategic, tactical, and operations objectives within resource limitations. Planning the BA resource allocation within the organizational structure of a firm is a starting place for the alignment of BA to best serve a firm's needs.

Aligning the BA resources requires a determination of the amount of resources a firm wants to invest. The outcome of the resource investment might identify only one individual to compute analytics for a

firm. Because of the varied skill sets in information systems, statistics, and operations research methods, a more common beginning for a BA initiative is the creation of a BA team organization structure possessing a variety of analytical and management skills. (We will discuss BA teams in **Section 4.1.2**.) Another way of aligning BA resources within an organization is to use a project structure. Most firms undertake projects, and some firms actually use a project structure for their entire organization. For example, consulting firms might view each client as a project (or product) and align their resources around the particular needs of that client. A project structure often necessitates multiple BA teams to deal with a wider variety of analytic needs. Even larger investments in BA resources might be required by firms that decide to establish a whole BA department containing all the BA resources for a particular organization. Although some firms create BA departments, the departments don't have to be large. Whatever the organization structure that is used, the role of BA is a staff (not line management) role in their advisory and consulting mission for the firm.

In general, there are different ways to structure an organization to align its BA resources to serve strategic plans. In organizations where functional departments are structured on a strict hierarchy, separate BA departments or teams have to be allocated to each functional area, as presented in **Figure 4.2**. This *functional organization structure* may have the benefit of stricter functional control by the VPs of an organization and greater efficiency in focusing on just the analytics within each specialized area. On the other hand, this structure does not promote the cross-department access that is suggested as a critical success factor for the implementation of a BA program.

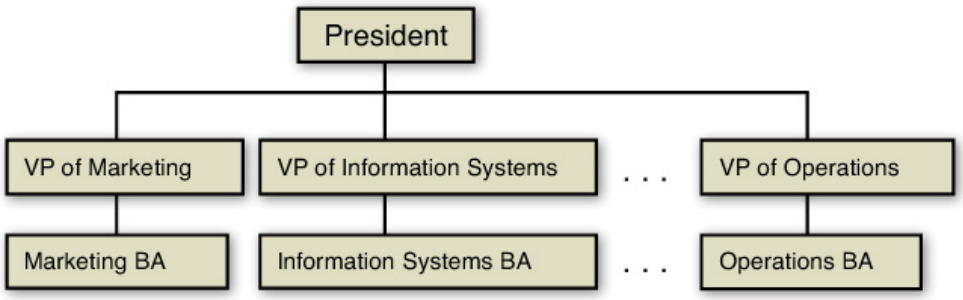


Figure 4.2 Functional organization structure with BA

The needs of each firm for BA sometimes dictate positioning BA within existing organization functional areas. Clearly, many alternative structures can house a BA grouping. For example, because BA provides information to users, BA could be included in the functional area of management information systems, with the *chief information officer* (CIO) acting as both the director of information systems (which includes database management) and the leader of the BA grouping.

An alternative organizational structure commonly found in large organizations aligns resources by project or product and is called a *matrix organization*. As illustrated in **Figure 4.3**, this structure allows the VPs some indirect control over their related specialists, which would include the BA specialists but also allows direct control by the project or product manager. This, similar to the functional organizational structure, does not promote the cross-department access suggested for a successful implementation of a BA program.

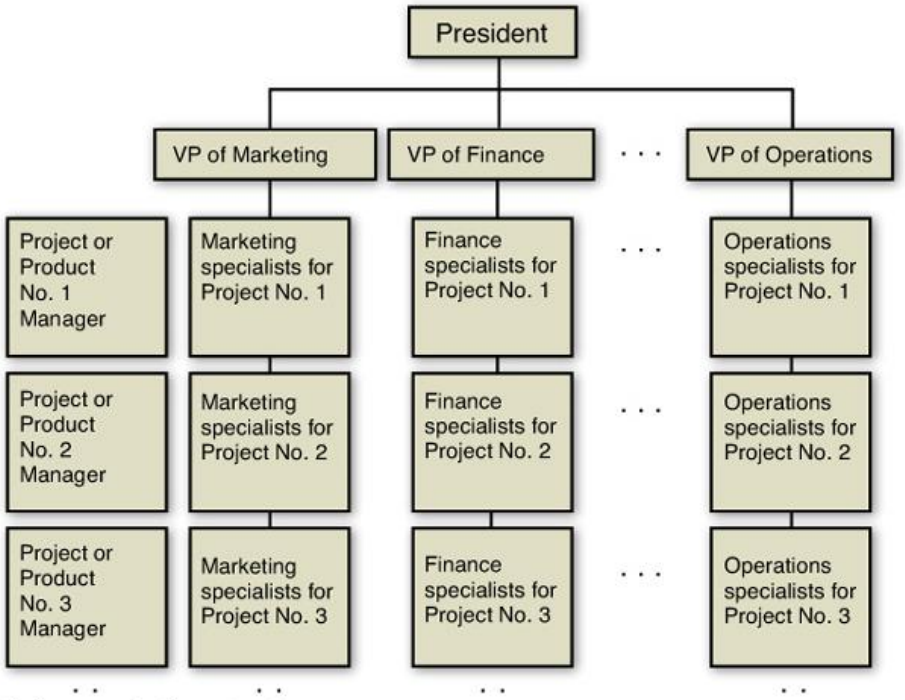


Figure 4.3 Matrix organization structure

The literature suggests that the organizational structure that best aligns BA resources is one in which a department, project, or team is formed in a staff structure where access to and from the BA grouping of resources permits access to all areas within a firm, as illustrated in **Figure 4.4** (**Laursen and Thorlund, 2010**, pp. 191–192; **Bartlett, 2013**, pp. 109–111; **Stubbs, 2011**, p. 68). The dashed line indicates a staff (not line management) relationship. This *centralized BA organization structure* minimizes investment costs by avoiding duplications found in both the functional and the matrix styles of organization structures. At the same time, it maximizes information flow between and across functional areas in the organization. This is a logical structure for a BA group in its advisory role to the organization. **Bartlett** (**2013**, pp. 109–110) suggests other advantages of a centralized structure like the one in **Figure 4.4**. These include a reduction in the filtering of information traveling upward through the organization, insulation from political interests, breakdown of the *siloed functional area* communication barriers, a more central platform for reviewing important analyses that require a broader field of specialists, analytic-based group decision-making efforts, separation of the line management leadership from potential clients (for example, the VP of marketing would not necessarily come between the BA group working on customer service issues for a department within marketing), and better connectivity between BA and all personnel within the area of problem solving.

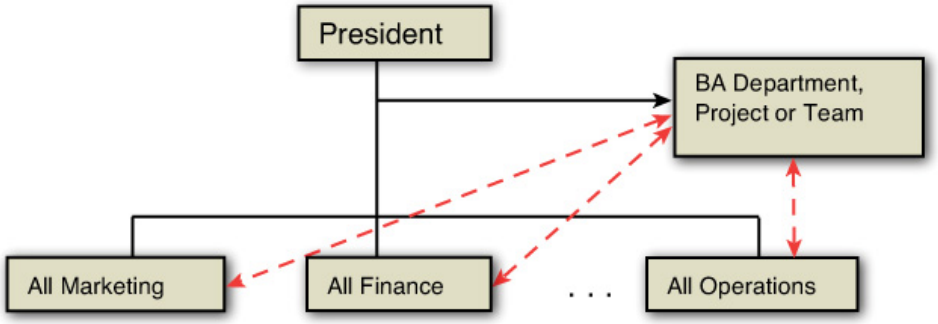


Figure 4.4 Centralized BA department, project, or team organization structure

Given the advocacy and logic recommending a centralized BA grouping, there are reasons for all BA groupings to be centralized. These reasons help explain why BA initiatives that seek to integrate and align BA resources into any type of BA group within the organization sometimes fail. The listing in Table 4.1 is not exhaustive, but it provides some of the important issues to consider in the process of structuring a BA group.

Reason	Description
Lack of Executive Sponsorship	Senior executive failure to recognize the value of BA and its importance eventually leads to a reduction in resources and eventual failure.
Limited Context Perception	There is an incorrect perception that analytics must be applied within a particular functional area in order to have the necessary validity to be applied to that area. Example: Financial regression analysis can only be applied correctly in the context of the finance area.
Belief of Physical Proximity	There is misperception that it takes physical proximity of the BA grouping in the business application area to be valid.
Lack of Leadership in BA Groupings	Without an advocate leader in the organization, as well as leaders in BA projects and teams to move the analysis to achieve desired goals, the entire BA effort will lead to eventual failure.
Lack of support	Without support for needed personnel, collecting data and technology to process the data will lead to failure.
Lack of Collaboration Across All Organizational Groups	Analytics that solve problems across multiple, functional areas are more likely to be accepted and successful than those that lack the cross-over into multiple organizational groups.
Lack of Skilled and Human Resources	BA departments, projects, or teams that don't have the skilled personnel to deal with the execution of analysis will eventually cause the failure of BA.
Inability to Delegate Responsibility	There is a desire to delegate responsibility to solve problems locally (a matter of trusting your own) rather than seeking help throughout the organization. This impedes the flow of problem solving efforts by an external BA department and impedes communication of information needed to successfully apply BA.
Lack of Integrated Processes	Information that is stored in silos and not shared makes it more difficult for BA analysis to succeed.

Table 4.1 Reasons for BA Initiative and Organization Failure

In summary, the organizational structure that a firm may select for the positioning of their BA grouping can either be aligned within an existing organizational structure, or the BA grouping can be separate, requiring full integration within all areas of an organization. While some firms may start with a number of small teams to begin their BA program, other firms may choose to start with a full-sized BA department. Regardless of the size of the investment in BA resources, it must be aligned to allow maximum information flow between and across functional areas to achieve the most benefits BA can deliver.

4.1.2. Teams

When it comes to getting the BA job done, it tends to fall to a BA team. For firms that employ BA teams the participants can be defined by the roles they play in the team effort. Some of the roles BA team participants undertake and their typical background are presented in Table 4.2.

Title or Function	Role Description	Background or Skills of Participant
Analytics Modeler	Develop and maintain predictive and forecasting models to provide insight.	Statistics, operation research, analytic modeling.
Analytics Process Designer	Develop and enforce reusable processes to reduce BA execution time.	Management consultant, process mapping, systems design.
Analytics Analyst	Respond to BA inquiries from functional areas within the firm to gain insight.	Reporting, problem solving, communicating, and providing customer service.
BA Team Head	Provide leadership to BA team, define strategies and tactics to ensure improved business performance, and interface with management.	BA manager or administrator.
Business Domain Expert	Provide business experience to ensure relevance of insight, help interpret business measures and the meaning of data.	Business experience in the area where the problem or opportunity exists.
Data Manager	Ensure data availability and access while minimizing costs.	Data modeling or warehousing, experience in data quality processes.
Implementation Specialist	Ensure rapid and robust model deployment to reduce time in interface.	Information system and data warehousing expertise, enterprise architecture experience.
Monitoring Analyst	Identify, establish, and enforce common analytics to be used to measure value and optimize effort.	Management and BA expert, predictive and financial modeling, process design, and team mentoring.

Source: Adapted from **Stubbs (2013)**, pp.137–149; **Stubbs (2011)** Table 3.3; **Laursen and Thorlund (2010)**, p.15.

Table 4.2 BA Team Participant Roles*

Aligning BA teams to achieve their tasks requires collaboration efforts from team members and from their organizations. Like BA teams, *collaboration* involves working with people to achieve a shared and explicit set of goals consistent with their mission. BA teams also have a specific mission to complete. Collaboration through teamwork is the means to accomplish their mission.

Team members' need for collaboration is motivated by changes in the nature of work (no more silos to hide behind, much more open environment, and so on), growth in professions (for example, interactive jobs tend to be more professional, requiring greater variety in expertise sharing), and the need to nurture innovation (creativity and innovation are fostered by collaboration with a variety of people sharing ideas). To keep one's job and to progress in any business career, particularly in BA, team members must encourage working with other members inside a team and out. For organizations, collaboration is motivated by the changing nature of information flow (that is, hierarchical flows tend to be downward, whereas in modern organizations, flow is in all directions) and changes in the scope of business operations (that is, going from domestic to global allows for a greater flow of ideas and information from multiple sources in multiple locations).

How does a firm change its culture of work and business operations to encourage collaboration? One way to affect the culture is to provide the technology to support a more open, cross-departmental information flow. This includes e-mail, instant messaging, *wikis* (collaboratively edited works, like Wikipedia), use of social media and networking through *Facebook* and *Twitter*, and encouragement of activities like collaborative writing, reviewing, and editing efforts. Other technology supporting collaboration includes webinars, audio and video conferencing, and even the use of iPads to enhance face-to-face communication. These can be tools that change the culture of a firm to be more open and communicative.

Reward systems should be put into place to reward team effort. Teams should be rewarded for their performance, and individuals should be rewarded for performance in a team. While middle-level managers build teams, coordinate their work, and monitor their performance, senior management should establish collaboration and teamwork as a vital function.

Despite the collaboration and best of intentions, BA teams sometimes fail. There are many reasons for this, but knowing some of the more common ones can help managers avoid them. Some of the more common reasons for team failure are presented in **Table 4.3**. They also represent issues that can cause a BA program to become unaligned and unproductive.

Reason for Failure	Descriptions
Lack of Communication	It is not enough to come up with valuable information for decision-making and to find business opportunities in data. That information must be shared with users, clients, and everyone within a firm for benefit to come from it. It is only when analytics show a tangible and beneficial outcome that they are considered business analytics (BA). If those results are not communicated on a continual basis, BA teams can be perceived to provide less value to the organization.
Failure to Deliver	Not every BA team will be able to deliver valued information if the team lacks the ability or resources to deliver needed answers and information. The greater the number of BA team failures, the greater are the chances that the team will be eliminated.
Lack of Justification	BA teams require resource allocations. Those allocations come from other departments that supposedly benefit from BA contributions. Without the role of BA and its potential contributions to a firm being clearly spelled out, users might not associate the ongoing efforts of a BA team as being worth the money spent on them.
Fail to Provide Value	BA teams have to sell their roles and suggested solutions or ideas. Without a clear understanding of value for potential users, the team faces a hard sell.
Inability to Prove Success	BA teams need to document and measure the impact of their ideas and suggestions. Without that proof, potential users might not support future BA efforts.

Source*: Adapted from **Flynn (2008) pp. 99–106 and **Stubbs (2011)** p. 89.

Table 4.3 Reasons for BA Team Failures*

4.2. Management Issues

Aligning organizational resources is a management function. There are general management issues that are related to a BA program, and some are specifically important to operating a BA department, project, or team. The ones covered in this section include establishing an information policy, outsourcing business analytics, ensuring data quality, measuring business analytics contribution, and managing change.

4.2.1. Establishing an Information Policy

There is a need to manage information. This is accomplished by establishing an *information policy* to structure rules on how information and data are to be organized and maintained and who is allowed to view the data or change it. The information policy specifies organizational rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying all types of information and data. It defines the specific procedures and accountabilities that identify which users and organizational units can share information, where the information can be distributed, and who is responsible for updating and maintaining the information.

In small firms, business owners might establish the information policy. For larger firms, *data administration* may be responsible for the specific policies and procedures for data management (Siegel and Shim, 2003, p. 280). Responsibilities could include developing the information policy, planning data collection and storage, overseeing database design, developing the data dictionary, as well as monitoring how information systems specialists and end user groups use data.

A more popular term for many of the activities of data administration is *data governance*, which includes establishing policies and processes for managing the availability, usability, integrity, and security of the data employed in businesses. It is specifically focused on promoting data privacy, data security, data quality, and compliance with government regulations.

Such information policy, data administration, and data governance must be in place to guard and ensure data is managed for the betterment of the entire organization. These steps are also important in the creation of database management systems (see **Chapter 3**, "**What Resource Considerations Are Important to Business Analytics?**") and their support of BA tasks.

4.2.2. Outsourcing Business Analytics

Outsourcing can be defined as a strategy by which an organization chooses to allocate some business activities and responsibilities from an internal source to an external source (**Schniederjans, et al., 2005**, pp. 3–4). Outsourcing business operations is a strategy that an organization can use to implement a BA program, run BA projects, and operate BA teams. Any business activity can be outsourced, including BA. Outsourcing is an important BA management activity that should be considered as a viable alternative in planning an investment in any BA program.

BA is a staff function that is easier to outsource than other line management tasks, such as running a warehouse. To determine if outsourcing is a useful option in BA programs, management needs to balance the advantages of outsourcing with its disadvantages. Some of the advantages of outsourcing BA include those listed in **Table 4.4**.

Advantage of Outsourcing BA	Description
Less Expensive	Maintaining a fully functioning BA department when analytics might only be useful periodically may be more expensive than occasionally hiring an outside consulting BA firm to solve a problem.
Superior Analytics	The pool of analytic capabilities is always going to be greater outside a firm.
More Staffing Flexibility	Staff positions are often the first cut in economy downturns. Using consultants is easier and less expensive than hiring full-time BA staff. Outsourcing permits more flexibility to add and reduce BA services as needed.
New Knowledge	Experienced BA consultants bring a variety of knowledge and experience from having worked with many other firms. That type of experience may be of great competitive advantage.

Table 4.4 Advantages of Outsourcing BA

Nevertheless, there are disadvantages of outsourcing BA. Some of the disadvantages to outsourcing are presented in **Table 4.5**.

Disadvantages of Outsourcing BA	Description
Loss of Control	Once outsourced, most of the control of a BA project is under the control of the outsourcing firm. The client firm might lose not only control, but also opportunities for new and unique information, which the outsourcing firm may not share with the client.
Difficulties in Managing the Relationship	Client firms may find it difficult to communicate with the outsourcing firm because of distance, differing culture, language issues, and more. The lack of management could cause substantial problems with customer service and product quality.
Weakens Innovation	Having outsourced a client firm's internal experts, the remaining collaboration within the firm's personnel is reduced, and that reduces the opportunity for innovation efforts through shared collaboration.
Risk of Information	Outsourcing staff are exposed to client proprietary information, including innovations in analytics. This information could be shared with other competing firms, placing the client firm at risk.
Worthless Analytics	Sometimes outsourcing partners are less capable than internal analysts, wasting time and money.

Table 4.5 Disadvantages of Outsourcing BA

Managing outsourcing of BA does not have to involve the entire department. Most firms outsource projects or tasks found to be too costly to assign internally. For example, firms outsource cloud computing services to outside vendors (Laudon and Laudon, 2012, p. 511), and other firms outsource software development or maintenance of legacy programs to offshore firms in low-wage areas of the world to cut costs (Laudon and Laudon, 2012, p. 192).

Outsourcing BA can also be used as a strategy to bring BA into an organization (**Schniederjans, et al., 2005**, pp. 24–27). Initially, to learn how to operate a BA program, project, or team, an outsource firm can be hired for a limited, contracted time period. The client firm can then learn from the outsourcing firm's

experience and instruction. Once the outsourcing contract is over, the client firm can form its own BA department, project, or team.

4.2.3. Ensuring Data Quality

Business analytics, if relevant, is based on data assumed to be of high quality. *Data quality* refers to accuracy, precision, and completeness of data. High-quality data is considered to correctly reflect the real world in which it is extracted. Poor quality data caused by data entry errors, poorly maintained databases, out-of-date data, and incomplete data usually leads to bad decisions and undermines BA within a firm. Organizationally, the database management systems (DBMS, mentioned in **Chapter 3**) personnel are managerially responsible for ensuring data quality. Because of its importance and the possible location of the BA department outside of the management information systems department (which usually hosts the DBMS), it is imperative that whoever leads the BA program should seek to ensure data quality efforts are undertaken.

Ideally, a properly designed database with organization-wide data standards and efforts taken to avoid duplication or inconsistent date elements should have high-quality data. Unfortunately, times are changing, and more organizations allow customers and suppliers to enter data into databases via the Web directly. As a result, most of the quality problems originate from data input such as misspelled names, transposed numbers, or incorrect or missing codes.

An organization needs to identify and correct faulty data and establish routines and procedures for editing data in the database. The analysis of data quality can begin with a *data quality audit*, where a structured survey or inspection of accuracy and level of completeness of data is undertaken. This audit may be of the entire database, just a sample of files, or a survey of end users for perceptions of the data quality. If during the data quality audit files are found that have errors, a process called *data cleansing* or *data scrubbing* is undertaken to eliminate or repair data. Some of the areas in a data file that should be inspected in the audit and suggestions on how to correct them are presented in **Table 4.6**.

Data Inspection Items	Description and Cleansing/Scrubbing Recommendation
Current Data	Check to make sure the data is current. If it is out of date, remove it.
Completeness	Check to see if there is missing data. If more than 50% is missing, remove the entire file from the database.
Relevance	Check to see if the data is no longer relevant for the purpose for which it was collected. If it's no longer relevant, consider removing it from the database.
Duplication	Check to see if duplicate data files exist in the database. Remove duplicate data.
Outliers	Check for extreme values (outliers) in quantitative data files for possible errors in data coding. Remove from the data file any suspected of being in error, or repair the data.
Inconsistent Values	If data fields contain both characters and real numbers data where only characters or numbers should be, explore repairing the data.
Coding	If suspicious or unknown coding of data exists in data files, remove from the database or repair the coding of data.

Table 4.6 Quality Data Inspection Items and Recommendations

4.2.4. Measuring Business Analytics Contribution

The investment in BA must continually be justified by communicating the BA contribution to the organization for ongoing projects. This means that performance analytics should be computed for every BA project and BA team initiative. These analytics should provide an estimate of the tangible and intangible values being delivered to the organization. This should also involve establishing a communication strategy to promote the value being estimated.

Measuring the value and contributions that BA brings to an organization is essential to helping the firm understand why the application of BA is worth the investment. Some BA contribution estimates can be computed using standard financial methods, such as *payback period* (how long it takes for the initial costs are returned by profit) or return on investment (ROI) (see Schniederjans, et al., 2010, pp. 90–132), where dollar values or quantitative analysis is possible. When intangible contributions are a major part of the contribution being delivered to the firm, other methods like cost/benefit analysis (see Schniederjans, et al., 2010, pp. 143–158), which include intangible benefits, should be used.

The continued measurement of value that BA brings to a firm is not meant to be self-serving, but it aids the organization in aligning efforts to solve problems and find new business opportunities. By continually running BA initiatives, a firm is more likely to identify internal activities that should and can be enhanced by employing optimization methodologies during the Prescriptive step of the BA process introduced in **Chapter 1**, "**What Are Business Analytics?**" It can also help identify underperforming assets. In addition, keeping track of investment payoffs for BA initiatives can identify areas in the organization that should have a higher priority for analysis. Indeed, past applications and allocations of BA resources that have shown significant contributions can justify priorities established by the BA leadership about where there should be allocated analysis efforts within the firm. They can also help acquire increases in data support, staff hiring, and further investments in BA technology.

4.2.5. Managing Change

Wells (2000) found that what is critical in changing organizations is organizational culture and the use of change management. *Organizational culture* is how an organization supports cooperation, coordination, and empowerment of employees (**Schermerhorn 2001**, p. 38). *Change management* is defined as an approach for transitioning the organization (individuals, teams, projects, departments) to a changed and desired future state (Laudon and Laudon, 2012, pp. 540–542). Change management is a means of implementing change in an organization, such as adding a BA department (**Schermerhorn 2001**, pp. 382–390). Changes in an organization can be either planned (a result of specific and planned efforts at change with direction by a change leader) or unplanned (spontaneous changes without direction of a change leader). The application of BA invariably will result in both types of changes because of BA's specific problem-solving role (a desired, planned change to solve a problem) and opportunity finding exploratory nature (i.e., unplanned new knowledge opportunity changes) of BA. Change management can also target almost everything that makes up an organization (see **Table 4.7**).

Change Target	Description
Culture	This represents the changing values and norms of the individuals and groups that make up the organization. BA has to sell itself in some situations, build trust, and alter decision-making. It often requires a different culture of thinking about decision-making.

Organization Structure	This is the changing organizational lines of authority and communication. The cross-departmental nature of BA positions may provide information that changes the organization and alters relationships and tasks.
Personnel	BA information about the need for human resource changes in attitudes and skills can mandate changes that permit an organization to achieve higher business performance levels.
Tasks	BA analysis might find that some job designs, specifications, and descriptions that employees perform need to have their objectives and goals changed to achieve higher business performance levels.
Technology	BA analysis might find information system technology used in the design and workflow that integrate employees and equipment into operating systems and require change to achieve higher business performance levels.

**Source*: Adapted from Figure 7 in Schniederjans and Cao (2002), pp. 261.

Table 4.7 Change Management Targets*

It is not possible to gain the benefits of BA without change. The intent is change that involves finding new and unique information on which change should take place in people, technology systems, or business conduct. By instituting the concept of change management within an organization, a firm can align resources and processes to more readily accept changes that BA may suggest. Instituting the concept of change management in any firm depends on the unique characteristics of that firm. There are, though, a number of activities in common with successful change management programs, and they apply equally to changes in BA departments, projects, or teams. Some of these activities that lead to change management success are presented as best practices in **Table 4.8**.

Best Practice	Description
Champion	Change is scary business for some, and a strong leader for change can champion the change effort, calming fears and explaining the need for change. The champion also helps direct efforts, motivate change, and keep the change activities on track.
Clearly Stated Goals	Any type of change should be clearly defined, including what the changes are, which personnel have to change, and what the processes involve and how they affect technology. This would also include deadlines needed to keep the change effort on track.
Good Communication	To avoid resistance to change (a natural norm to anything that is new), it is useful to help those facing the change understand its value through effective and repeated communications, keeping them informed on progress and easing fears.
Measured Performance	Any goals stated prior to the launch of change can be used to measure performance during the changeover period. Seeing business performance improve with changes can motivate further change and support by those impacted.
Senior Management Support	Critical to all BA departments, projects, or teams is the need for senior management to support change efforts. Sometimes that support is in direct dollars, and sometimes it's in lending authority to get resources needed for BA work.

Table 4.8 Change Management Best Practices

Summary

Structuring a BA department, undertaking a BA project, or setting up a BA team within an organization can largely determine successfulness in aligning resources to achieve information-sharing goals. In this chapter, several organization structures (functional, matrix, and centralized) were discussed as possible homes for BA resource groupings. The role of BA teams as an important organizational resource aligning tool was also presented. In addition, this chapter discussed reasons for BA organization and team failures. Other managerial issues included in this chapter were establishing an information policy, outsourcing business analytics, ensuring data quality, measuring business analytics contribution, and managing change.

Once a firm has set up the internal organization for a BA department, program, or project, the next step is to undertake BA. In the next chapter, we begin the first of the three chapters devoted to detailing how to undertake the three steps of the BA process.

Discussion Questions

1. The literature in management information systems consistently suggests that a decentralized approach to resource allocation is the most efficient. Why then do you think the literature in BA suggests that the opposite—a centralized organization—is the best structure?

2. Why is collaboration important to BA?

3. Why is organization culture important to BA?

4. How does establishing an information policy affect BA?

5. Under what circumstances is outsourcing BA good for the development of BA in an organization?

6. Why do we have to measure BA contributions to an organization?

7. How does data quality affect BA?

8. What role does change management play in BA?

5. What Are Descriptive Analytics?

Chapter objectives:

• Explain why we need to visualize and explore data.

• Describe statistical charts and how to apply them.

• Describe descriptive statistics useful in the descriptive business analytics (BA) process.

• Describe the differences in SPSS descriptive software printouts from those covering the comparable subject in Excel.

• Describe sampling methods useful in BA and where to apply them.

• Describe what sampling estimation is and how it can aid in the BA process.

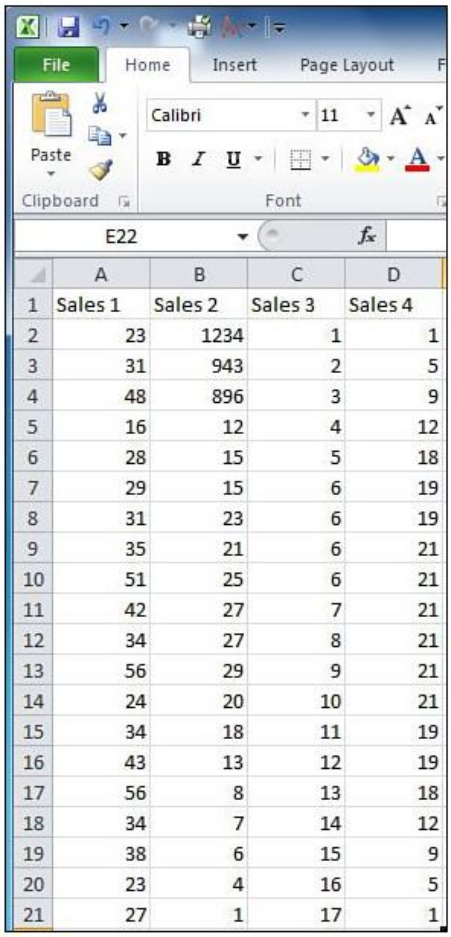
• Describe the use of confidence intervals and probability distributions.

• Explain how to undertake the descriptive analytics step in the BA process.

5.1. Introduction

In any BA undertaking, referred to as *BA initiatives* or *projects*, a set of objectives is articulated. These objectives are a means to align the BA activities to support strategic goals. The objectives might be to seek out and find new business opportunities, to solve operational problems the firm is experiencing, or to grow the organization. It is from the objectives that exploration via BA originates and is in part guided. The directives that come down, from the strategic planners in an organization to the BA department or analyst, focus the tactical effort of the BA initiative or project. Maybe the assignment will be one of exploring internal marketing data for a new marketing product. Maybe the BA assignment will be focused on enhancing service quality by collecting engineering and customer service information. Regardless of the type of BA assignment, the first step is one of exploring data and revealing new, unique, and relevant information to help the organization advance its goals. Doing this requires an exploration of data.

This chapter focuses on how to undertake the first step in the BA process: *descriptive analytics*. The focus in this chapter is to acquaint readers with more common descriptive analytic tools used in this step and available in SPSS and Excel software. The treatment here is not computational but informational regarding the use and meanings of these analytic tools in support of BA. For purposes of illustration, we will use the data set in **Figure 5.1** representing four different types of product sales (Sales 1, Sales 2, Sales 3, and Sales 4).



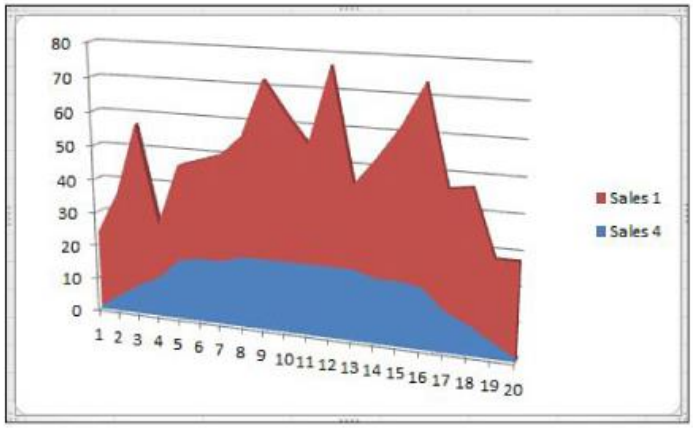
	A	B	C	D
1	Sales 1	Sales 2	Sales 3	Sales 4
2	23	1234	1	1
3	31	943	2	5
4	48	896	3	9
5	16	12	4	12
6	28	15	5	18
7	29	15	6	19
8	31	23	6	19
9	35	21	6	21
10	51	25	6	21
11	42	27	7	21
12	34	27	8	21
13	56	29	9	21
14	24	20	10	21
15	34	18	11	19
16	43	13	12	19
17	56	8	13	18
18	34	7	14	12
19	38	6	15	9
20	23	4	16	5
21	27	1	17	1

Figure 5.1 Illustrative sales data sets

5.2. Visualizing and Exploring Data

There is no single best way to explore a data set, but some way of conceptualizing what the data set looks like is needed for this step of the BA process. Charting is often employed to visualize what the data might reveal.

When determining the software options to generate charts in SPSS or Excel, consider that each software can draft a variety of charts for the selected variables in the data sets. Using the data in **Table 5.1**, charts can be created for the illustrative sales data sets. Some of these charts are discussed in **Table 5.1** as a set of exploratory tools that are helpful in understanding the informational value of data sets. The chart to select depends on the objectives set for the chart.

Type of Chart	Application Notes	Chart Example
Area	• Overlay more than one variable at a time. • Ideal for contrasting two variables. • Example: Overlaying different product sales to show improvement. • Note also the 3D effect, which is possible with most of the charts listed in this table.	
Bar	• Can be horizontal, vertical, cone, or cyclically shaped and multidimensional with overlaying variables. • Ideal for showing comparative improvement over time. • Example: Bars showing productivity of one person versus another.	
Column	• Same as a bar chart. • Note how this chart clearly reveals a positive trend upward.	
Line	• Ideal for showing linear trend and other linear or nonlinear appearance. • Best applied with time-series data with time as the X-axis.	

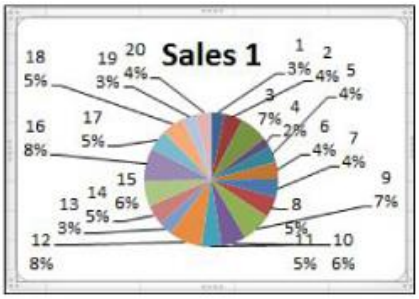
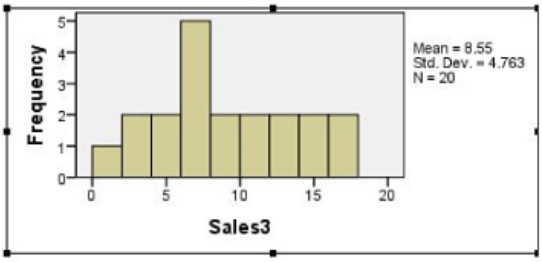
Pie	• Useful in conceptualizing proportions. • Various other versions, like the donut chart (with a hollow center), can also be used. • Useful in situations where the number of variables is limited (not like the illustration to the right).	
Scatter	• Useful when patterns are observed in the data sets. • Useful when outliers are observed in the data that may need to be cleansed out. • Outline trends that a linear chart can augment.	
Histogram	• Ideal to help reveal frequency distributions in variable data sets. • Reduces the size of data by grouping data points into frequencies.	

Table 5.1 Statistical Charts Useful in BA

The charts presented in Table 5.1 reveal interesting facts. The area chart is able to clearly contrast the magnitude of the values in the two variable data sets (Sales 1 and Sales 4). The column chart is useful in revealing the almost perfect linear trend in the Sales 3 data, whereas the scatter chart reveals an almost perfect nonlinear function in Sales 4 data. Additionally, the cluttered pie chart with 20 different percentages illustrates that all charts can or should be used in some situations. The best practices suggest charting should be viewed as an exploratory activity of BA. BA analysts should run a variety of charts and see which ones reveal interesting and useful information. Those charts can be further refined to drill down to more detailed information and more appropriate charts related to the objectives of the BA initiative.

Of course, a cursory review of the Sales 4 data in Figure 5.1 makes the concave appearance of the data in the scatter chart in Table 5.1 unnecessary. But most BA problems involve big data—so large as to make it impossible to just view it and make judgment calls on structure or appearance. This is why descriptive statistics can be employed to view the data in a parameter-based way in the hopes of better understanding the information that the data has to reveal.

5.3. Descriptive Statistics

When selecting the option of *descriptive statistics* in SPSS or Excel, a number of useful statistics are automatically computed for the variables in the data sets. Some of these descriptive statistics are discussed in Table 5.2 as exploratory tools that are helpful in understanding the informational value of data sets.

Statistics	Computation (in Data Set)	Application Area	Example	Application Notes
N or Count	Number of values.	Any.	Sample size of a company's transactions during a month.	Useful in knowing how many items were used in the statistics computations.
Sum	Total of the values in the entire data set.	Any.	Total sales for a company.	Useful in knowing the total value.
Mean	Average of all values.	Any.	Average sales per month.	Useful in capturing the central tendency of the data set.
Median	Midpoint value in the data set arranged from high to low.	Finding the midpoint in the distribution of data.	Total income for citizens of a country.	Useful in finding the point where 50 percent of the data is above and below.
Mode	Most common value in the data set.	Where values are highly repeated in the data set.	Fixed annual salaries where a limited number of wage levels are used.	Useful in declaring a common value in highly repetitive data sets.
Maximum/ Minimum	Largest and smallest values, respectively.	To conceptualize the spread of the data's distribution.	Largest and smallest sales in a day.	Useful in providing a scope or end points in the data.
Range	Difference between the max and min values.	A crude estimate of the spread of the data's distribution.	Spread of sales in units during a month.	Useful as a simple estimate of dispersion.
Standard deviation	Square root of the average of the differences squared between the mean and all other values in the data set.	A precise estimate of the spread of the data's distribution from a mean value in terms of the units used in its computation.	Standard deviation in dollars from mean sales.	The smaller the value, the less the variation and the more predictable using the data set.

Variance	Average differences squared between the mean and all other values.	A variance estimate of the spread of the data's distribution from a mean value, not in terms of the units used in its computation.	Measure of variance that is best used when compared with another variance computed on the same data set.	The smaller the value, the less the variation and the more predictable the data set.
(Coefficient of) Skewedness	Positive or negative values. If value sign is +, distribution is positively skewed; if -, it is negatively skewed. The larger the value, the greater it is skewed.	Measure of the degree of asymmetry of data about a mean.	As the age of residents in a country becomes older, the population age distribution becomes more negatively skewed.	The closer the value is to 0, the better the symmetry. A positively skewed distribution has its largest allocation to the left, and a negative distribution to the right.
(Coefficient of) Kurtosis	Value where less than 3 means a flat distribution and more than 3 means a peaked distribution.	Measure of the degree of spread vertically in a distribution about a mean. Also, it reveals a positive and a negative symmetry depending on its sign.	Distribution of customers at lunch and dinner times peaks and then flattens out.	The closer the value is to 2, the less is the kurtosis (peaking or flattening in the distribution).
Standard Error (of the Mean)	Mean of the sample standard deviation (that is, a standard deviation adjusted to reflect a sample size).	Standard deviation of a sampling distribution.	Standard deviation in dollars from mean sales based on a sample.	The smaller the value, the less the variation and the more predictable the sample data set.
Sample Variance	Same as variance, but adjusted for sample sizes.	Variance estimate of the spread of the sampling data distribution.	Measure of variance when sampling is used for collection purposes.	The smaller the value, the less the variation and the more predictable the sample data set.

Table 5.2 Descriptive Statistics Useful in BA

Fortunately, we do not need to compute these statistics to know how to use them. Computer software provides these descriptive statistics where they're needed or requested. The SPSS descriptive statistics for the illustrative sales data sets are presented in Table 5.3, and the Excel descriptive statistics are presented in Table 5.4.

	N	Range	Min	Max	Sum	Mean		Std. Dev.	Variance	Skewedness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Sales 1	20	40	16	56	703	35.15	2.504	11.198	125.397	.490	.512	–.429	.992
Sales 2	20	1233	1	1234	3344	167.20	83.686	374.254	140065.853	2.241	.512	3.636	.992
Sales 3	20	16	1	17	171	8.55	1.065	4.763	22.682	.272	.512	–.988	.992
Sales 4	20	20	1	21	292	14.60	1.603	7.170	51.411	-.824	.512	–.825	.992
Valid N	20												

Table 5.3 SPSS Descriptive Statistics

	A	B	C	D	E
1	*Statistics*	*Sales1*	*Sales2*	*Sales3*	*Sales4*
2					
3	Mean	35.15	167.2	8.55	14.6
4	Standard Error	2.503970531	83.68567758	1.064931428	1.603286099
5	Median	34	19	7.5	18.5
6	Mode	34	15	6	21
7	Standard Deviation	11.19809664	374.2537276	4.762518131	7.17011341
8	Sample Variance	125.3973684	140065.8526	22.68157895	51.41052632
9	Kurtosis	-0.429071744	3.636015862	-0.988175871	-0.825197874
10	Skewness	0.490453685	2.240917321	0.27209513	-0.824089881
11	Range	40	1233	16	20
12	Minimum	16	1	1	1
13	Maximum	56	1234	17	21
14	Sum	703	3344	171	292
15	Count	20	20	20	20

Table 5.4 Excel Descriptive Statistics

Looking at the data sets for the four variables in Figure 5.1 and at the statistics in Tables 5.3 and 5.4, there are some obvious conclusions based on the detailed statistics from the data sets. It should be no surprise that Sales 2, with a few of the largest values and mostly smaller ones making up the data set, would have the largest variance statistics (standard deviation, sample variance, range, maximum/minimum). Also, Sales 2 is highly, positively skewed (Skewedness > 1) and highly peaked (Kurtosis >3). Note the similarity of the mean, median, and mode in Sales 1 and the dissimilarity in Sales 2. These descriptive statistics provide a more precise basis to envision the behavior of the data. Referred to as *measures of central tendency,* the mean, median, and mode can also be used to clearly define the direction of a skewed distribution. A negatively skewed distribution orders these measures such that mean < median < mode, and a positive skewed distribution orders them such that mode < median < mean.

So what can be learned from these statistics? There are many observations that could be drawn from this data. Keep in mind that, in dealing with the big data sets, one would only have the charts and the statistics to serve as a guide in determining what the data looks like. Yet, from these statistics, one can begin describing the data set. So in the case of Sales 2, it can be predicted that the data set is positively skewed and peaked. Note in Figure 5.2 that the histogram of Sales 2 is presented. The SPSS chart also overlays a normal distribution (a bell-shaped curve) to reflect the positioning of the mean (highest point on the curve, 167.2) and how the data appears to fit the normal distribution (not very well in this situation). As expected, the distribution is positively distributed with a substantial variance between the large values in the data set and the many more smaller valued data points.

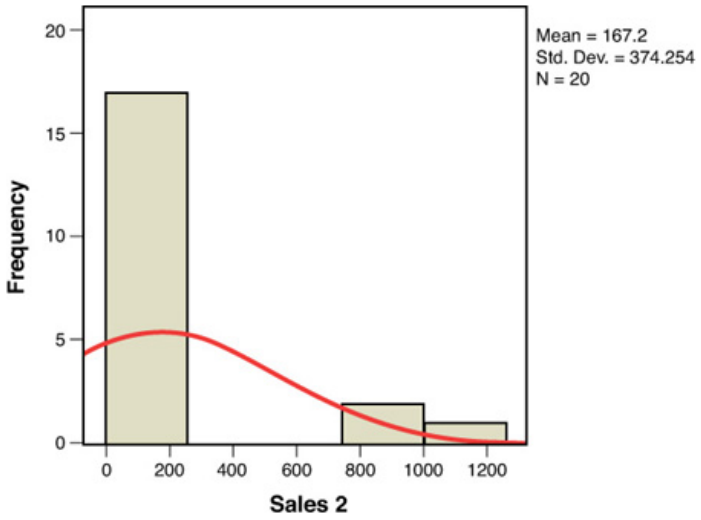


Figure 5.2 SPSS histogram of Sales 2 data

We also know that substantial variance in the data points making up the data set is highly diverse—so much so that it would be difficult to use this variable to predict future behavior or trends. This type of information may be useful in the further steps of the BA process as a means of weeding out data that will not help predict anything useful. Therefore, it would not help an organization improve its operations.

Sometimes big data files become so large they cannot be manipulated by certain statistical software systems. In these instances, a smaller but representative sample of the data can be obtained if necessary. Obtaining the sample for accurate prediction of business behavior requires understanding the sampling process and estimation from that process.

5.4. Sampling and Estimation

The estimation of most business analytics requires sample data. In this section we discuss various types of sampling methods and follow that up with a discussion on how the samples are used in sampling estimation.

5.4.1. Sampling Methods

Sampling is an important strategy of handling large data. If data files are too big to be run by software or just too large to work with, the number of items in the data file can be sampled to provide a new data file that seeks to accurately represent the population from which it comes. In sampling data, there are three components that should be recognized: a population, a sample, and a *sample element* (the items that make up the sample). A firm's collection of customer service performance documents for one year could be designated as a population of customer service performance for that year. From that population, a sample of a lesser number of sample elements (the individual customer service documents) can be drawn to reduce the effort of working with the larger data. Several sampling methods can be used to arrive at a representative sample. Some of these sampling methods are presented in **Table 5.5**.

Sampling Method	Description	Application	Application Notes
Simple Random	Allows each sample element in a population to have an equal chance of selection.	Selecting customers based on their percentage of occurrence as a member of a particular race.	Sample size must be sufficient to avoid sampling bias.
Systematic Random (or Period)	Selects sample elements from a population based on a fixed number in an interval.	Selecting every fifth person leaving an airport to interview.	Assumes the sample elements order in the interval is presented in a random fashion; otherwise, it can result in sampling bias.
Stratified Random	Stage 1: Divide a population into groups (called strata); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people in each of three different economic strata.	Strata must be representative of the population, or it can result in sampling bias.
Cluster Random	Stage 1: Group sample elements geographically (called clusters); Stage 2: Apply simple random sampling.	Randomly selecting an equal number of people from voting districts.	Cluster must be representative of the population, or it can result in sampling bias.
Quota	Based on a fixed quota or number of sample elements.	Selecting the first 200 people who enter a store.	• Mainly used to save time and money. • Sample size must be sufficient to avoid sampling bias.
Judgment	Selects sample elements based on expert judgment.	Selecting candidates for an interview with a special offer based on their appearance.	Prone to bias without defined criteria for selection because of dependency on interviewer experience.

Table 5.5 Sampling Methods

The simple, systematic, stratified, and cluster random methods are based on some kind of probability of their occurrence in a population. Quota and judgment methods are nonprobability-based tools. Although the randomization process in some methods helps ensure representative samples being drawn from the population, sometimes because of cost or time constraints, nonprobability methods are the best choice for sampling.

Which sampling method should be selected for a particular BA analysis? It depends on the nature of the sample. As mentioned in the application notes in **Table 5.5**, the size of the population, the size of the sample, the area of application (geography, strata, ordering of the data, and so on), and even the researchers running the data collection effort impact the particular methodology selected. A best practices approach might begin with a determination of any constraints (time allowed and costs) that might limit the selection of a sample collection effort. That may narrow the choice to something like a quota method. Another best practices recommendation is to start with the objective(s) of the BA project and use them as a guide in the selection of the sampling method. For example, suppose the objective of a BA analysis is to increase

sales of a particular product. This might lead to random sampling of customers or even a stratified sample by income levels, if income is important to the end results of the analysis. Fortunately, there is software to make the data collection process easier and less expensive.

Data file software can be used with the methods mentioned earlier to quickly collect a sample. For example, SPSS permits simple, systematic, stratified, and cluster random methods, among others. Using this software requires a designation of the number of sample elements in each stratum (for example, selected 2 for each stratum in this example). In **Table 5.6**, SPSS has defined seven strata for the Sales 4 data. The logic of this stratification can be observed by looking at the Sales 4 data in **Figure 5.1**. The additional SPSS printout in **Figure 5.3** shows the specific sample elements that were randomly selected in each stratum, as well as totals and their percentages in the resulting sample. For example, only 0.33, or 33 percent, of the "21" strata sample elements were randomly selected by the SPSS program.

Summary for Stage 1

		Number of Units Sampled		Proportion of Units Sampled	
		Requested	Actual	Requested	Actual
Sales4 =	1	2	2	100.0%	100.0%
	5	2	2	100.0%	100.0%
	9	2	2	100.0%	100.0%
	12	2	2	100.0%	100.0%
	18	2	2	100.0%	100.0%
	19	2	2	50.0%	50.0%
	21	2	2	33.3%	33.3%

Table 5.6 SPSS Stratifications of Sample for Sales 4 Variable

Sales4	InclusionProbability_1_	SampleWeightCumulative_1_	PopulationSize_1_	SampleSize_1_
1	1.00	1.00	2	2
5	1.00	1.00	2	2
9	1.00	1.00	2	2
12	1.00	1.00	2	2
18	1.00	1.00	2	2
19	.	.	.	.
19	.50	2.00	4	2
21	.	.	.	.
21	.	.	.	.
21	.33	3.00	6	2
21	.	.	.	.
21	.33	3.00	6	2
21	.	.	.	.
19	.50	2.00	4	2
19	.	.	.	.
18	1.00	1.00	2	2
12	1.00	1.00	2	2
9	1.00	1.00	2	2
5	1.00	1.00	2	2
1	1.00	1.00	2	2

Figure 5.3 SPSS stratified sample of Sales 4 variable identified

Excel also permits simple random and periodic sampling. For example, Figure 5.4 shows the Excel input and printout results for the Sales 4 data. In this example, a random sample of 5 values is requested from the sample elements of 20. The resulting 5 sample elements that were randomly selected are presented in the lower-right side of Figure 5.4.

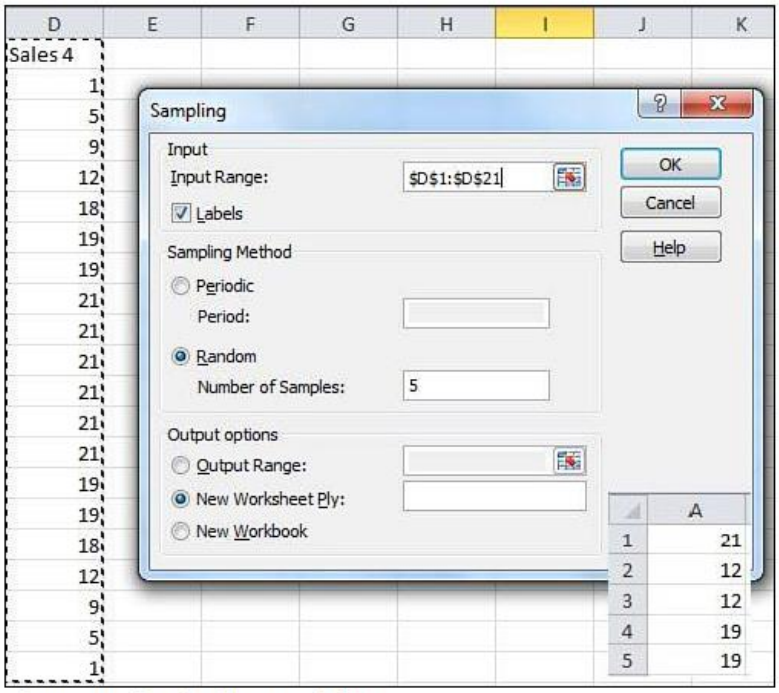


Figure 5.4 Excel random sample of Sales 4 variable

5.4.2. Sampling Estimation

Invariably, using any sampling method can cause errors in the sample results. Most of the statistical methods listed in Table 5.2 are formulated for population statistics. Once sampling is introduced into any statistical analysis, the data must be treated as a sample and not as a population. Many statistical techniques, such as standard error of mean and sample variance, incorporate mathematical correction factors to adjust descriptive analysis statistical tools to compensate for the possibility of sampling error.

One of the methods of compensating for error is to show some degree of confidence in any sampling statistic. The confidence in the sample statistics used can be expressed in a *confidence interval*, which is an interval estimate about the sample statistics. In general, we can express this interval estimate as follows:

Confidence interval = (sample statistic) ± [(confidence coefficient) × (standard error of the estimate)]

The *sample statistic* in the confidence interval can be any measure or proportion from a sample that is to be used to estimate a population parameter, such as a measure of central tendency like a mean. The *confidence coefficient* is set as a percentage to define the degree of confidence to accurately identify the correct sample statistic. The larger the confidence coefficient, the more likely the population mean from the sample will fall within the confidence interval. Many software systems set a 95 percent *confidence level* as the default confidence coefficient, although any percentage can be used. Both SPSS and Excel permit the user to enter a desired percentage. The *standard error of the estimate* in the preceding expression can be any statistical estimate, including proportions used to estimate a population parameter. For example, using a mean as the sample statistic, we have the following interval estimate expression:

Confidence interval = mean ± [(95 percent) × (standard error of the mean)]

The output of this expression consists of two values that form high and low values defining the confidence interval. The interpretation of this interval is that the true population mean represented by the sample has a 95 percent chance of falling in the interval. In this way, there is still a 5 percent chance that the true population mean will not fall in the interval due to sampling error. Because the standard error of the mean is based on variation statistics (standard deviation), the larger the variance statistics used in this expression, the wider the confidence interval and the less precise the sample mean value, which results in a good estimate for the true population mean.

Both SPSS and Excel compute confidence intervals when analyzing various statistical measures and tests. For example, the SPSS printout in **Table 5.7** is of the 95 percent confidence interval for the Sales 1 variable. With a sample mean value of 35.15, the confidence interval suggests there is a 95 percent chance that the true population mean falls between 29.91 and 40.39. When trying to ascertain if the sample is of any value, this kind of information can be of great significance. For example, knowing with 95 percent certainty there is at least a mean of 29.91 might make the difference between continuing to sell a product or not because of a needed requirement for a breakeven point in sales.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean	95 Percent Confidence Interval of the Difference	
Sales 1	20	35.15	11.198	2.504	Lower	Upper
					29.91	40.39

Table 5.7 SPSS 95 Percent Confidence Intervals for Sales 1 Variable

Confidence intervals are also important for demonstrating the accuracy of some forecasting models. For example, confidence intervals can be created about a regression equation model forecast to see how far off the estimates might be if the model is used to predict future sales. For additional discussion on confidence intervals, see **Appendix A**, "**Statistical Tools**."

5.5. Introduction to Probability Distributions

By taking samples, one seeks to reveal population information. Once a sample is taken on which to base a forecast or a decision, there is the possibility that it may not accurately capture the population information. No single sample can assure an analyst that the true population information has been captured. Confidence interval statistics are employed to reflect the possibility of error from the true population information.

To utilize the confidence interval formula expressed in **Section 5.4**, a confidence coefficient percentage (95 percent) is set as a way to express the possibility that the sample statistics used to represent the population statistics may have a potential for error. The confidence coefficient used in the confidence interval is usually referred to as a *Z value*. It is spatially related to the area (expressed as a percentage or frequency) representing the probability under the curve of a distribution. The *sample standard normal distribution* is the bell-shaped curve illustrated in **Figure 5.5**. This distribution shows the relationship of the Z value to the area under the curve. The Z value is the number of standard error of the means.

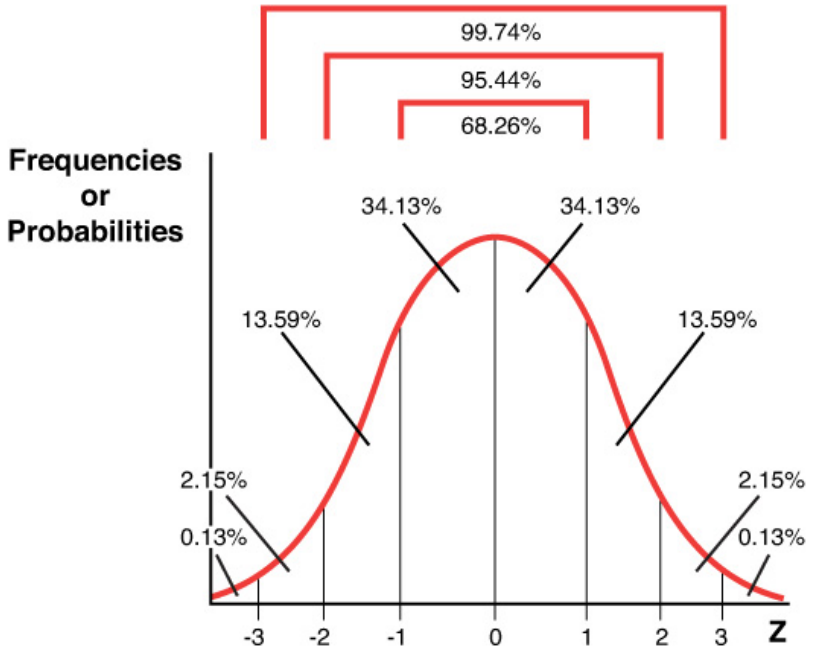


Figure 5.5 Standard normal probability distribution

The *confidence coefficient* is related to the Z values, which divide the area under a normal curve into probabilities. Based on the *central limit theorem*, we assume that all sampling distributions of sufficient size are normally distributed with a standard deviation equal to the standard error of the estimate. This means that an interval of plus or minus two standard errors of the estimate (whatever the estimate is) has a 95.44 percent chance of containing the true or actual population parameter. Plus or minus three standard errors of the estimate has a 99.74 percent chance of containing the true or actual population parameter. So the Z value represents the number of standard errors of the estimate. Table 5.8 has selected Z values for specific confidence levels representing the probability that the true population parameter is within the confidence interval and represents the percentage of area under the curve in that interval.

Confidence Level	Related Z-Value
0.60	0.253
0.70	0.524
0.80	0.842
0.90	1.282
0.95	1.645
0.99	2.327
0.999	3.080

Table 5.8 Selected Z-values and Confidence Levels

The important BA use of the probability distributions and confidence intervals is that they suggest an assumed parameter based on a sample that has properties that will allow analysts to predict or forecast with some assessed degree of statistical accuracy. In other words, BA analysts can, with some designated confidence level, use samples from large databases to accurately predict population parameters.

Another important value to probability distributions is that they can be used to compute probabilities that certain outcomes like success with business performance may occur. In the exploratory descriptive analytics step of the BA process, assessing the probabilities of some events occurring can be a useful strat-

egy to guide subsequent steps in an analysis. Indeed, probability information may be very useful in weighing the choices an analyst faces in any of the steps of the BA process. Suppose, for example, the statistics from the Sales 1 variable in Table 5.8 are treated as a sample to discover the probability of sales greater than one standard error of the mean above the current mean of 35.15. In Figure 5.6, the mean (35.15) and standard error of the mean (2.504) statistics are included at the bottom of the standard sampling normal distribution. By adding one standard error of the mean to the sample mean, the resulting value is 37.654. The sum of the area (the shaded region in Figure 5.6) representing the total probability beyond 37.654 is a probability of 15.87 (13.59+2.15+0.13). So there is only a 15.87 percent probability that sales will exceed 37.654 based on the sample information for the Sales 1 variable.

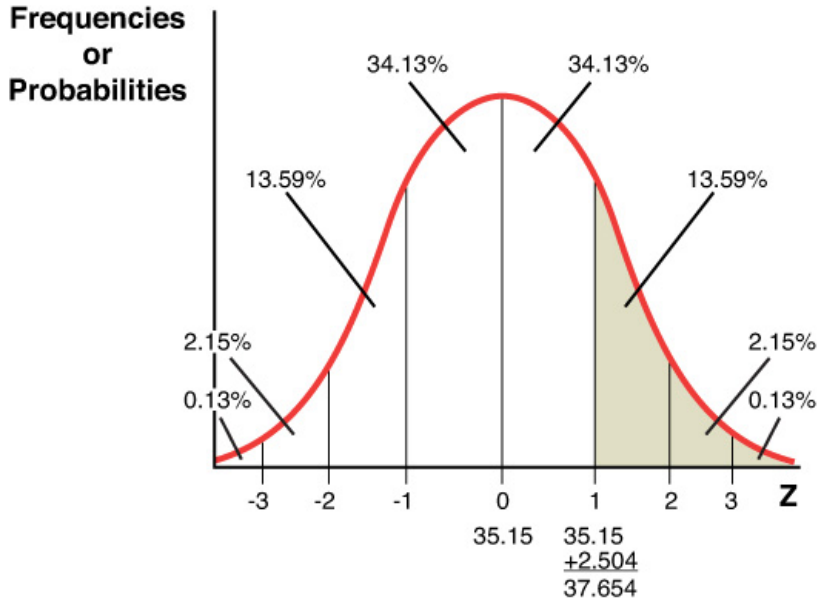


Figure 5.6 Probability function example

The ability to assess probabilities using this approach is applicable to other types of probability distributions. For a review of probability concepts and distributions, probability terminology, and probability applications, see Appendix A.

5.6. Marketing/Planning Case Study Example: Descriptive Analytics Step in the BA Process

In the last section of this chapter and in Chapters 6, "What Are Predictive Analytics?" and 7, "What Are Prescriptive Analytics?" an ongoing marketing/planning case study of the relevant BA step discussed in those chapters will be presented to illustrate some of the tools and strategies used in a BA problem analysis. This is the first installment of the case study dealing with the descriptive analytics step in BA. The predictive analytics step (in Chapter 6) and prescriptive analytics step (in Chapter 7) will continue with this ongoing case study.

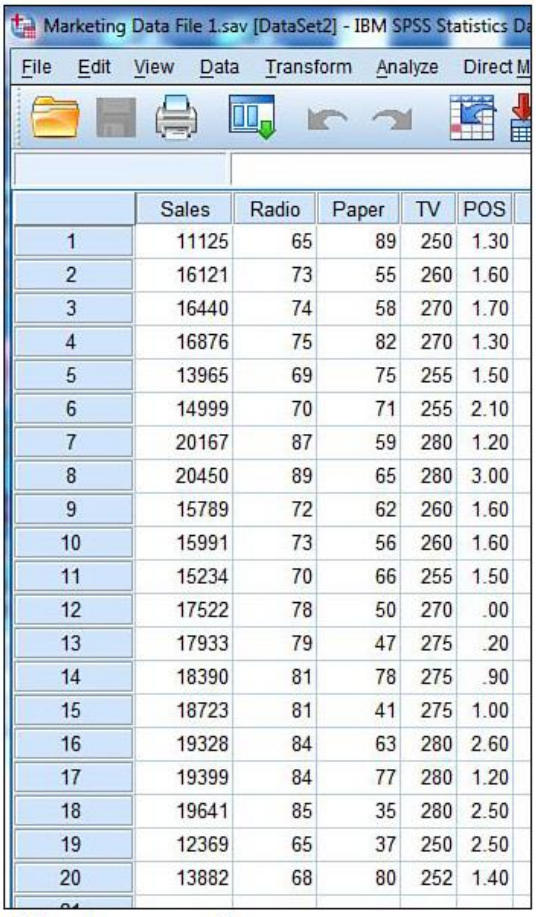
5.6.1. Case Study Background

A firm has collected a random sample of monthly sales information on a service product offered infrequently and only for a month at a time. The sale of this service product occurs only during the month that the promotion efforts are allocated. Basically, promotion funds are allocated at the beginning or during the month, and whatever sales occur are recorded for that promotion effort. There is no spillover of pro-

motion to another month, because monthly offerings of the service product are independent and happen randomly during any particular year. The nature of the product does not appear to be impacted by seasonal or cyclical variations, which prevents forecasting and makes planning the budget difficult.

The firm promotes this service product by using radio commercials, newspaper ads, television commercials, and point-of-sale (POS) ad cards. The firm has collected the sales information as well as promotion expenses. Because the promotion expenses are put into place before the sales take place and on the assumption that the promotion efforts impact products, the four promotion expenses can be viewed as predictive data sets (or what will be the predictive variables in a forecasting model). Actually, in terms of modeling this problem, product sales is going to be considered the dependent variable, and the other four data sets represent independent or predictive variables.

These five data sets, in thousands of dollars, are present in the SPSS printout shown in **Figure 5.7**. What the firm would like to know is, given a fixed budget of $350,000 for promoting this service product, when offered again, how best should budget dollars be allocated in the hope of maximizing future estimated months' product sales? This is a typical question asked of any product manager and marketing manager's promotion efforts. Before allocating the budget, there is a need to understand how to estimate future product sales. This requires understanding the behavior of product sales relative to sales promotion. To begin to learn about the behavior of product sales to promotion efforts, we begin with the first step in the BA process: descriptive analytics.



	Sales	Radio	Paper	TV	POS
1	11125	65	89	250	1.30
2	16121	73	55	260	1.60
3	16440	74	58	270	1.70
4	16876	75	82	270	1.30
5	13965	69	75	255	1.50
6	14999	70	71	255	2.10
7	20167	87	59	280	1.20
8	20450	89	65	280	3.00
9	15789	72	62	260	1.60
10	15991	73	56	260	1.60
11	15234	70	66	255	1.50
12	17522	78	50	270	.00
13	17933	79	47	275	.20
14	18390	81	78	275	.90
15	18723	81	41	275	1.00
16	19328	84	63	280	2.60
17	19399	84	77	280	1.20
18	19641	85	35	280	2.50
19	12369	65	37	250	2.50
20	13882	68	80	252	1.40

Figure 5.7 Data for marketing/planning case study

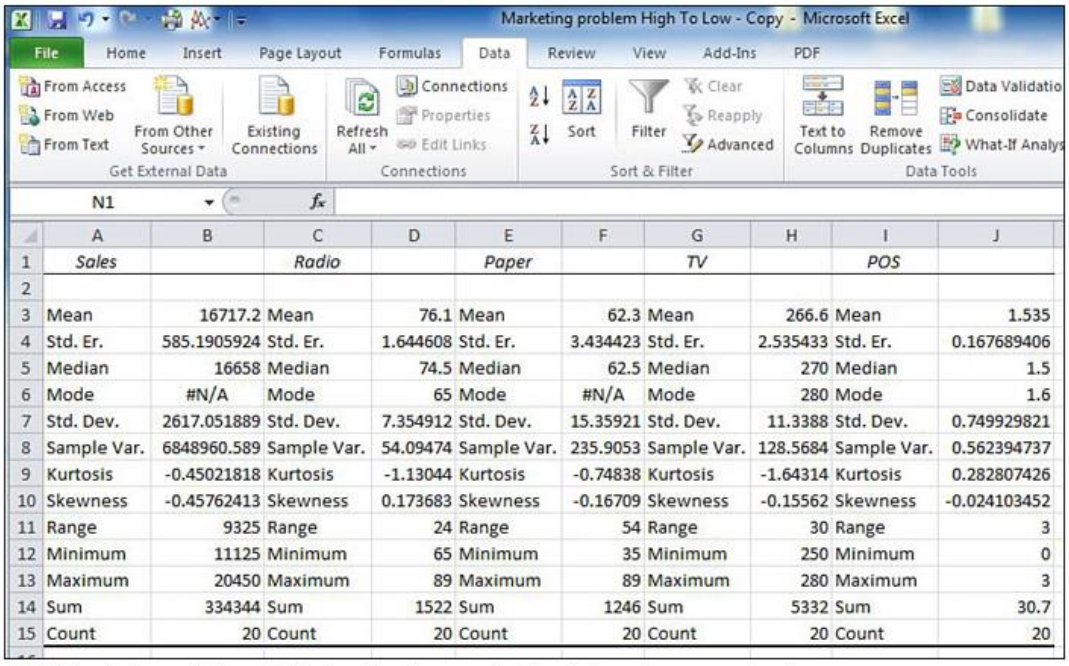
5.6.2. Descriptive Analytics Analysis

To begin conceptualizing possible relationships in the data, one might compute some descriptive statistics and graph charts of data (which will end up being some of the variables in the planned model). SPSS can be used to compute these statistics and charts. The SPSS software printout in Table 5.9 provides a typical set of basic descriptive statistics (means, ranges, standard deviations, and so on) and several charts. Similarly, Excel's printout in Figure 5.8 provides a basic set of descriptive statistics. Where values cannot be computed, a designation of #N/A is provided.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Radio	20	24	65	89	76.10	7.355	54.095
Paper	20	54	35	89	62.30	15.359	235.905
TV	20	30	250	280	266.60	11.339	128.568
POS	20	3	0	3	1.54	.750	.562
Sales	20	9325	11125	20450	16717.20	2617.052	6848960.589
Valid N (listwise)	20						

Table 5.9 SPSS Descriptive Statistics for the Marketing/Planning Case Study



	A	B	C	D	E	F	G	H	I	J
1	*Sales*		*Radio*		*Paper*		*TV*		*POS*	
2										
3	Mean	16717.2	Mean	76.1	Mean	62.3	Mean	266.6	Mean	1.535
4	Std. Er.	585.1905924	Std. Er.	1.644608	Std. Er.	3.434423	Std. Er.	2.535433	Std. Er.	0.167689406
5	Median	16658	Median	74.5	Median	62.5	Median	270	Median	1.5
6	Mode	#N/A	Mode	65	Mode	#N/A	Mode	280	Mode	1.6
7	Std. Dev.	2617.051889	Std. Dev.	7.354912	Std. Dev.	15.35921	Std. Dev.	11.3388	Std. Dev.	0.749929821
8	Sample Var.	6848960.589	Sample Var.	54.09474	Sample Var.	235.9053	Sample Var.	128.5684	Sample Var.	0.562394737
9	Kurtosis	-0.45021818	Kurtosis	-1.13044	Kurtosis	-0.74838	Kurtosis	-1.64314	Kurtosis	0.282807426
10	Skewness	-0.45762413	Skewness	0.173683	Skewness	-0.16709	Skewness	-0.15562	Skewness	-0.024103452
11	Range	9325	Range	24	Range	54	Range	30	Range	3
12	Minimum	11125	Minimum	65	Minimum	35	Minimum	250	Minimum	0
13	Maximum	20450	Maximum	89	Maximum	89	Maximum	280	Maximum	3
14	Sum	334344	Sum	1522	Sum	1246	Sum	5332	Sum	30.7
15	Count	20	Count	20	Count	20	Count	20	Count	20

Figure 5.8 Excel descriptive statistics for the marketing/planning case study

Remember, this is the beginning of an exploration that seeks to describe the data and get a handle on what it may reveal. This effort may take some exploration to figure out the best way to express data from a file or database, particularly as the size of the data file increases. In this simple example, the data sets are small but can still reveal valuable information if explored well.

In Figure 5.9, five typical SPSS charts are presented. Respectively, these charts include a bar chart (sales), an area chart (radio), a line chart (paper), a pie chart (TV), and a dot chart (POS). These charts are inter-

esting, but they're not very revealing of behavior that will help in understanding future sales trends that may be hiding in this data. Figure 5.10 presents the comparable Excel charts. The only real difference is that the pie chart is called a donut chart.

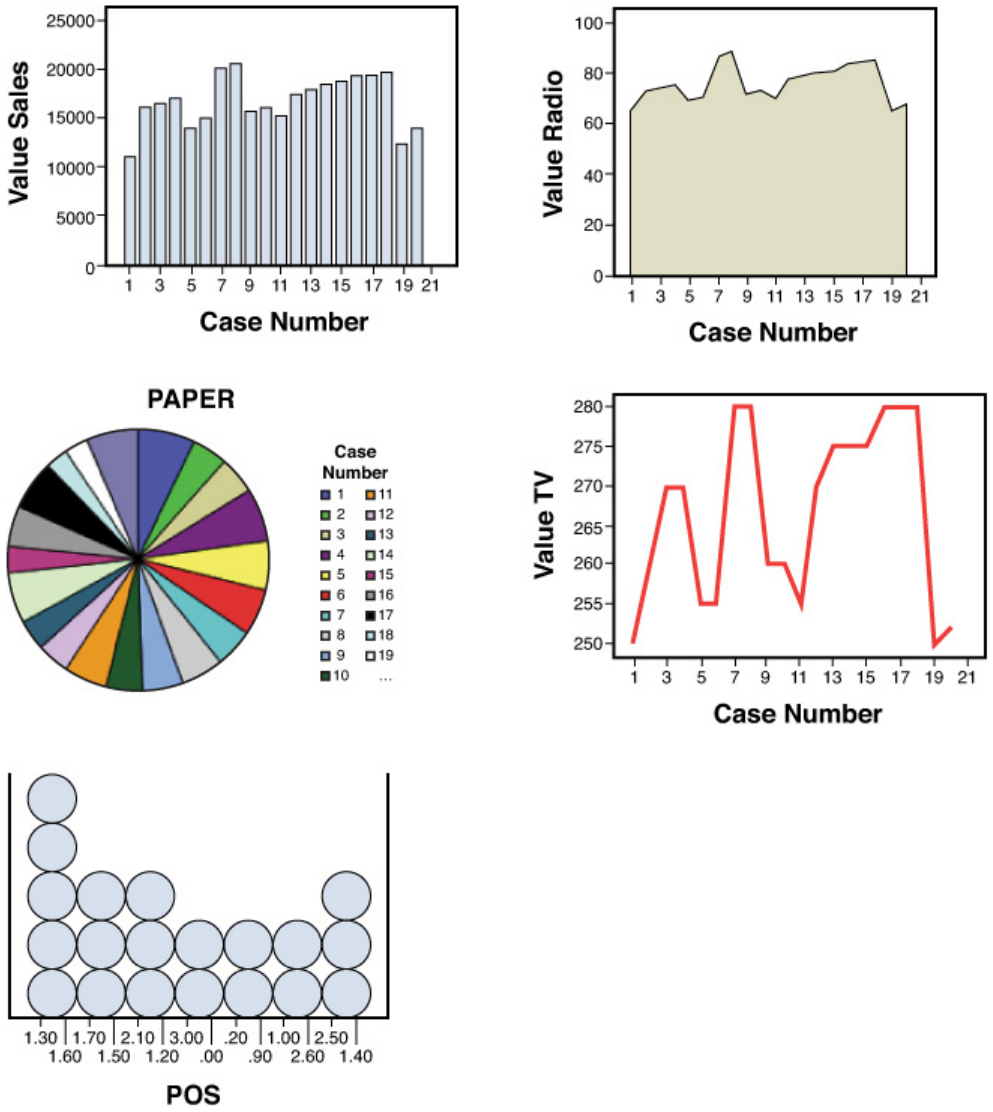


Figure 5.9 Preliminary SPSS charts for the marketing/planning case study

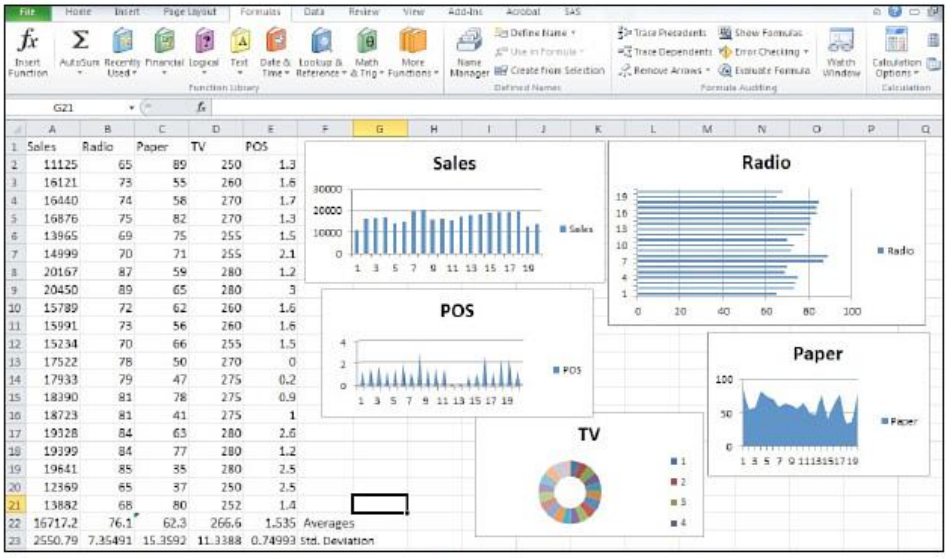


Figure 5.10 Preliminary Excel charts for the marketing/planning case study

To expedite the process of revealing potential relational information, think in terms of what one is specifically seeking. In this instance, it is to predict the future sales of the service product. That means looking for a graph to show a trend line. One type of simple graph that is related to trend analysis is a line chart. Using SPSS again, line charts can be computed for each of the five data sets. These charts are presented in Figure 5.11. The vertical axis consists of the dollar values, and the horizontal axis is the number ordering of observations as listed in the data sets. The comparable Excel charts are presented in Figure 5.12.

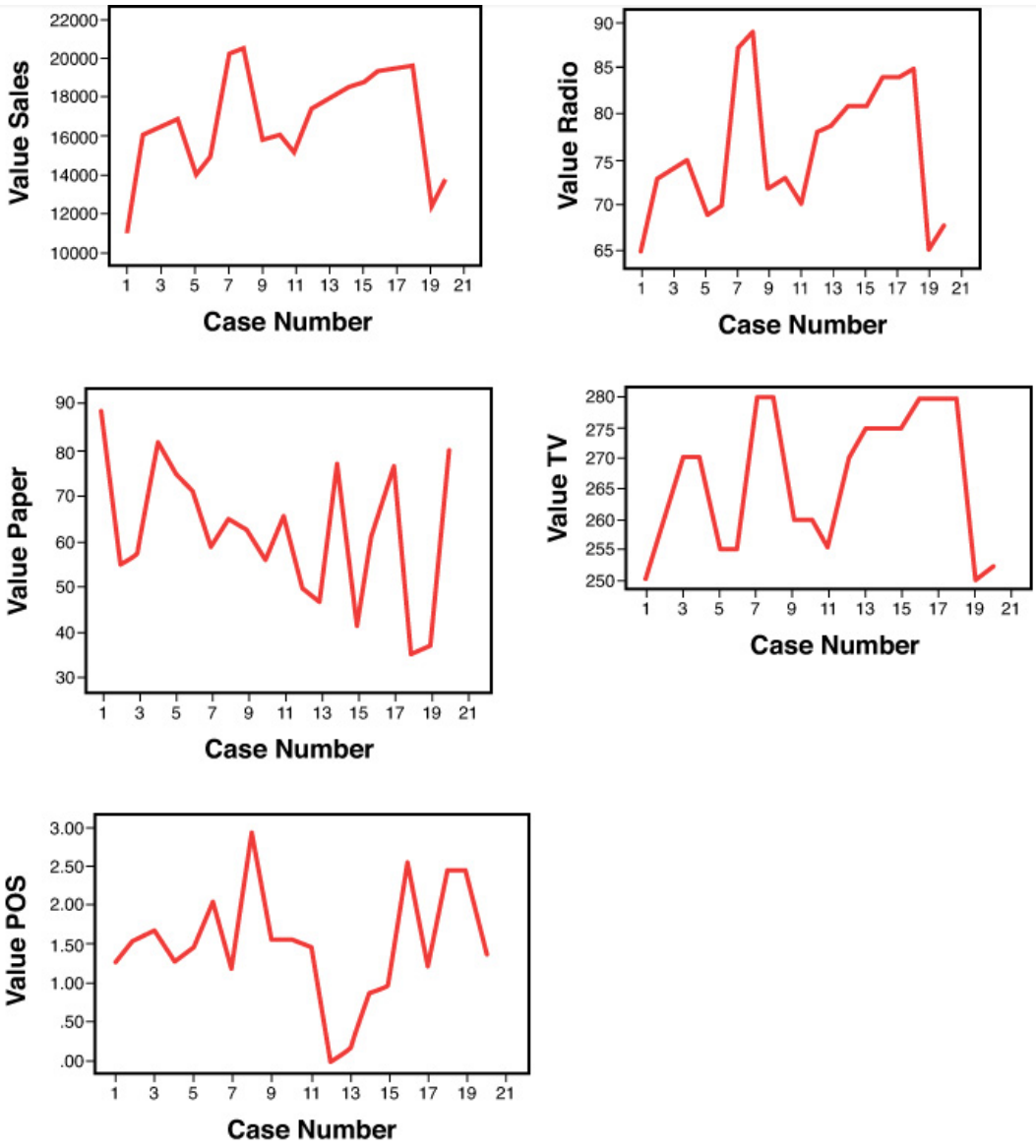
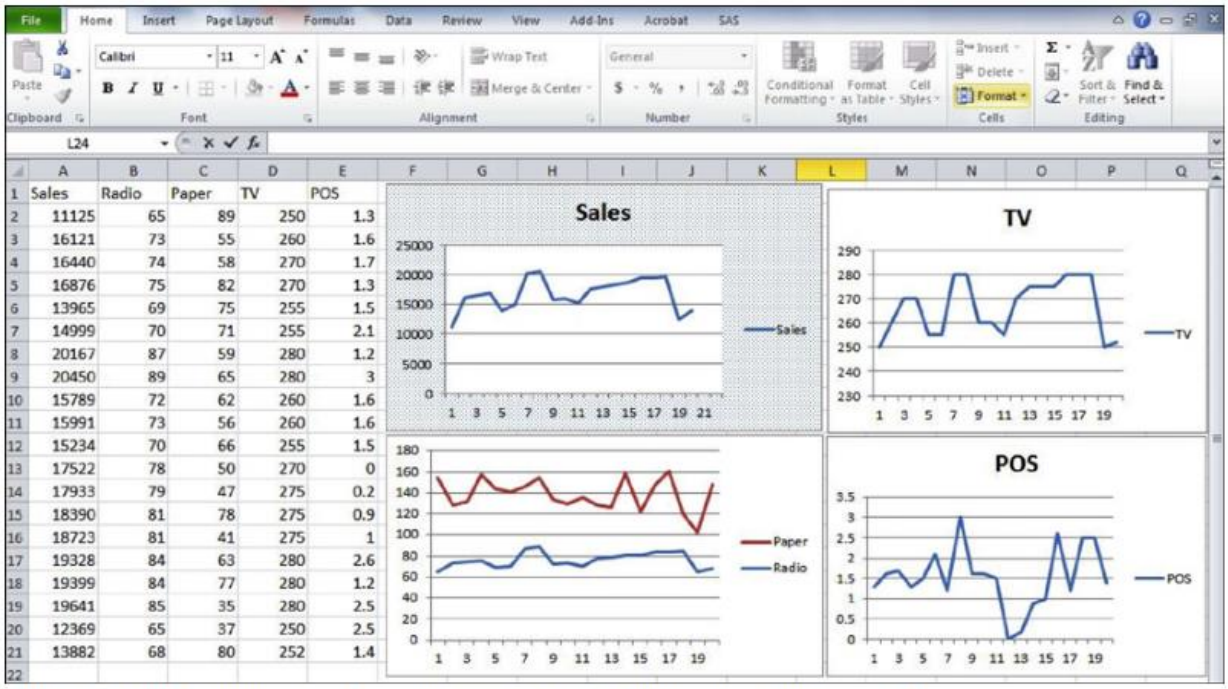


Figure 5.11 Preliminary SPSS line charts for the marketing/planning case study



Sales	Radio	Paper	TV	POS
11125	65	89	250	1.3
16121	73	55	260	1.6
16440	74	58	270	1.7
16876	75	82	270	1.3
13965	69	75	255	1.5
14999	70	71	255	2.1
20167	87	59	280	1.2
20450	89	65	280	3
15789	72	62	260	1.6
15991	73	56	260	1.6
15234	70	66	255	1.5
17522	78	50	270	0
17933	79	47	275	0.2
18390	81	78	275	0.9
18723	81	41	275	1
19328	84	63	280	2.6
19399	84	77	280	1.2
19641	85	35	280	2.5
12369	65	37	250	2.5
13882	68	80	252	1.4

Figure 5.12 Preliminary Excel line charts for the marketing/planning case study

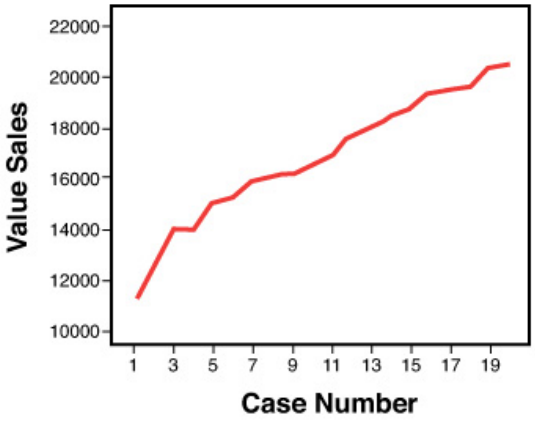
While providing a less confusing graphic presentation of the up-and-down behavior of the data, the charts in these figures still do not clearly reveal any possible trend information. Because the 20 months of data are not in any particular order and are not related to time, they are independent values that can be reordered in any way. Reordering data or sorting it can be a part of the descriptive analytics process when needed. Because trend is usually an upward or downward linear behavior, one might be able to observe a trend in the product sales data set if that data is reordered from low to high (or high to low). Reordering the sales by moving the 20 rows of data around such that sales is arranged from low to high is presented in **Figure 5.13**. Using this reordered data set, the SPSS results are illustrated in the new line charts in **Figure 5.14**. The comparable Excel charts are presented in **Figure 5.15**.

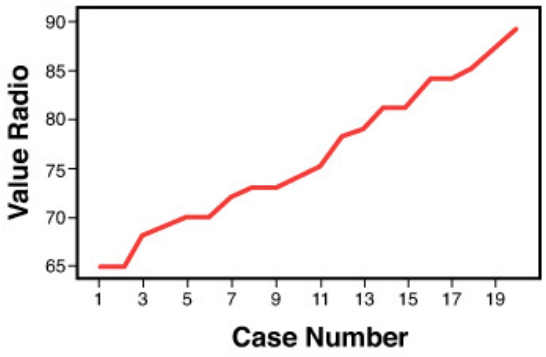
*Untitled2 [DataSet4] - IBM SPSS Statistics Data Editor

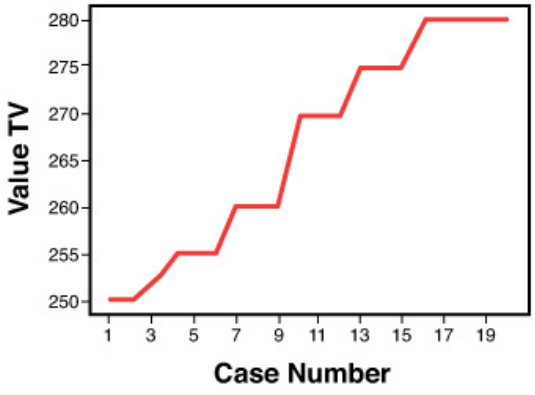
File Edit View Data Transform Analyze Direct Marketing Gra

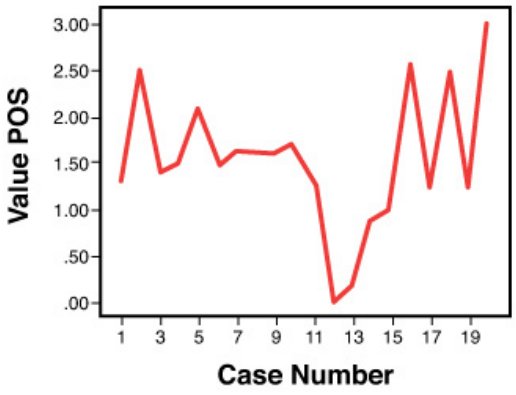
	Sales	Radio	Paper	TV	POS
1	11125	65	89	250	1.30
2	12369	65	37	250	2.50
3	13882	68	80	252	1.40
4	13965	69	75	255	1.50
5	14999	70	71	255	2.10
6	15234	70	66	255	1.50
7	15789	72	62	260	1.60
8	15991	73	56	260	1.60
9	16121	73	55	260	1.60
10	16440	74	58	270	1.70
11	16876	75	82	270	1.30
12	17522	78	50	270	.00
13	17933	79	47	275	.20
14	18390	81	78	275	.90
15	18723	81	41	275	1.00
16	19328	84	63	280	2.60
17	19399	84	77	280	1.20
18	19641	85	35	280	2.50
19	20167	87	59	280	1.20
20	20450	89	65	280	3.00
21					

Figure 5.13 Reordered data in line charts for the marketing/planning case study









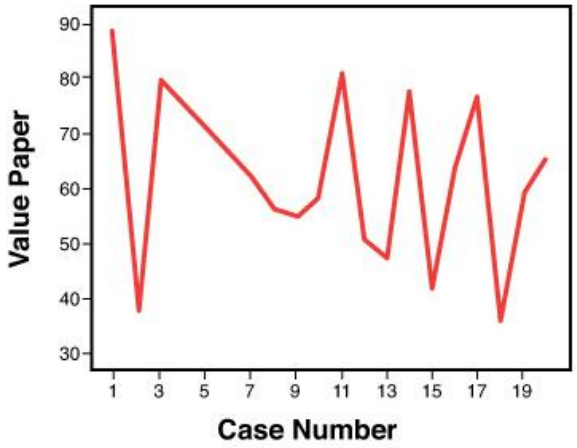


Figure 5.14 SPSS line charts based on reordered data for the marketing/planning case study

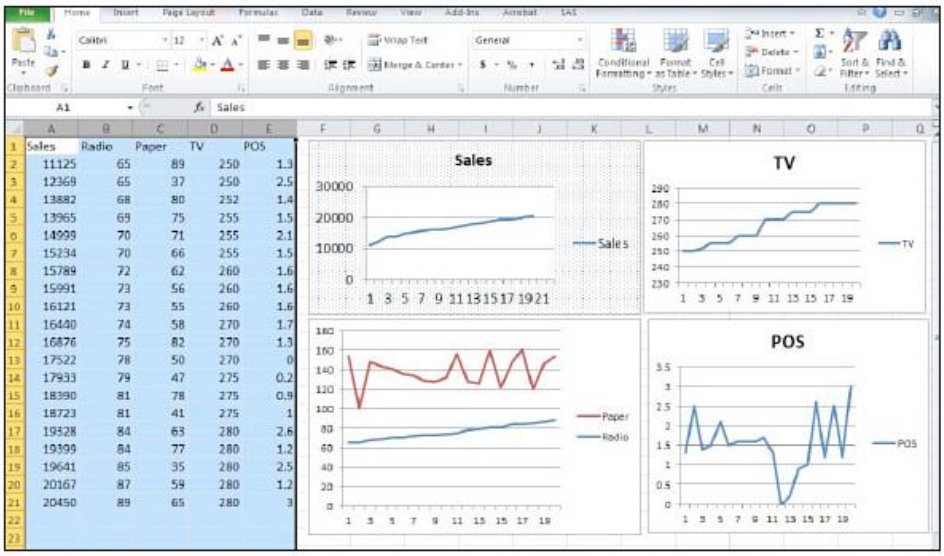


Figure 5.15 Excel line charts based on reordered data for the marketing/planning case study

Given the low to high reordering of the product sales as a guide, some of the other four line charts suggest a relationship with product sales. Both radio and TV commercials appear to have a similar low to high trending relationship that matches with product sales. This suggests these two will be good predictive variables for product sales, whereas newspaper and POS ads are still considerably volatile in their charted relationships with product sales. Therefore, these two latter variables might not be useful in a model seeking to predict product sales. They cannot be ruled out at this point in the analysis, but they are suspected of adding little to a model for accurately forecasting product sales. Put another way, they appear to add unneeded variation that may take away from the accuracy of the model. Further analysis is called for to explore in more detail and sophistication the best set of predictive variables to predict the relationships in product sales.

In summary, for this case study, the descriptive analytics analysis has revealed a potential relationship between radio and TV commercials and future product sales, and it questions the relationship of newspaper and POS ads to sales. The managerial ramifications of these results might suggest discontinuing investing in newspaper and POS ads and more productively allocate funds to radio and TV commercials. Before such a reallocation can be justified, more analysis is needed. The next step in the analysis, predictive analytics, will be presented in the last section of **Chapter 6**.

Summary

This chapter discussed data visualization and exploration. In particular, this chapter described and illustrated graphic and statistical methods useful in the descriptive analytics step of the BA process. Illustrations of both SPSS and Excel printouts of graphs, charts, and statistical methods were presented. In addition, sampling methods were described, along with the available software applications from SPSS and Excel. Sampling estimation was also discussed, as was its connection to sampling distributions for purposes of error estimation in measures of central tendency. Finally, this chapter presented the first installment of a case study illustrating the descriptive analytics step of the BA process. The remaining installments will be presented in **Chapters 6** and 7.

Several of the appendixes of this book are designed to augment the chapter material by including technical, mathematical, and statistical tools. For both greater understanding of the methodologies discussed in this chapter and a basic review of statistical and other quantitative methods, a review of the appendixes mentioned in this chapter is recommended.

The results of the descriptive analytics step of the BA process create an exploratory foundation on which further analysis can be based. In **Chapter 6**, we continue with the second step of the BA process: predictive analytics.

Discussion Questions

1. Why is it important to explore data with graphs and charts?

2. What is the difference between skewedness and kurtosis?

3. Why would we ever want to use a sample if we have population information?

4. Is there a way to determine skewedness from the ordering of the mean, median, and mode measures of central tendency?

5. Which of the sampling methods listed in **Table 5.5** is the best, and why?

6. In setting the confidence level, why not just set one that is low enough for the population parameter to be assured of inclusion?

Problems

1. Using either SPSS or Excel, draw a line graph of the Sales 2 distribution from the data in **Figure 5.1**. Does the kurtosis statistic in **Table 5.4** make sense given your graph? Does the positioning of the mean, median, and mode support the skewedness statistic in **Table 5.4**? Explain the answer to both questions.

2. Using either SPSS or Excel, draw a scatter diagram of the Sales 3 distribution from the data in **Figure 5.1**. Based on the statistics in **Table 5.4**, how would you judge its skewedness: highly or slightly? Does the positioning of the mean, median, and mode support the skewedness statistic in **Table 5.4**? Explain the answer to both questions.

3. Using SPSS or Excel, make a random sample on Sales 3 distribution from the data in **Figure 5.1**. Using the software, determine four items from the data set for sampling purposes. Which specific values should be selected from the data set?

4. Using SPSS or Excel, make a random sample on Sales 2 distribution from the data in **Figure 5.1**. Using the software, determine six items from the data set for sampling purposes. Which specific values should be selected from the data set?

5. With a mean value of 50 and a standard error of the mean of 12, what is the 90 percent confidence interval for this problem?

6. With a mean value of 120 and a standard error of the mean of 20, what is the 99 percent confidence interval for this problem?

7. A firm has computed its mean sales for a new product to be 2,000 units for the year, with a standard error of the mean of 56. The firm would like to know if the probability of its mean sales for next year (based on this year) will be above 2,112. What is the probability?

8. The Homes Golf Ball Company has made a number of different golf products over the years. Research on thousands of balls revealed the mean flight distance of its Maximum Fly golf ball product to be 450 yards, with a standard error of the mean of 145 yards. The company is hoping to improve the product to fly an additional 290 yards. What is the probability of the improvement from 450 to 740 yards?

6. What Are Predictive Analytics?

Chapter objectives:

• Explain what logic-driven models are used for in business analytics (BA).

• Describe what a cause-and-effect diagram is used for in BA.

• Explain the difference between logic-driven and data-driven models.

• Explain how data mining can aid in BA.

• Explain why neural networks can be helpful in determining both associations and classification tasks required in some BA analyses.

• Explain how clustering is undertaken in BA.

• Explain how step-wise regression can be useful in BA.

• Explain how to use R-Squared adjusted statistics in BA.

6.1. Introduction

In **Chapter 1**, "**What Are Business Analytics?**" we defined predictive analytics as an application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in the descriptive analytic analysis. Knowing that relationships exist explains why one set of independent variables (predictive variables) influences dependent variables like business performance. **Chapter 1** further explained that the purpose of the descriptive analytics step is to position decision makers to build predictive models designed to identify and predict future trends.

Picture a situation in which big data files are available from a firm's sales and customer information (responses to differing types of advertisements, customer surveys on product quality, customer surveys on supply chain performance, sale prices, and so on). Assume also that a previous descriptive analytic analysis suggests there is a relationship between certain customer variables, but there is a need to precisely establish a quantitative relationship between sales and customer behavior. Satisfying this need requires exploration into the big data to first establish whether a measurable, quantitative relationship does in fact exist and then develop a statistically valid model in which to predict future events. This is what the predictive analytics step in BA seeks to achieve.

Many methods can be used in this step of the BA process. Some are just to sort or classify big data into manageable files in which to later build a precise quantitative model. As previously mentioned in **Chapter 3**, "**What Resource Considerations Are Important to Support Business Analytics?**" predictive modeling and analysis might consist of the use of methodologies, including those found in forecasting, sampling and estimation, statistical inference, data mining, and regression analysis. A commonly used methodology is multiple regression. (See Appendixes **A**, "**Statistical Tools**," and **E**, "**Forecasting**," for a discussion on multiple regression and ANOVA testing.) This methodology is ideal for establishing whether a statistical relationship exists between the predictive variables found in the descriptive analysis

and the dependent variable one seeks to forecast. An example of its use will be presented in the last section of this chapter.

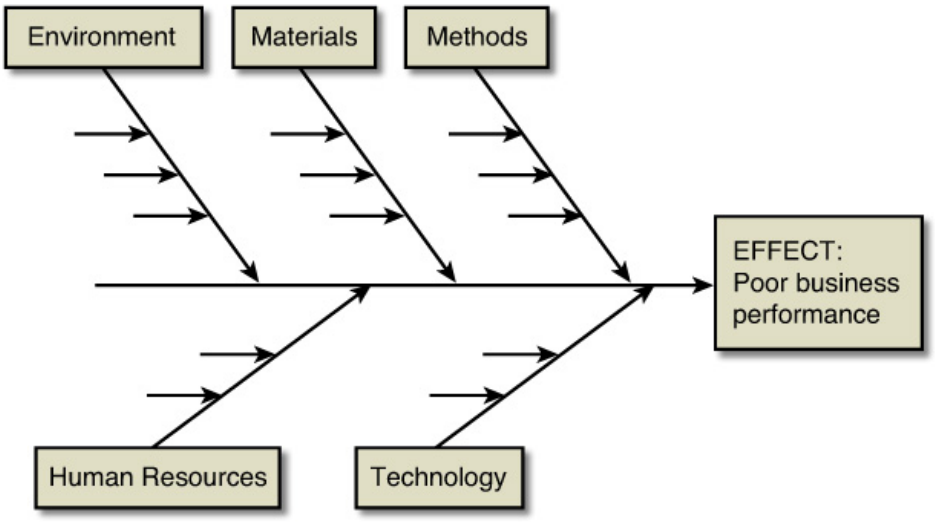
Although single or multiple regression models can often be used to forecast a trend line into the future, sometimes regression is not practical. In such cases, other forecasting methods, such as exponential smoothing or smoothing averages, can be applied as predictive analytics to develop needed forecasts of business activity. (See **Appendix E**.) Whatever methodology is used, the identification of future trends or forecasts is the principle output of the predictive analytics step in the BA process.

6.2. Predictive Modeling

Predictive modeling means developing models that can be used to forecast or predict future events. In business analytics, models can be developed based on logic or data.

6.2.1. Logic-Driven Models

A *logic-driven model* is one based on experience, knowledge, and logical relationships of variables and constants connected to the desired business performance outcome situation. The question here is how to put variables and constants together to create a model that can predict the future. Doing this requires business experience. Model building requires an understanding of business systems and the relationships of variables and constants that seek to generate a desirable business performance outcome. To help conceptualize the relationships inherent in a business system, diagramming methods can be helpful. For example, the *cause-and-effect diagram* is a visual aid diagram that permits a user to hypothesize relationships between potential causes of an outcome (see **Figure 6.1**). This diagram lists potential causes in terms of human, technology, policy, and process resources in an effort to establish some basic relationships that impact business performance. The diagram is used by tracing contributing and relational factors from the desired business performance goal back to possible causes, thus allowing the user to better picture sources of potential causes that could affect the performance. This diagram is sometimes referred to as a *fishbone diagram* because of its appearance.



Source*: Adapted from Figure 5 in **Schniederjans et al. (2014), p. 201.

Figure 6.1 Cause-and-effect diagram*

Another useful diagram to conceptualize potential relationships with business performance variables is called the *influence diagram*. According to **Evans** (**2013**, pp. 228–229), influence diagrams can be useful to conceptualize the relationships of variables in the development of models. An example of an influence diagram is presented in **Figure 6.2**. It maps the relationship of variables and a constant to the desired business performance outcome of profit. From such a diagram, it is easy to convert the information into a quantitative model with constants and variables that define profit in this situation:

Profit = Revenue − Cost, or

Profit = (Unit Price × Quantity Sold) − [(Fixed Cost) + (Variable Cost × Quantity Sold)], or

P = (UP × QS) − [FC + (VC × QS)]

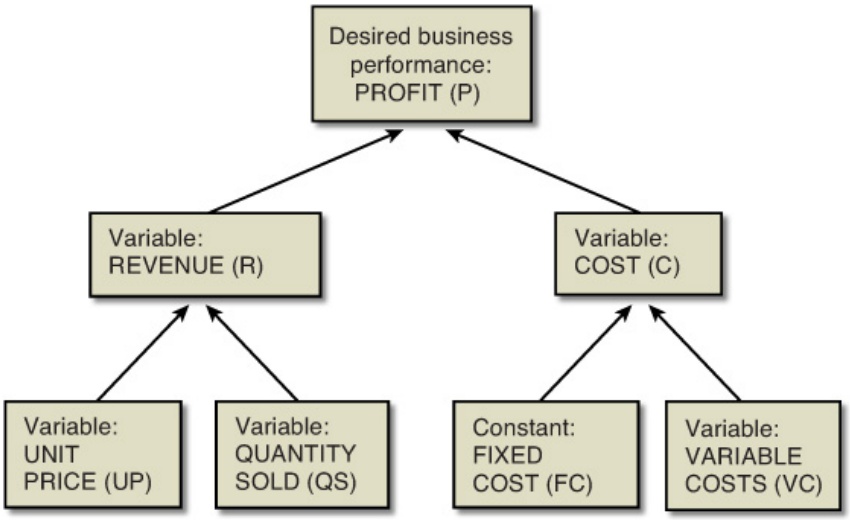


Figure 6.2 An influence diagram

The relationships in this simple example are based on fundamental business knowledge. Consider, however, how complex cost functions might become without some idea of how they are mapped together. It is necessary to be knowledgeable about the business systems being modeled in order to capture the relevant business behavior. Cause-and-effect diagrams and influence diagrams provide tools to conceptualize relationships, variables, and constants, but it often takes many other methodologies to explore and develop predictive models.

6.2.2. Data-Driven Models

Logic-driven modeling is often used as a first step to establish relationships through *data-driven models* (using data collected from many sources to quantitatively establish model relationships). To avoid duplication of content and focus on conceptual material in the chapters, most of the computational aspects and some computer usage content are relegated to the appendixes. In addition, some of the methodologies are illustrated in the case problems presented in this book. Please refer to the Additional Information column in **Table 6.1** to obtain further information on the use and application of the data-driven models.

Data-Driven Models	Possible Applications	Additional Information
Sampling and Estimation	Generate statistical confidence intervals to define limitations and boundaries on future forecasts for other forecasting models.	Chapter 5, "What Are Descriptive Analytics?," Appendix A, "Statistical Tools," Appendix E, "Forecasting."
Regression Analysis	(1) Create a predictive equation useful for forecasting time series forecasts. (2) Weed out predictive variables in forecasting models that add little to predicting values. (3) Generate a trend line for forecasting.	Chapter 6, "What Are Predictive Analytics?," Chapter 8, "A Final Case Study Illustration," Appendix E.

Correlation Analysis	(1) Assess variable relationships. (2) Weed out predictive variables in forecasting models that add little to predicting values.	Chapter 6, Appendix E.
Probability Distributions	(1) Estimate trend behavior that follows certain types of probability distributions. (2) Conduct statistical tests to confirm significance of variables.	Chapter 5, Appendix A.
Predictive Modeling and Analysis	Fit linear and nonlinear models to data to use the models for forecasting.	Appendix A, Appendix E.
Forecasting Models	Those listed in this table and others such as smoothing models can be used to forecast values.	Appendix E.
Simulation	Project future behavior in variables by simulating the past behavior found in probability distributions.	Appendix F, "Simulation."

Table 6.1 Data-Driven Models

6.3. Data Mining

As mentioned in **Chapter 3**, "**What Resource Considerations are Important to Support Business Analytics?**" *data mining* is a discovery-driven software application process that provides insights into business data by finding hidden patterns and relationships in big or small data and inferring rules from them to predict future behavior. These observed patterns and rules guide decision-making. This is not just numbers, but text and social media information from the Web. For example, **Abrahams et al. (2013)** developed a set of text-mining rules that automobile manufacturers could use to distill or mine specific vehicle component issues that emerge on the Web but take months to show up in complaints or other damaging media. These rules cut through the mountainous data that exists on the Web and are reported to provide marketing and competitive intelligence to manufacturers, distributors, service centers, and suppliers. Identifying a product's defects and quickly recalling or correcting the problem before customers experience a failure reduce customer dissatisfaction when problems occur.

6.3.1. A Simple Illustration of Data Mining

Suppose a grocery store has collected a big data file on what customers put into their baskets at the market (the collection of grocery items a customer purchases at one time). The grocery store would like to know if there are any associated items in a typical market basket. (For example, if a customer purchases product A, she will most often associate it or purchase it with product B.) If the customer generally purchases product A and B together, the store might only need to advertise product A to gain both product A's and B's sales. The value of knowing this association of products can improve the performance of the store by reducing the need to spend money on advertising both products. The benefit is real if the association holds true.

Finding the association and proving it to be valid requires some analysis. From the descriptive analytics analysis, some possible associations may have been uncovered, such as product A's and B's association. With any size data file, the normal procedure in data mining would be to divide the file into two parts. One is referred to as a training data set, and the other as a validation data set. The *training data set* develops the association rules, and the *validation data set* tests and proves that the rules work. Starting with the training data set, a common data mining methodology is *what-if analysis* using logic-based software. Excel and SPSS both have what-if logic-based software applications, and so do a number of other software vendors (see **Chapter 3**). These software applications allow logic expressions. (For example, if product A is present, then is product B present?) The systems can also provide frequency and probability information to show the strength of the association. These software systems have differing capabilities, which permit users to deterministically simulate different scenarios to identify complex combinations of associations between product purchases in a market basket.

Once a collection of possible associations is identified and their probabilities are computed, the same logic associations (now considered association rules) are reran using the validation data set. A new set of probabilities can be computed, and those can be statistically compared using hypothesis testing methods to determine their similarity. Other software systems compute correlations for testing purposes to judge the strength and the direction of the relationship. In other words, if the consumer buys product A first, it could be referred to as the *Head* and product B as the *Body* of the association. (**Nisbet et al., 2009**, p. 128). If the same basic probabilities are statistically significant, it lends validity to the association rules and their use for predicting market basket item purchases based on groupings of products.

6.3.2. Data Mining Methodologies

Data mining is an ideal predictive analytics tool used in the BA process. We mentioned in **Chapter 3** different types of information that data mining can glean, and **Table 6.2** lists a small sampling of data mining methodologies to acquire different types of information. Some of the same tools used in the descriptive analytics step are used in the predictive step but are employed to establish a model (either based on logical connections or quantitative formulas) that may be useful in predicting the future.

Types of Information	Description	Sample of Data Mining Methodologies
Association	Occurrence linked to a single event.	Association rules (for example, if-then analysis), correlation analysis, neural networks.
Classification	Pattern that describes the group an item belongs to. Found by examining previous classified existing items and inferring a set of rules that guide the classification process.	Discriminant analysis, logistics regression, neural networks.
Clustering	Similar to classification when no groups have yet been defined. Helps discover different groupings within data.	Hierarchical clustering, K-mean clustering.
Forecasting	Used to predict values that can identify patterns in customer behavior.	Regression analysis, correlation analysis.

Sequence	Event that is linked over time.	Lag correlation analysis, cause-and-effect diagrams.

Table 6.2 Types of Information and Data Mining Methodologies

Several computer-based methodologies listed in **Table 6.2** are briefly introduced here. *Neural networks* are used to find associations where connections between words or numbers can be determined. Specifically, neural networks can take large volumes of data and potential variables and explore variable associations to express a beginning variable (referred to as an *input layer*), through middle layers of interacting variables, and finally to an ending variable (referred to as an *output*). More than just identifying simple one-on-one associations, neural networks link multiple association pathways through big data like a collection of nodes in a network. These nodal relationships constitute a form of classifying groupings of variables as related to one another, but even more, related in complex paths with multiple associations (**Nisbet et al., 2009**, pp. 128–138). SPSS has two versions of neural network software functions: *Multilayer Perceptron* (MLP) and *Radial Basis Function* (RBF). Both procedures produce a predictive model for one or more dependent variables based on the values of the predictive variables. Both allow a decision maker to develop, train, and use the software to identify particular traits (such as bad loan risks for a bank) based on characteristics from data collected on past customers).

*Discriminant anal*ysis is similar to a multiple regression model except that it permits continuous independent variables and a categorical dependent variable. The analysis generates a regression function whereby values of the independent variables can be incorporated to generate a predicted value for the dependent variable. Similarly, *logistic regression* is like multiple regression. Like discriminant analysis, its dependent variable can be categorical. The independent variables, though, in logistic regression can be either continuous or categorical. For example, in predicting potential outsource providers, a firm might use a logistic regression, in which the dependent variable would be to either classify an outsource provider as rejected (represented by the value of the dependent variable being zero) or classify the outsource provider as acceptable (represented by the value of one for the dependent variable).

Hierarchical clustering is a methodology that establishes a hierarchy of clusters that can be grouped by the hierarchy. Two strategies are suggested for this methodology: agglomerative and divisive. The *agglomerative strategy* is a bottom-up approach, where one starts with each item in the data and begins to group them. The *divisive strategy* is a top-down approach, where one starts with all the items in one group and divides the group into clusters. How the clustering takes place can involve many different types of algorithms and differing software applications. One method commonly used is to employ a Euclidean distance formula that looks at the square root of the sum of distances between two variables, their differences squared. Basically, the formula seeks to match up variable candidates that have the least squared error differences. (In other words, they're closer together.)

K-mean clustering is a classification methodology that permits a set of data to be reclassified into *K groups*, where *K* can be set as the number of groups desired. The algorithmic process identifies initial candidates for the *K* groups and then interactively searches other candidates in the data set to be averaged into a mean value that represents a particular *K* group. The process of selection is based on maximizing the distance from the initial *K* candidates selected in the initial run through the list. Each run or iteration through the data set allows the software to select further candidates for each group.

The K-mean clustering process provides a quick way to classify data into differentiated groups. To illustrate this process, use the sales data in **Figure 6.3** and assume these are sales from individual customers. Suppose a company wants to classify the sales customers into high and low sales groups.

Time	Sale
1	13444
2	12369
3	15322
4	13965
5	14999
6	15234
7	12999
8	15991
9	16121
10	18654
11	16876
12	17522
13	17933
14	15233
15	18723
16	13855
17	19399
18	16854
19	20167
20	18654

Figure 6.3 Sales data for cluster classification problem

The SPSS K-Mean cluster software can be found in Analyze > Classify > K-Means Cluster Analysis. Any integer value can designate the *K* number of clusters desired. In this problem set, *K*=2. The SPSS printout of this classification process is shown in **Table 6.3**. The solution is referred to as a *Quick Cluster* because it initially selects the first two high and low values. The Initial Cluster Centers table listed the initial high (20167) and a low (12369) value from the data set as the clustering process begins. As it turns out, the software divided the customers into nine high sales customers with a group mean sales of 18,309 and eleven low sales customers with a group mean sales of 14,503.

Quick Cluster

Initial Cluster Centers

	Cluster	
	1	2
Sale	20167	12369

Final Cluster Centers

	Cluster	
	1	2
Sale	18309	14503

Number of Cases in each Cluster

Cluster	1	9.000
	2	11.000
Valid		20.000
Missing		.000

Table 6.3 SPSS K-Mean Cluster Solution

Consider how large big data sets can be. Then realize this kind of classification capability can be a useful tool for identifying and predicting sales based on the mean values.

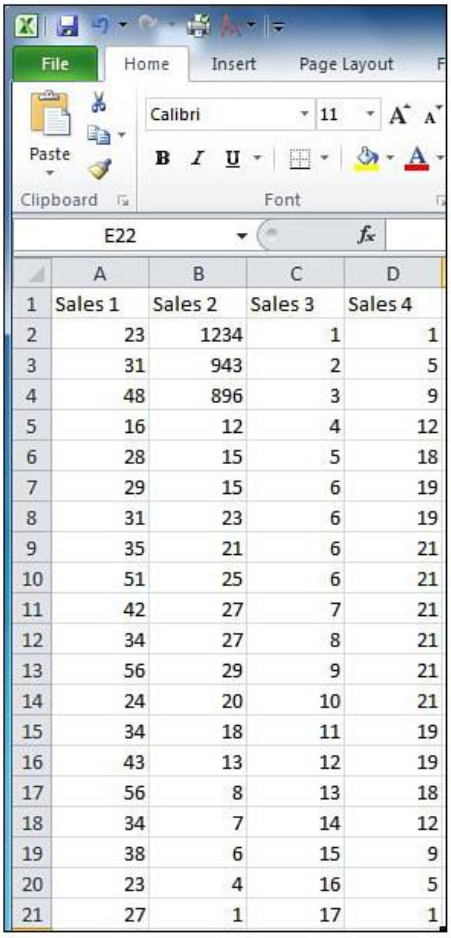
There are so many BA methodologies that no single section, chapter, or even book can explain or contain them all. The analytic treatment and computer usage in this chapter have been focused mainly on conceptual use. For a more applied use of some of these methodologies, note the case study that follows and some of the content in the appendixes.

6.4. Continuation of Marketing/Planning Case Study Example: Prescriptive Analytics Step in the BA Process

In the last sections of **Chapters 5**, **6**, and 7, an ongoing marketing/planning case study of the relevant BA step discussed in those chapters is presented to illustrate some of the tools and strategies used in a BA problem analysis. This is the second installment of the case study dealing with the predictive analytics analysis step in BA. The prescriptive analysis step coming in **Chapter 7**, "**What Are Prescriptive Analytics?**" will complete the ongoing case study.

6.4.1. Case Study Background Review

The case study firm had collected a random sample of monthly sales information presented in **Figure 6.4** listed in thousands of dollars. What the firm wants to know is, given a fixed budget of $350,000 for promoting this service product, when offered again, how best should the company allocate budget dollars in hopes of maximizing the future estimated month's product sales? Before making any allocation of budget, there is a need to understand how to estimate future product sales. This requires understanding the behavior of product sales relative to sales promotion efforts using radio, paper, TV, and point-of-sale (POS) ads.



	A	B	C	D
1	Sales 1	Sales 2	Sales 3	Sales 4
2	23	1234	1	1
3	31	943	2	5
4	48	896	3	9
5	16	12	4	12
6	28	15	5	18
7	29	15	6	19
8	31	23	6	19
9	35	21	6	21
10	51	25	6	21
11	42	27	7	21
12	34	27	8	21
13	56	29	9	21
14	24	20	10	21
15	34	18	11	19
16	43	13	12	19
17	56	8	13	18
18	34	7	14	12
19	38	6	15	9
20	23	4	16	5
21	27	1	17	1

Figure 6.4 Data for marketing/planning case study

The previous descriptive analytics analysis in **Chapter 5** revealed a potentially strong relationship between radio and TV commercials that might be useful in predicting future product sales. The analysis also revealed little regarding the relationship of newspaper and POS ads to product sales. So although radio and TV commercials are most promising, a more in-depth predictive analytics analysis is called for to accurately measure and document the degree of relationship that may exist in the variables to determine the best predictors of product sales.

6.4.2. Predictive Analytics Analysis

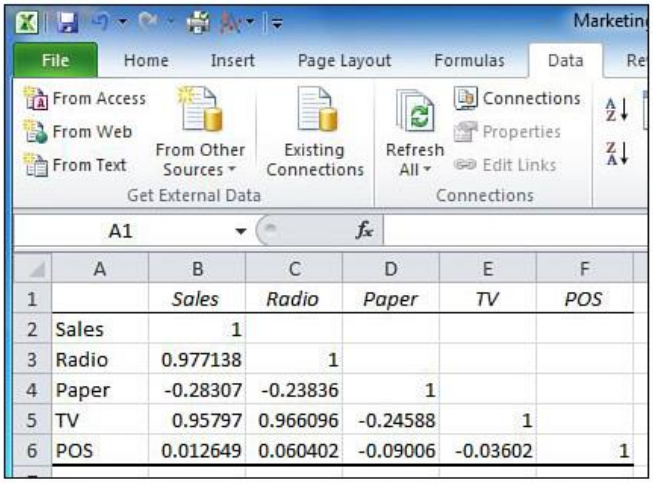
An ideal multiple variable modeling approach that can be used in this situation to explore variable importance in this case study and eventually lead to the development of a predictive model for product sales is correlation and multiple regression. We will use both Excel and IBM's SPSS statistical packages to compute the statistics in this step of the BA process.

First, we must consider the four independent variables—radio, TV, newspaper, POS—before developing the model. One way to see the statistical direction of the relationship (which is better than just comparing graphic charts) is to compute the Pearson correlation coefficients *r* between each of the independent variables with the dependent variable (product sales). The SPSS correlation coefficients and their levels of significance are presented in **Table 6.4**. The comparable Excel correlations are presented in **Figure 6.5**. Note: They do not include the level of significance but provide correlations between all the variables being considered. The larger the Pearson correlation (regardless of the sign) and the smaller the *Significance test* values (these are t-tests measuring the significance of the Pearson *r* value; see **Appendix A**), the more significant the relationship. Both radio and TV are statistically significant correlations, whereas at a 0.05 level of significance, paper and POS are not statistically significant.

Statistic	Radio	Paper	TV	POS
Pearson Correlation *r* with Product Sales	.977	–.283	.958	.013
Significance Test (1-Tailed)*	.000	.113	.000	.479

*Values of 0.05 or less would designate a significant relationship with product sales

Table 6.4 SPSS Pearson Correlation Coefficients: Marketing/Planning Case Study



	A	B	C	D	E	F
1		*Sales*	*Radio*	*Paper*	*TV*	*POS*
2	Sales	1				
3	Radio	0.977138	1			
4	Paper	-0.28307	-0.23836	1		
5	TV	0.95797	0.966096	-0.24588	1	
6	POS	0.012649	0.060402	-0.09006	-0.03602	1

Figure 6.5 Excel Pearson correlation coefficients: marketing/planning case study

Although it can be argued that the positive or negative correlation coefficients should not automatically discount any variable from what will be a predictive model, the negative correlation of newspapers sug-

gests that as a firm increases investment in newspaper ads, it will decrease product sales. This does not make sense in this case study. Given the illogic of such a relationship, its potential use as an independent variable in a model is questionable. Also, this negative correlation poses several questions that should be considered. Was the data set correctly collected? Is the data set accurate? Was the sample large enough to have included enough data for this variable to show a positive relationship? Should it be included for further analysis? Although it is possible that a negative relationship can statistically show up like this, it does not make sense in this case. Based on this reasoning and the fact that the correlation is not statistically significant, this variable (i.e., newspaper ads) will be removed from further consideration in this exploratory analysis to develop a predictive model.

Some researchers might also exclude POS based on the insignificance (p=0.479) of its relationship with product sales. However, for purposes of illustration, continue to consider it a candidate for model inclusion. Also, the other two independent variables (radio and TV) were both found to be significantly related to product sales, as reflected in the correlation coefficients in the tables.

At this point, there is a dependent variable (product sales) and three candidate independent variables (POS, TV, and Radio) in which to establish a predictive model that can show the relationship between product sales and those independent variables. Just as a line chart was employed to reveal the behavior of product sales and the other variables in the descriptive analytic step, a statistical method can establish a linear model that combines the three predictive variables. We will use multiple regression, which can incorporate any of the multiple independent variables, to establish a relational model for product sales in this case study. Multiple regression also can be used to continue our exploration of the candidacy of the three independent variables.

The procedure by which multiple regression can be used to evaluate which independent variables are best to include or exclude in a linear model is called *step-wise multiple regression*. It is based on an evaluation of regression models and their validation statistics—specifically, the multiple correlation coefficients and the F-ratio from an ANOVA. SPSS software and many other statistical systems build in the step-wise process. Some are called *backward step-wise regression* and some are called *forward step-wise regression*. The backward step-wise regression starts with all the independent variables placed in the model, and the step-wise process removes them one at a time based on worst predictors first until a statistically significant model emerges. The forward step-wise regression starts with the best related variable (using correction analysis as a guide), and then step-wise adds other variables until adding more will no longer improve the accuracy of the model. The forward step-wise regression process will be illustrated here manually. The first step is to generate individual regression models and statistics for each independent variable with the dependent variable one at a time. These three models are presented in **Tables 6.5**, **6.6**, and **6.7** for the POS, radio, and TV variables, respectively. The comparable Excel regression statistics are presented in **Tables 6.8**, **6.9** and **6.10** for the POS, radio, and TV variables, respectively.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.013[a]	.000	-.055	2688.55013	.000	.003	1	18	.958

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16649.445	1398.322		11.907	.000
	POS	44.140	822.471	.013	.054	.958

a. Dependent Variable: Sales

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20819.162	1	20819.162	.003	.958[b]
	Residual	130109432.038	18	7228301.780		
	Total	130130251.200	19			

a. Dependent Variable: Sales / b. Predictors: (Constant), POS

Table 6.5 SPSS POS Regression Model: Marketing/Planning Case Study

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.977[a]	.955	.952	571.64681	.955	380.220	1	18	.000

a. Predictors: (Constant), Radio

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9741.921	1362.939		-7.148	
	Radio	347.689	17.831	.977	19.499	

a. Dependent Variable: Sales

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	124248209.906	1	124248209.906	380.220	.000[b]
	Residual	5882041.294	18	326780.072		
	Total	130130251.200	19			

a. Dependent Variable: Sales / b. Predictors: (Constant), Radio

Table 6.6 SPSS Radio Regression Model: Marketing/Planning Case Study

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.958[a]	.918	.913	$771.31951	.918	200.731	1	18	.000

a. Predictors: (Constant), TV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-42229.208	4164.121		-10.141	.000
	TV	221.104	15.606	.958	14.168	.000

a. Dependent Variable: Sales

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	119421442.977	1	119421442.977	200.731	.000[b]
	Residual	10708808.223	18	594933.790		
	Total	130130251.200	19			

a. Dependent Variable: Sales/ b. Predictors: (Constant), TV

Table 6.7 SPSS TV Regression Model: Marketing/Planning Case Study

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.0126486
R Square	0.00015999
Adjusted R Square	-0.0553867
Standard Error	2688.55013
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	20819.162	20819.16	0.00288	0.957791029
Residual	18	130109432	7228302		
Total	19	130130251.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	16649.4448	1398.322032	11.90673	5.72E-10	13711.67923	19587.21	13711.6792	19587.21039
POS	44.14019	822.4712269	0.053668	0.957791	-1683.807738	1772.0881	-1683.80774	1772.088118

Table 6.8 Excel POS Regression Model: Marketing/Planning Case Study

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.977138075
R Square	0.954798817
Adjusted R Square	0.95228764
Standard Error	571.646807
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	124248209.9	124248210	380.2197	1.492E-13
Residual	18	5882041.294	326780.072		
Total	19	130130251.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-9741.92148	1362.939419	-7.1477289	1.17E-06	-12605.35	-6878.492	-12605.351	-6878.49202
Radio	347.68885	17.83090866	19.4992222	1.49E-13	310.2275	385.150199	310.227501	385.150199

Table 6.9 Excel Radio Regression Model: Marketing/Planning Case Study

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9579703
R Square	0.917707
Adjusted R Square	0.9131352
Standard Error	771.31951
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	119421443	1.19E+08	200.7306	3.336E-11
Residual	18	10708808.22	594933.8		
Total	19	130130251.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-42229.21	4164.121037	-10.1412	7.19E-09	-50977.7	-33480.714	-50977.702	-33480.7145
TV	221.10431	15.60595543	14.16794	3.34E-11	188.31741	253.8912	188.317411	253.8912023

Table 6.10 Excel TV Regression Model: Marketing/Planning Case Study

The computer printouts in the tables provide a variety of statistics for comparative purposes. Discussion will be limited here to just a few. The R-Square statistics are a precise proportional measure of the variation that is explained by the independent variable's behavior with the dependent variable. The closer the R-Square to 1.00, the more of the variation is explained, and the better the predictive variable. The three variables' R-Squares are 0.000 (POS), 0.955 (radio), and 0.918 (TV). Clearly, radio is the best predictor variable of the three, followed by TV and, without almost any relationship, POS. This latter result was expected based on the prior Pearson correlation. What it is suggesting is that only 0.082 percent (1.000–0.918) of the variation in product sales is explained by TV commercials.

From ANOVA, the F-ratio statistic is useful in actually comparing the regression model's capability to predict the dependent variable. As R-Square increases, so does the F-ratio because of the way in which they are computed and what is measured by both. The larger the F-ratio (like the R-Square statistic), the greater the statistical significance in explaining the variable's relationships. The three variables' F-ratios from the ANOVA tables are 0.003 (POS), 380.220 (radio), and 200.731 (TV). Both radio and TV are statistically significant, but POS has an insignificant relationship. To give some idea of how significant the relationships are, assuming a level of significance where α=0.01, one would only need a cut-off value for the F-ratio of 8.10 to designate it as being significant. Not exceeding that F-ratio (as in the case of POS at 0.003) is the same as saying that the coefficient in the regression model for POS is no different from a value of zero (no contribution to Product Sales). Clearly, the independent variables radio and TV appear to have strong relationships with the dependent variable. The question is whether the two combined or even three variables might provide a more accurate forecasting model than just using the one best variable like radio.

Continuing with the step-wise multiple regression procedure, we next determine the possible combinations of variables to see if a particular combination is better than the single variable models computed previously. To measure this, we have to determine the possible combinations for the variables and compute their regression models. The combinations are (1) POS and radio, (2) POS and TV, (3) POS, radio, and TV, and (4) radio and TV.

The resulting regression model statistics are summarized and presented in Table 6.11. If one is to base the selection decision solely on the R-Square statistic, there is a tie between the POS/radio/TV and the radio/TV combination (0.979 R-Square values). If the decision is based solely on the F-ratio value from ANOVA, one would select just the radio/TV combination, which one might expect of the two most significantly correlated variables.

To aid in supporting a final decision and to ensure these analytics are the best possible estimates, an additional statistic can be considered. That tie breaker is the R-Squared (Adjusted) statistic, which is commonly used in multiple regression models.

Variable Combination	R-Square	R-Square (Adjusted)	F-Ratio
POS/radio	0.957	0.952	188.977
POS/TV	0.920	0.911	97.662
POS/radio/TV	0.979	0.951	123.315
Radio/TV	0.979	0.953	192.555

Table 6.11 SPSS Variable Combinations and Regression Model Statistics: Marketing/Planning Case Study

The *R-Square Adjusted* statistic does not have the same interpretation as R-Square (a precise, proportional measure of variation in the relationship). It is instead a comparative measure of suitability of alternative independent variables. It is ideal for selection between independent variables in a multiple regression model. The R-Square adjusted seeks to take into account the phenomenon of the R-Square automatically increasing when additional independent variables are added to the model. This phenomenon is like a painter putting paint on a canvas, where more paint additively increases the value of the painting. Yet by continually adding paint, there comes a point at which some paint covers other paint, diminishing the value of the original. Similarly, statistically adding more variables should increase the ability of the model to capture what it seeks to model. On the other hand, putting in too many variables, some of which may be poor predictors, might bring down the total predictive ability of the model. The R-Square adjusted statistic provides some information to aid in revealing this behavior.

The value of the R-Square adjusted statistic can be negative, but it will always be less than or equal to that of the R-Square in which it is related. Unlike R-Square, the R-Square adjusted increases when a new independent variable is included only if the new variable improves the R-Square more than would be expected in the absence of any independent value being added. If a set of independent variables is introduced into a regression model one at a time in forward step-wise regression using the highest correlations ordered first, the R-Square adjusted statistic will end up being equal to or less than the R-Square value of the original model. By systematic experimentation with the R-Square adjusted recomputed for each added variable or combination, the value of the R-Square adjusted will reach a maximum and then decrease. The multiple regression model with the largest R-Square adjusted statistic will be the most accurate combination of having the best fit without excessive or unnecessary independent variables. Again, just putting all the variables into a model may add unneeded variability, which can decrease its accuracy. Thinning out the variables is important.

Finally, in the step-wise multiple regression procedure, a final decision on the variables to be included in the model is needed. Basing the decision on the R-Square adjusted, the best combination is radio/TV. The SPSS multiple regression model and support statistics are presented in Table 6.12, and the Excel model is shown in Table 6.13.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.979[a]	.958	.953	568.87547	.958	192.555	2	17	.000

a. Predictors: (Constant), TV, Radio

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-17150.455	6,965.591		-2.462	.025
	Radio	275.691	68.728	.775	4.011	.001
	TV	48.341	44.580	.209	1.084	.293

a. Dependent Variable: Sales

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	124628723.140	2	62314361.570	192.555	.000[b]
	Residual	5501528.060	17	323619.298		
	Total	130130251.200	19			

a. Dependent Variable: Sales / b. Predictors: (Constant), TV, Radio

Table 6.12 SPSS Best Variable Combination Regression Model and Statistics: Marketing/Planning Case Study

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.97863319
R Square	0.95772291
Adjusted R Square	0.95274914
Standard Error	568.875468
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	124628723.1	62314362	192.5545	2.09842E-12
Residual	17	5501528.06	323619.3		
Total	19	130130251.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-17150.4554	6965.590997	-2.46217	0.024791	-31846.56777	-2454.343	-31846.5678	-2454.342976
Radio	275.69065	68.72801022	4.011329	0.000905	130.6872233	420.694077	130.687223	420.6940765
TV	48.3405736	44.5804165	1.084345	0.293351	-45.71588363	142.397031	-45.7158836	142.3970308

Table 6.13 Excel Best Variable Combination Regression Model and Statistics: Marketing/Planning Case Study

Although there are many other additional analyses that could be performed to validate this model, we will use the SPSS multiple regression model in **Table 6.12** for the firm in this case study. The forecasting model can be expressed as follows:

$Y_p = -17150.455 + 275.691\ X_1 + 48.341\ X_2$

where:

Y_p = the estimated number of dollars of product sales

X_1 = the number of dollars to invest in radio commercials

X_2 = the number of dollars to invest in TV commercials

Because all the data used in the model is expressed as dollars, the interpretation of the model is made easier than using more complex data. The interpretation of the multiple regression model suggests that for every dollar allocated to radio commercials (represented by X_1), the firm will receive $275.69 in product sales (represented by Y_p in the model). Likewise, for every dollar allocated to TV commercials (represented by X_2), the firm will receive $48.34 in product sales.

A caution should be mentioned on the results of this case study. Many factors might challenge a result, particularly those derived from using powerful and complex methodologies like multiple regression. As such, the results may not occur as estimated, because the model is not reflecting past performance. What is being suggested here is that more analysis can always be performed in questionable situations. Also, additional analysis to confirm a result should be undertaken to strengthen the trust that others must have in the results to achieve the predicted higher levels of business performance.

In summary, for this case study, the predictive analytics analysis has revealed a more detailed, quantifiable relationship between the generation of product sales and the sources of promotion that best predict sales. The best way to allocate the $350,000 budget to maximize product sales might involve placing the entire budget into radio commercials because they give the best return per dollar of budget. Unfortunately, there are constraints and limitations regarding what can be allocated to the different types of promotional methods. Optimizing the allocation of a resource and maximizing business performance necessitate the use of special business analytic methods designed to accomplish this task. This requires the additional step of prescriptive analytics analysis in the BA process, which will be presented in the last section of **Chapter 7**.

Summary

This chapter dealt with the predictive analytics step in the BA process. Specifically, it discussed logic-driven models based on experience and aided by methodologies like the cause-and-effect and the influence diagrams. This chapter also defined data-driven models useful in the predictive step of the BA analysis. A further discussion of data mining was presented. Data mining methodology such as neural networks, discriminant analysis, logistic regression, and hierarchical clustering was described. An illustration of K-mean clustering using Excel was presented. Finally, this chapter discussed the second installment of a case study illustrating the predictive analytics step of the BA process. The remaining installment of the case study will be presented in **Chapter 7**.

Once again, several of this book's appendixes are designed to augment the chapter material by including technical, mathematical, and statistical tools. For both a greater understanding of the methodologies discussed in this chapter and a basic review of statistical and other quantitative methods, a review of the appendixes is recommended.

As previously stated, the goal of using predictive analytics is to generate a forecast or path for future improved business performance. Given this predicted path, the question now is how to exploit it as fully as possible. The purpose of the prescriptive analytics step in the BA process is to serve as a guide to fully maximize the outcome in using the information provided by the predictive analytics step. The subject of **Chapter 7** is the prescriptive analytics step in the BA process.

Discussion Questions

1. Why is predictive analytics analysis the next logical step in any business analytics (BA) process?

2. Why would one use logic-driven models to aid in developing data-driven models?

3. How are neural networks helpful in determining both associations and classification tasks required in some BA analyses?

4. Why is establishing clusters important in BA?
5. Why is establishing associations important in BA?

6. How can F-tests from the ANOVA be useful in BA?

Problems

1. Using the equation developed in this chapter for predicting dollar product sales (note below), what is the forecast for dollar product sales if the firm could invest $70,000 in radio commercials and $250,000 in TV commercials?

$Y_p = -17150.455 + 275.691\ X_1 + 48.341\ X_2$

where:

Y_p = the estimated number of dollars of product sales

X_1 = the number of dollars to invest in radio commercials

X_2 = the number of dollars to invest in TV commercials

2. Using the same formula as in Question 1, but now using an investment of $100,000 in radio commercials and $300,000 in TV commercials, what is the prediction on dollar product sales?

3. Assume for this problem the following table would have held true for the resulting marketing/planning case study problem. Which combination of variables is estimated here to be the best predictor set? Explain why.

Variable Combination	R-Square	R-Square (Adjusted)	F-Ratio
POS/radio	0.057	0.009	2.977
POS/TV	0.120	0.100	3.662
POS/radio/TV	0.179	0.101	4.315
Radio/TV	0.879	0.853	122.555

4. Assume for this problem that the following table would have held true for the resulting marketing/planning case study problem. Which of the variables is estimated here to be the best predictor? Explain why.

Statistic	Radio	Paper	TV	POS
Pearson correlation r with product sales	.127	.083	.208	.013
Significance test (1-tailed)*	.212	.313	.192	.479

5. Given the coefficients table that follows, what is the resulting regression model for TV and product sales? Is TV a good predictor of product sales according to this SPSS printout? Explain.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.912[a]	.900	.913	$771.31951	.918	200.731	1	18	.000

a. Predictors: (Constant), TV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-45000.000	4164.121		-10.141	.000
	TV	200.000	15.606	.958	14.168	.000

a. Dependent Variable: Sales

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	119421442.977	1	119421442.977	200.731	.000[b]
	Residual	10708808.223	18	594933.790		
	Total	130130251.200	19			

a. Dependent Variable: Sales/ b. Predictors: (Constant), TV

7. What Are Prescriptive Analytics?

Chapter objectives:

• List and describe the commonly used prescriptive analytics in the business analytics (BA) process.

• Explain the role of case studies in prescriptive analytics.

• Explain how curve fitting can be used in prescriptive analytics.

• Explain how to formulate a linear programming model.

• Explain the value of linear programming in the prescriptive analytics step of BA.

7.1. Introduction

After undertaking the descriptive and predictive analytics steps in the BA process, one should be positioned to undertake the final step: prescriptive analytics analysis. The prior analysis should provide a forecast or prediction of what future trends in the business may hold. For example, there may be significant statistical measures of increased (or decreased) sales, profitability trends accurately measured in dollars for new market opportunities, or measured cost savings from a future joint venture.

If a firm knows where the future lies by forecasting trends, it can best plan to take advantage of possible opportunities that the trends may offer. Step 3 of the BA process, prescriptive analytics, involves the application of decision science, management science, or operations research methodologies to make best use of allocable resources. These are mathematically based methodologies and algorithms designed to take variables and other parameters into a quantitative framework and generate an optimal or near-optimal solution to complex problems. These methodologies can be used to optimally allocate a firm's limited resources to take best advantage of the opportunities it has found in the predicted future trends. Limits on human, technology, and financial resources prevent any firm from going after all the opportunities. Using prescriptive analytics allows the firm to allocate limited resources to optimally or near-optimally achieve the objectives as fully as possible.

In **Chapter 3**, "**What Resource Considerations Are Important to Support Business Analytics?**" the relationships of methodologies to the BA process were expressed as a function of certification exam content. The listing of the prescriptive analytic methodologies as they are in some cases utilized in the BA process is again presented in **Figure 7.1** to form the basis of this chapter's content.

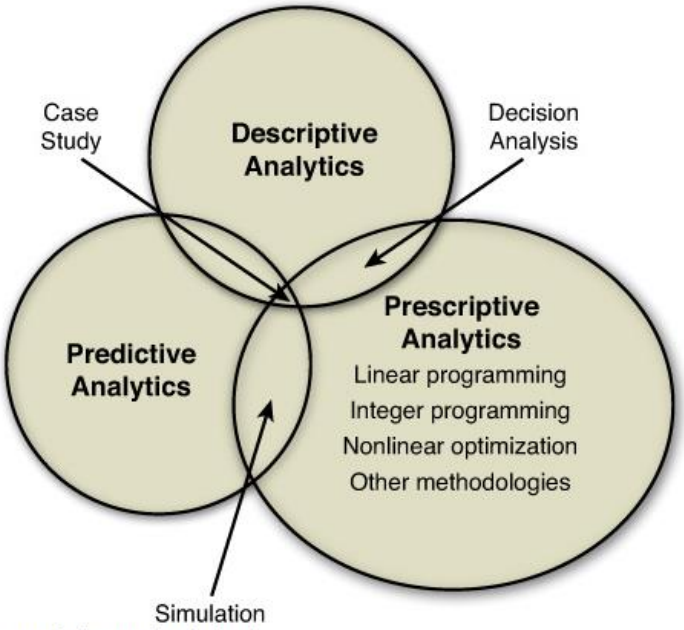


Figure 7.1 Prescriptive analytic methodologies

7.2. Prescriptive Modeling

The listing of prescriptive analytic methods and models in **Figure 7.1** is but a small grouping of many operations research, decision science, and management science methodologies that are applied in this step of the BA process. The explanation and use of most of the methodologies in **Table 7.1** are explained throughout this book. (See Additional Information column in **Table 7.1**.)

Data-Driven Models	Possible Applications	Additional Information
Linear Programming	A general-purpose modeling methodology is applied to multiconstrained, multivariable problems when an optimal solution is sought. It is ideal for complex and large-scale problems where limited resources are being allocated to multiple uses. Examples include allocating advertising budgets to differing media, allocating human and technology resources to product production, and optimizing blends of mixing ingredients to minimize costs of food products.	Chapters 7 and 8, "A Final Case Study Illustration" Appendix B, "Linear Programming" Appendix C, "Duality and Sensitivity Analysis in Linear Programming"
Integer Programming	This is the same as LP, but it permits decision variables to be integer values. Examples include allocating stocks to portfolios, allocating personnel to jobs, and allocating types of crops to farm lands.	Appendix D, "Integer Programming"
Nonlinear Optimization	A large class of methodologies and algorithms is used to analyze and solve for optimal or near-optimal solutions when the behavior of the data is nonlinear. Examples include solving for optimized allocations of human, technology, and systems whose data appears to form a cost or profit function that is quadratic, cubic, or nonlinear in some way.	Chapters 7 and 8 Appendix E, "Forecasting"

Optimization	used to analyze and solve for optimal or near-optimal solutions when the behavior of the data is nonlinear. Examples include solving for optimized allocations of human, technology, and systems whose data appears to form a cost or profit function that is quadratic, cubic, or nonlinear in some way.	Appendix E, "Forecasting"
Decision Analysis	A set of methodologies, models, or principles is used to analyze and guide decision-making when multiple choices face the decision maker in differing decision environments (for example, certainty, risk, and uncertainty). Examples include selecting one from a set of computer systems, trucks, or site locations for a service facility.	Appendix G, "Decision Theory"
Case Studies	A learning aid provides practical experience by offering real or hypothetical case studies of real-world applications of BA. For example, case studies can simulate the issues and challenges in an actual problem setting. This kind of simulation can prep decision makers to anticipate and prepare for what has been predicted to occur by the predicted analytics step in the BA process. For example, a case study discussion on how to cope with organization growth might provide a useful decision-making environment for a firm whose analytics have predicted growth in the near future.	This is beyond the scope of this book. See Sekaran and Bougie (2013); Adkins (2006).
Simulation	This methodology can be used in prescriptive analysis in situations where parameters are probabilistic, nonlinear or just too complex to use with other optimizations models that require deterministic or linear behavior. For example, a bank might want to simulate the transactions they currently use to process a loan application to determine if changes in the process might reduce time and improve performance. The simulation model might be used to test alternative process scenarios.	Appendix F, "Simulation"
Others Methodologies	The areas of operations research, decision sciences, and management science combine the application of mathematics, engineering, and computer science to offer a broad listing of prescriptive methodologies. These other methodologies include network modeling, project scheduling, dynamic programming, queuing models, decision support systems, heuristics, artificial intelligence, expert systems, Markov processes, decision tree analysis, game theory, goal programming, nonlinear programming, reliability analysis, genetic programming, and data envelopment analysis, just to name a few. There are virtually no application limitations of the collection of these methodologies.	These are outside the scope of this book. See Hillier, F.S. (2014); Cooper et al. (2013); Rothlauf (2013); Liebowitz (2014); Albright and Winston (2014).

Table 7.1 Select Prescriptive Analytic Models

7.3. Nonlinear Optimization

The prescriptive methodologies in **Table 7.1** are explained in detail in the referenced chapters and appendixes, but nonlinear optimization will be discussed here. When business performance cost or profit functions become too complex for simple linear models to be useful, exploration of nonlinear functions is a standard practice in BA. Although the predictive nature of exploring for a mathematical expression to denote a trend or establish a forecast falls mainly in the predictive analytics step of BA, the use of the nonlinear function to optimize a decision can fall in the prescriptive analytics step.

As mentioned previously, there are many mathematical programing nonlinear methodologies and solution procedures designed to generate optimal business performance solutions. Most of them require careful estimation of parameters that may or may not be accurate, particularly given the precision required of a solution that can be so precariously dependent upon parameter accuracy. This precision is further complicated in BA by the large data files that should be factored into the model-building effort.

To overcome these limitations and be more inclusive in the use of large data, regression software can be applied. As illustrated in **Appendix E**, Curve Fitting software can be used to generate predictive analytic models that can also be utilized to aid in making prescriptive analytic decisions.

For purposes of illustration, SPSS's Curve Fitting software will be used in this chapter. Suppose that a resource allocation decision is being faced whereby one must decide how many computer servers a service facility should purchase to optimize the firm's costs of running the facility. The firm's predictive analytics effort has shown a growth trend. A new facility is called for if costs can be minimized. The firm has a history of setting up large and small service facilities and has collected the 20 data points in **Figure 7.2**. Whether there are 20 or 20,000 items in the data file, this SPSS function fits the data based on regression mathematics to a nonlinear line that best minimizes the distance from the data items to the line. The software then converts the line into a mathematical expression useful for forecasting.

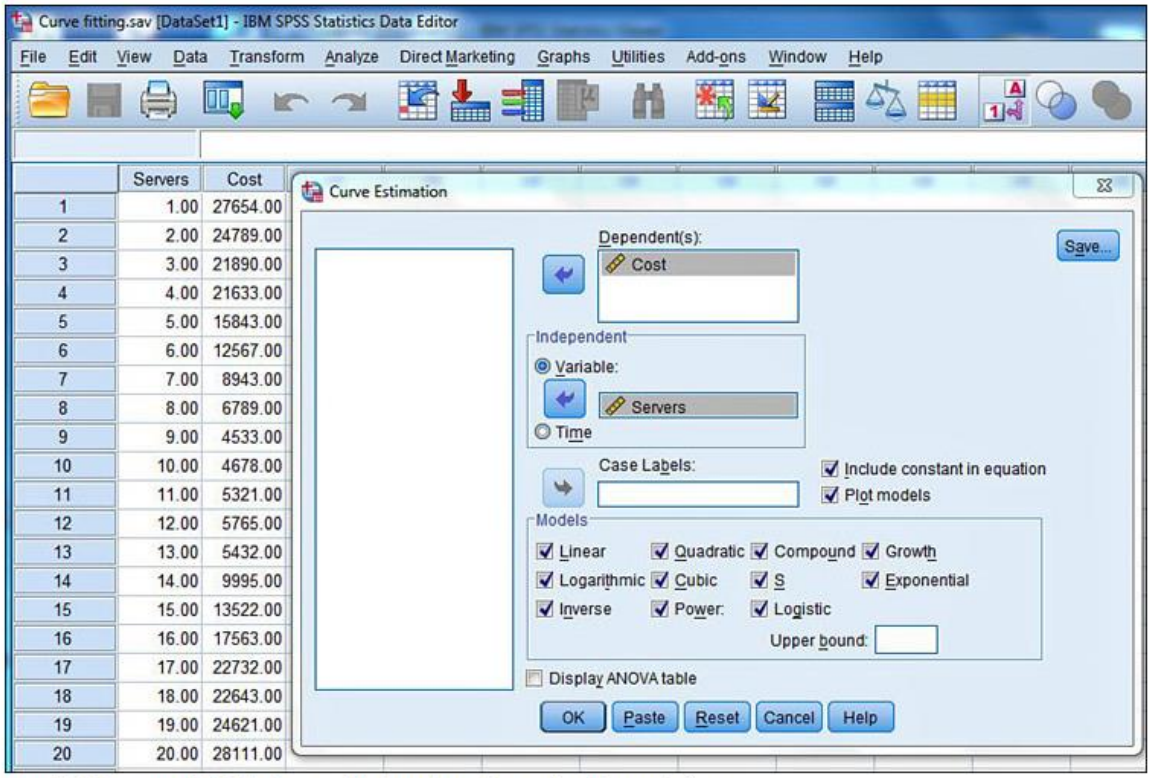


Figure 7.2 Data and SPSS Curve Fitting function selection window

In this server problem, the basic data has a u-shaped function, as presented in Figure 7.3. This is a classic shape for most cost functions in business. In this problem, it represents the balancing of having too few servers (resulting in a costly loss of customer business through dissatisfaction and complaints with the service) or too many servers (excessive waste in investment costs as a result of underutilized servers). Although this is an overly simplified example with little and nicely ordered data for clarity purposes, in big data situations, cost functions are considerably less obvious.

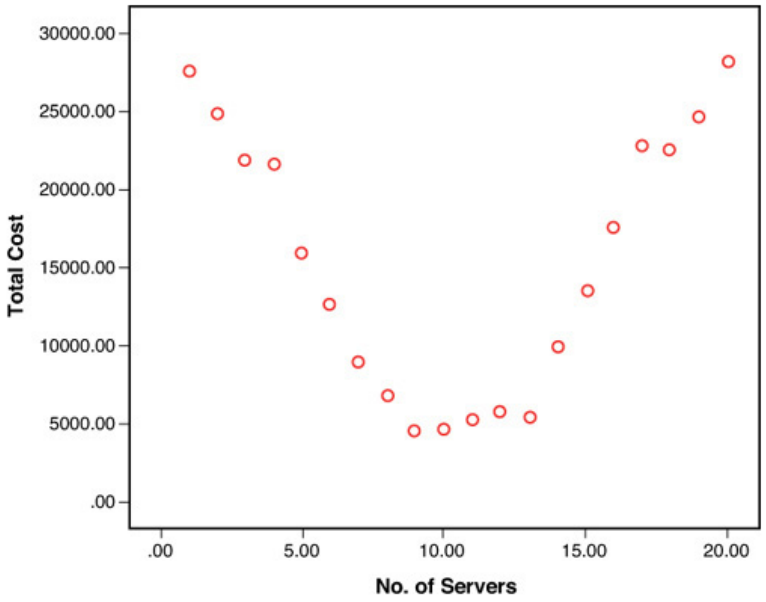


Figure 7.3 Server problem basic data cost function

The first step in using the curve-fitting methodology is to generate the best-fitting curve to the data. By selecting all the SPSS models in Figure 7.2, the software applies each point of data using the regression

process of minimizing distance from a line. The result is a series of regression models and statistics, including ANOVA and other testing statistics. It is known from the previous illustration of regression that the adjusted R-Square statistic can reveal the best estimated relationship between the independent (number of servers) and dependent (total cost) variables. These statistics are presented in **Table 7.2**. The best adjusted R-Square value (the largest) occurs with the quadratic model, followed by the cubic model. The more detailed supporting statistics for both of these models are presented in **Table 7.3**. The graph for all the SPSS curve-fitting models appears in **Figure 7.4**.

Linear Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.034	.001	-.054	8687.290

The independent variable is Servers.

Logarithmic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.267	.071	.020	8376.020

Inverse Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.435	.189	.144	7825.696

Quadratic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.965	.931	.923	2342.315

Cubic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.965	.932	.919	2404.009

Compound Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.025	.001	-.055	.677

Power Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.247	.061	.009	.657

S Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.381	.145	.098	.627

Growth Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.025	.001	-.055	.677

Exponential Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.025	.001	-.055	.677

Logistic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.025	.001	-.055	.677

Table 7.2 Adjusted R-Square Values of All SPSS Models

Quadratic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.965	.931	.923	2342.315

The independent variable is Servers.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1266704838.323	2	633352419.161	115.440	.000
Residual	93269442.877	17	5486437.816		
Total	1359974281.200	19			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Servers	-5589.432	382.188	-3.909	-14.625	.000
Servers ** 2	268.449	17.678	4.058	15.186	.000
(Constant)	35417.772	1742.639		20.324	.000

Cubic Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.965	.932	.919	2404.009

The independent variable is Servers.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1267506094.832	3	422502031.611	73.107	.000
Residual	92468186.368	16	5779261.648		
Total	1359974281.200	19			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Servers	-5954.738	1056.596	-4.164	-5.636	.000
Servers ** 2	310.895	115.431	4.700	2.693	.016
Servers ** 3	-1.347	3.619	-.399	-.372	.715
(Constant)	36133.696	2625.976		13.760	.000

Table 7.3 Quadratic and Cubic Model SPSS Statistics

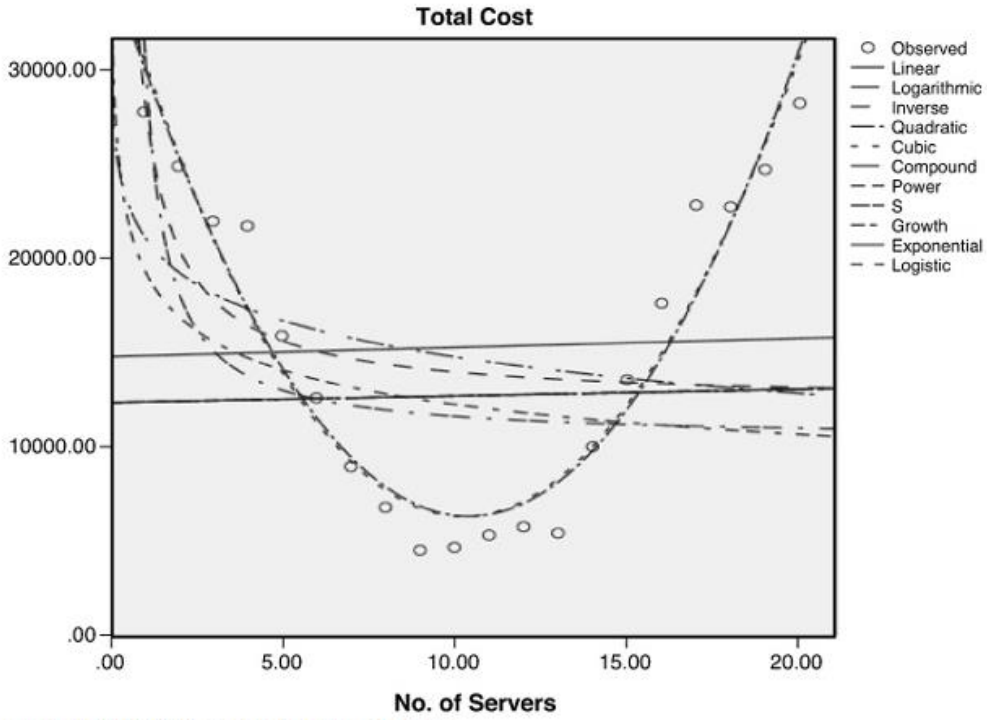


Figure 7.4 Graph of all SPSS curve-fitting models

From Table 7.3, the resulting two statistically significant curve-fitted models follow:

$Y_p = 35417.772 - 5589.432\ X + 268.445\ X^2$ [Quadratic model]

$Y_p = 36133.696 - 5954.738\ X + 310.895\ X^2 - 1.347\ X^3$ [Cubic model]

where:

Y_p = the forecasted or predicted total cost, and

X = can be the number of computer servers.

For purposes of illustration, we will use the quadratic model. In the next step of using the curve-fitting models, one can either use calculus to derive the cost minimizing value for X (number of servers) or perform a deterministic simulation where values of X are substituted into the model to compute and predict the total cost (Y_p). The calculus-based approach is presented in the "**Addendum**" section of this chapter.

As a simpler solution method to finding the optimal number of servers, simulation can be used. Representing a deterministic simulation (see **Appendix F**, **Section F.2.1**), the resulting costs of servers can be computed using the quadratic model, as presented in **Figure 7.5**. These values were computed by plugging the number of server values (1 to 20) into the Y_p quadratic function one at a time to generate the predicted values for each of the server possibilities. Note that the lowest value in these predicted values occurs with the acquisition of 10 servers at $6357.952, and the next lowest is at 11 servers at $6415.865. In the actual data in **Figure 7.2**, the minimum total cost point occurs at 9 servers at $4533, whereas the next lowest total cost is $4678 occurring at 10 servers. The differences are due to the estimation process of curve fitting. Note in **Figure 7.3** that the curve that is fitted does not touch the lowest 5 cost values. Like regression in general, it is an estimation process, and although the ANOVA statistics in the quadratic model demonstrate a strong relationship with the actual values, there is some error. This process provides a near-optimal solution but does not guarantee one.

	A	G
1	1	30096.79
2	2	25312.69
3	3	21065.48
4	4	17355.16
5	5	14181.74
6	6	11545.2
7	7	9445.553
8	8	7882.796
9	9	6856.929
10	10	6367.952
11	11	6415.865
12	12	7000.668
13	13	8122.361
14	14	9780.944
15	15	11976.42
16	16	14708.78
17	17	17978.03
18	18	21784.18
19	19	26127.21
20	20	31007.13

Figure 7.5 Predicted total cost in server problem for each server alternative

Like all regression models, curve fitting is an estimation process and has risks, but the supporting statistics, like ANOVA, provide some degree of confidence in the resulting solution.

Finally, it must be mentioned that many other nonlinear optimization methodologies exist. Some, like quadratic programming, are considered constrained optimization models (like LP). These topics are beyond the scope of this book. For additional information on nonlinear programming, see **King and Wallace (2013)**, **Betts (2009)**, and **Williams (2013)**. Other methodologies, like the use of calculus in this chapter, are useful in solving for optimal solutions in unconstrained problem settings. For additional information on calculus methods, see **Spillers and MacBain (2009)**, **Luptacik (2010)**, and **Kwak and Schniederjans (1987)**.

7.4. Continuation of Marketing/Planning Case Study Example: Prescriptive Step in the BA Analysis

In Chapter 5, "What Are Descriptive Analytics?" and Chapter 6, "What Are Predictive Analytics?" an ongoing marketing/planning case study was presented to provide an illustration of some of the tools and strategies used in a BA problem analysis. This is the third and final installment of the case study dealing with the prescriptive analytics step in BA.

7.4.1. Case Background Review

The predictive analytics analysis in Chapter 6 revealed a statistically strong relationship between radio and TV commercials that might be useful in predicting future product sales. The ramifications of these results suggest a better allocation of funds away from paper and POS ads to radio and TV commercials. Determining how much of the $350,000 budget should be allocated between the two types of commercials requires the application of an optimization decision-making methodology.

7.4.2. Prescriptive Analysis

The allocation problem of the budget to purchase radio and TV commercials is a multivariable (there are two media to consider), constrained (there are some limitations on how one can allocate the budget funds), optimization problem (BA always seeks to optimize business performance). Many optimization methods could be employed to determine a solution to this problem. Considering the singular objective of maximizing estimated product sales, linear programming (LP) is an ideal methodology to apply in this situation. To employ LP to model this problem, use the six-step LP formulation procedure explained in Appendix B.

7.4.2.1. Formulation of LP Marketing/Planning Model

In the process of exploring the allocation options, a number of limitations or constraints on placing radio and TV commercials were observed. The total budget for all the commercials was set at a maximum of $350,000 for the next monthly campaign. To receive the price discount on radio commercials, a minimum budget investment in radio of $15,000 is required, and to receive the price discount on TV commercials, a minimum of $75,000 is necessary. Because the radio and TV stations are owned by the same corporation, there is an agreement that for every dollar of radio commercials required, the client firm must purchase $2 in TV commercials. Given these limitations and the modeled relationship found in the previous predictive analysis, one can formulate the budget allocation decision as an LP model using a five-step LP formulation procedure (see Appendix B, Section B.4.1):

1. Determine the type of problem—This problem seeks to maximize dollar product sales by determining how to allocate budget dollars over radio and TV commercials. For each dollar of radio commercials estimated with the regression model, $275.691 will be received, and for each dollar of TV commercials, $48.341 will be received. Those two parameters are the product sales values to maximize. Therefore, it will be a maximization model.

2. Define the decision variables—The decision variables for the LP model are derived from the multiple regression model's independent variables. The only adjustment is the monthly timeliness of the allocation of the budget:

X_1= the number of dollars to invest in radio commercials for the next monthly campaign

X_2=the number of dollars to invest in TV commercials for the next monthly campaign

3. Formulate the objective function—Because the multiple regression model defines the dollar sales as a linear function with the two independent variables, the same dollar coefficients from the regression model can be used as the contribution coefficients in the objective function. This results in the following LP model objective function:

Maximize: $Z = 275.691 X_1 + 48.341 X_2$

4. Formulate the constraints—Given the information on the limitations in this problem, there are four constraints:

Constraint 1—No more than $350,000 is allowed for the total budget to allocate to both radio (X_1) and TV (X_2) commercials. So add $X_1 + X_2$ and set it less than or equal to 350,000 to formulate the first constraint as follows:

$X_1 + X_2 \leq 350000$

Constraint 2—To get a discount on radio (X_1) commercials, a minimum of $15,000 must be allocated to radio. The constraint for this limitation follows:

$X_1 \geq 15000$

Constraint 3—Similar to Constraint 2, to get a discount on TV (X_2) commercials, a minimum of $75,000 must be allocated to TV. The constraint for this limitation follows:

$X_2 \geq 75000$

Constraint 4—This is a blending problem constraint (see **Appendix B**, **Section B.6.3**). What is needed is to express the relationship as follows:

$$\frac{X_1}{1} = \frac{X_2}{2}$$

which is to say, for each one unit of X_1, one must acquire two units of X_2. Said differently, the ratio of one unit of X_1 is equal to two units of X_2. Given the expression, use algebra to cross-multiply such that:

$2 X_1 = X_2$

Convert it into an acceptable constraint with a constant on the right side and the variables on the left side as follows:

$2 X_1 - X_2 = 0$

5. State the Nonnegativity and Given Requirements—With only two variables, this formal requirement in the formulation of an LP model is expressed as follows:

$X_1, X_2 \geq 0$

Because these variables are in dollars, they do not have to be integer values. (They can be any real or cardinal number.) The complete LP model formulation is given here:

$$\begin{aligned} \text{Maximize: } Z &= 275.691\,X_1 + 48.341\,X_2 \\ \text{Subject to: } & X_1 + X_2 \leq 350000 \\ & X_1 \geq 15000 \\ & X_2 \geq 75000 \\ & 2X_1 - X_2 = 0 \\ \text{and } & X_1, X_2 \geq 0 \end{aligned}$$

7.4.2.2. Solution for the LP Marketing/Planning Model

From **Appendix B**, one knows that both Excel and LINGO software can be used to run the LP model and solve the budget allocation in this marketing/planning case study problem. For purposes of brevity, discussion will be limited to just LINGO. As presented in **Appendix B**, *LINGO* is a mathematical programming language and software system. It allows the fairly simple statement of the LP model to be entered into a single window and run to generate LP solutions.

LINGO opens with a blank window for entering whatever type of model is desired. After entering the LP model formulation into the LINGO software, the resulting data entry information is presented in **Figure 7.6**.

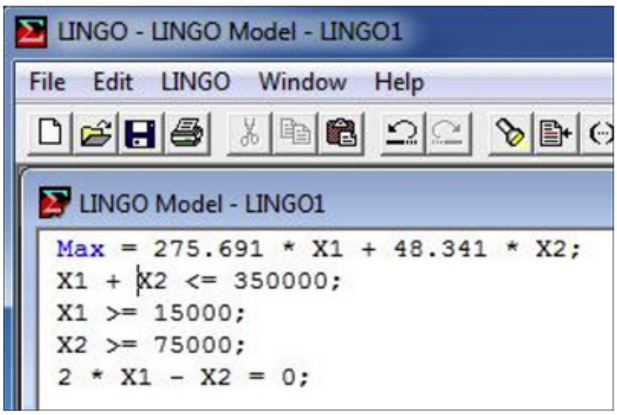


Figure 7.6 LINGO LP model entry requirements: marketing/planning case study

There are several minor differences in the model entry requirements over the usual LP model formulation. These differences are required to run a model in LINGO. These include (1) using the term "Max" instead of "Maximize," (2) dropping off "Subject to" and "and" in the model formulation, (3) placing an asterisk and a space between unknowns and constant values in the objective and constraint functions where multiplication is required, (4) ending each expression with a semicolon, and (5) omitting the nonnegativity requirements, which aren't necessary.

Having entered the model into LINGO, a single click on the SOLVE option in the bar at the top of the window generates a solution. The marketing budget allocation LP model solution is found in **Figure 7.7**.

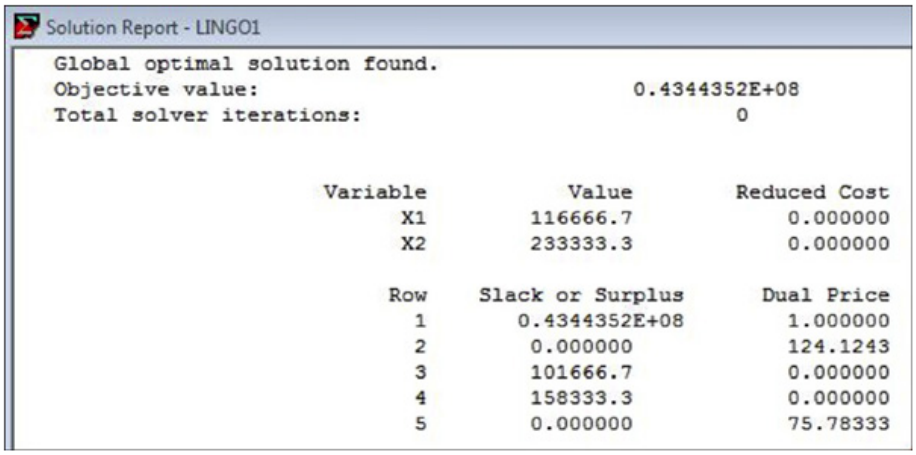


Figure 7.7 LINGO LP model solution: marketing/planning case study

As it turns out, the optimal distribution of the $350,000 promotion budget is to allocate $116,666.70 to radio commercials and $233,333.30 to TV commercials. The resulting Z value, which in this model is the total predicted product sales in dollars, is 0.4344352E+08, or $43,443,524. Comparing that future estimated month's product sales with the average current monthly product sales of $16,717,200 presented in **Figure 7.7**, it does appear that the firm in this case study will optimally maximize future estimated monthly product sales if it allocates the budget accordingly (that is, if the multiple regression model estimates and the other parameters in the LP model hold accurate and true).

In summary, the prescriptive analytics analysis step brings the prior statistical analytic steps into an applied decision-making process where a potential business performance improvement is shown to better this organization's ability to use its resources more effectively. The management job of monitoring performance and checking to see that business performance is in fact improved is a needed final step in the BA analysis. Without proof that business performance is improved, it's unlikely that BA would continue to be used.

7.4.2.3. Final Comment on the Marketing/Planning Model

Although the LP solution methodology used to generate an allocation solution guarantees an optimal LP solution, it does not guarantee that the firm using this model's solution will achieve the results suggested in the analysis. Like any estimation process, the numbers are only predictions, not assurances of outcomes. The high levels of significance in the statistical analysis and the added use of other conformational statistics (R-Square, adjusted R-Square, ANOVA, and so on) in the model development provide some assurance of predictive validity. There are many other methods and approaches that could have been used in this case study. Learning how to use more statistical and decision science tools helps ensure a better solution in the final analysis.

Summary

This chapter discussed the prescriptive analytics step in the BA process. Specifically, this chapter revisited and briefly discussed methodologies suggested in BA certification exams. An illustration of nonlinear optimization was presented to demonstrate how the combination of software and mathematics can generate useful decision-making information. Finally, this chapter presented the third installment of a marketing/planning case study illustrating how prescriptive analytics can benefit the BA process.

We end this book with a final application of the BA process. Once again, several of the appendixes are designed to augment this chapter's content by including technical, mathematical, and statistical tools. For both a greater understanding of the methodologies discussed in this chapter and a basic review of statistical and other quantitative methods, a review of the appendixes and chapters is recommended.

Addendum

The *differential calculus* method for finding the minimum cost point on the quadratic function that follows involves a couple of steps. It finds the zero slope point on the cost function (the point at the bottom of the u-shaped curve where a line could be drawn that would have a zero slope). There are limitations to its use, and qualifying conditions are required to prove minimum or maximum positions on a curve. The quadratic model in the server problem follows:

$Y_p = 35417.772 - 5589.432\ X + 268.445\ X^2$ [Quadratic model]

Step 1. Given the quadratic function above, take its first derivative:

$d(Y_p) = -5589.432 + 536.89\ X$

Step 2. Set the derivative function equal to zero and solve for X.

$0 = -5589.432 + 536.89\ X$

$X = 10.410758$

Slightly more than ten servers should be purchased at the resulting optimally minimized cost value. This approach provides a near-optimal solution but does not guarantee one. For additional information on the application of calculus, see **Field, M.J. (2012)** and **Dobrushkin, V.A. (2014)**.

Discussion Questions

1. How are prescriptive and descriptive analytics related?

2. How can we use simulation in both predictive and prescriptive analytics?

3. Why in the server problem were there so few statistically significant models?

4. Does it make sense that the resulting quadratic model in **Figure 7.4** did not touch the lowest cost data points in the data file? Explain.

5. What conditions allowed the application of LP?

Problems

1. A computer services company sells computer services to industrial users. The company's analytics officer has predicted the need for growth to meet competitive pressures. To implement this strategy, upper management has determined that the company would tactically expand its sales and service organization. In this expansion, new districts would be defined and newly hired or appointed managers would be

placed in charge to establish and run the new districts. The first job of the new district managers would be to select the sales people and staff support employees for their districts. To aid the new district managers in deciding on the number of sales people and staffers to hire, the company researched existing office operations and made a number of analytic-based observations, which they passed on to the new district managers. A new manager's district should, at the very least, have 14 sales people and 4 staffers to achieve adequate customer service. Research has indicated that a district manager could adequately manage the equivalent of no more than 32 employees. Sales people are twice as time consuming to manage as staffers. The district manager was assigned part of the floor in an office building for operations. This space could house no more than 20 sales people and staffers. The district manager had some discretion regarding budgetary limitations. A total payroll budget for sales people and staffers was set at $600,000. The manufacturing company's policy in developing a new territory would be to pay sales people a fixed salary instead of commissions and salary. The yearly salary of a beginning sales person would be $36,000, whereas a staffer would receive $18,000. All the sales people and staffers being hired for this district would be new with the company, and as such, would start with the basic salaries mentioned. Finally, the source of prospective sales people and staffers would be virtually unlimited in the district and pose no constraint on the problem situation. What is the LP formulation of this model?

2. (This problem requires computer support.) What is the optimal answer to the problem formulated in Problem 1?

3. A trucking firm must transport exactly 900, 800, 700, and 1,000 units of a product to four cities: A, B, C, and D. The product is manufactured and supplied in two other cities, X and Y, in the exact amounts to match the total demand. The production of units from the two cities is 1,900 and 1,500 units, respectively, to X and Y. The cost per unit to transport the product between the manufacturing plants in cities X and Y and the demand market cities A, B, C, and D are given here:

	Demand Market			
Supply Plant	**A**	**B**	**C**	**D**
X	.65	.70	.80	.90
Y	.60	.60	.80	.70

For example, in the table, $0.65 is the cost to ship one unit from Supply Plant X to Demand Market A. The trucking firm needs to know how many units should be shipped from each supply city to each demand city in such a way that it minimizes total costs. Hint: This is a multidimensional decision variable problem (see **Section B.6.4** in **Appendix B**). What is the LP model formulation for this problem?

4. (This problem requires computer support.) What is the optimal answer to the problem formulated in Problem 3?

8. A Final Business Analytics Case Problem

Chapter objectives:

• Provide a capstone business analytics (BA) overview within a case study problem.

• Show the step-wise connections of the descriptive, predictive, and prescriptive steps in the BA process.

8.1. Introduction

In **Parts I**, "**What Are Business Analytics?**" and **II**, "**Why Are Business Analytics Important?**" (**Chapters 1** through **3**), this book explained what BA is about and why it is important to business organization decision-making. In **Part III**, "**How Can Business Analytics Be Applied?**" (**Chapters 4** through **7**), we explained and illustrated how BA can be applied using a variety of different concepts and methodologies. Completing **Part III**, we seek in this chapter a closing illustration of how the BA process can be applied by presenting a final case study. This case study is meant as a capstone learning experience on the business analytics process discussed throughout the book. Several of the concepts and methodologies presented in prior chapters and the appendixes will once again be applied here.

As will be seen in this case study, unique metrics and measures are sometimes needed in a BA setting to affect a solution to a problem or answer a question. Therefore, the methodologies and approach used in this chapter should be viewed as just one approach in obtaining the desired information.

Undertaking the analytic steps in the BA process (see **Chapter 1**, "**What Are Business Analytics?**") requires a beginning effort that preempts data collection efforts. This prerequisite to BA is to understand the business systems that are a part of the problem. When BA effort has been outsourced (see **Chapter 4**, "**How Do We Align Resources to Support Business Analytics within an Organization?**") or when it is completely performed in-house by a BA team (**Chapter 3**, "**What Resource Considerations Are Important to Support Business Analytics?**"), experienced managers must be brought into the process to provide the necessary systems behavior and general knowledge of operations needed to eventually model and explain how the business operates. In this case study, it is assumed that the staff or information is available. Based on this information, a BA project can be undertaken.

8.2. Case Study: Problem Background and Data

A Midwest US commercial manufacturing firm is facing a supply chain problem. The manufacturer produces and sells a single product, a general-purpose small motor as a component part to different customers who incorporate the motor into their various finished products. The manufacturer has a supply chain network that connects production centers located in St. Louis, Missouri, and Dallas, Texas, with six warehouse facilities that serve commercial customers located in Kansas City, Missouri; Chicago, Illinois; Houston, Texas; Oklahoma City, Oklahoma; Omaha, Nebraska; and Little Rock, Arkansas.

Part of the supply chain problem is the need to keep the cost of shipping motors to the customers as low as possible. The manufacturer adopted a lean management philosophy that seeks to match what it produces with what is demanded at each warehouse. The problem with implementing this philosophy is complicated by the inability to forecast the customer demand month to month. If the forecast of customer

demand is too low and not enough inventory is available (an underage of inventory), the manufacturer has to rush order motors that end up being costly to the manufacturer. If the forecast is too high and the manufacturer produces and ships unwanted inventory (an overage of inventory), the warehouse incurs wasteful storage costs. The management of the manufacturing firm has decided that an analytics-based procedure needs to be developed to improve overall business performance. This would be a procedure that analysts could use each month to develop an optimal supply chain schedule of shipments from the two supply centers to the six warehouse demand destinations that would minimize costs. A key part of this procedure would be to include a means to accurately forecast customer demand and an optimization process for shipping products from the manufacturing centers to the warehouse demand destinations.

The manufacturing firm created a small BA team to develop the procedure (see **Chapter 4**, **Section 4.1.1**). The BA team consists of a BA analyst (who would be responsible for using the procedure and heads the BA team), the supply chain general manager, the shipping manager (responsible for drafting the shipping schedule), and a warehouse manager (whose job it is to develop monthly forecasts).

8.3. Descriptive Analytics Analysis

Determining a procedure by which analyst teams can determine optimal shipments between supply sources and demand destinations requires differing types of data. There is supply, demand, and cost data required to plan shipments. The total manufactured supply of motors produced at the St. Louis and Dallas plants is determined once the forecast demand is established. The BA team established that there is ample capacity between both plants to satisfy the forecasted customer demand at the six warehouse demand destinations.

The BA team determined that the cost data for shipping a motor from the production centers to the customers depends largely on distance between the cities, where the items are trucked directly by the manufacturer to the warehouses. The cost data per motor shipped to a customer is given in **Table 8.1**. For example, it costs the manufacturer $4 per motor to ship from St. Louis to Kansas City. These cost values are routinely computed by the manufacturer's cost accounting department and are assumed by the BA team to be accurate.

Supply Center	Kansas City, Missouri	Chicago, Illinois	Houston, Texas	Oklahoma City, Oklahoma	Omaha, Nebraska	Little Rock, Arkansas
St. Louis, Missouri	$4	$6	$9	$8	$5	$6
Dallas, Texas	$5	$8	$2	$5	$8	$5

Table 8.1 Estimated Shipping Costs Per Motor

The present system of forecasting customer demand usually results in costly overages and underages shipped to the warehouses. In the past, the manufacturer would take a three value smoothing average to estimate the monthly demand. (See **Section E.6.1** in **Appendix E**, "**Forecasting**.") This evolved by taking the last three months of actual customer motor demand and averaging them to produce a forecast for the next month. The process was repeated each month for each of the six warehouses. Not making products available when customers demanded them caused lost sales, so the manufacturer would rush and ship products to customers at a loss. On the other hand, producing too much inventory meant needless production, inventory, and shipping costs.

To deal with the variability in customer demand forecasting, models for each warehouse's customer demand would need to be developed. The customer demand data on which to build the models was collected from prior monthly demand in motors. To determine which data to include in a final sample and which to exclude, a few simple rules were adopted to eliminate potentially useless and out-of-date data. Going back more than 27 months invited cyclical variations caused by changes in the economy that were no longer present, so that data was removed. Unfortunately, some of the data files were incomplete and required cleansing (see **Chapter 4**). The resulting time series data collected on warehouse customer monthly demand files is presented in **Table 8.2**. It was decided that the most recent three months (darkened months of 25, 26, and 27) would not be included in the model development, but instead would be used for validation purposes to confirm the forecasting accuracy of the resulting models. This is similar to what was referred to as a training data set and a validation data set (see **Section 6.3.1** in **Chapter 6**, "**What Are Predictive Analytics?**").

Month	Kansas City	Chicago	Houston	Oklahoma City	Omaha	Little Rock
1	3120	2130	3945	14020	5045	4610
2	3090	2290	4000	13890	5030	4630
3	3140	2405	4105	13785	5075	4650
4	3010	2580	4300	13575	5015	4680
5	2900	2635	4255	13345	5015	4700
6	2990	2690	4420	12990	5020	4750
7	3000	2740	4540	12340	5025	4800
8	3030	2780	4670	11850	5050	4865
9	3050	2890	4820	11010	5010	4910
10	2970	2940	4780	10015	5010	4980
11	2980	3000	4900	9875	5015	5000
12	2990	3020	5020	9005	5015	5010
13	3020	3120	5045	8880	5010	4950
14	3100	3180	4945	7990	5015	4900
15	2900	3210	4855	7345	5020	4845
16	3000	3270	4780	6920	5020	4800
17	3040	3455	4650	6745	5010	4785
18	3060	3575	4535	6010	5015	4740
19	2950	3765	4475	5670	5040	4700
20	2970	3810	4330	5345	5010	4695
21	2990	3910	4325	5110	5020	4690
22	3060	3990	4155	4760	5010	4680
23	3000	4010	4090	4320	5015	4670
24	3010	4030	4010	4030	5010	4660
25	**2980**	**4285**	**3720**	**3005**	**5010**	**4590**
26	**2965**	**4420**	**3530**	**2515**	**5010**	**4570**
27	**2,945**	**4,560**	**3,330**	**2,030**	**5,005**	**4,555**

Table 8.2 Actual Monthly Customer Demand in Motors

As a part of the descriptive analysis, summary statistics were generated from both Excel (Table 8.3) and SPSS (Table 8.4). The mean values provide some basis for a monthly demand rate, but at this point consideration of overall behavior within data distributions is required to more accurately capture relevant variation. To that end, other statistics can provide some picture of the distribution of the data. For example, the Kurtosis coefficient (see Chapter 5, "What Are Descriptive Analytics?") for Omaha's demand suggests a peaked distribution. This indicates that the variance about the mean is closely grouped toward the mean, implying a lack of variability in forecast values (a good thing). Note that the Standard Error statistic (see Chapter 5, Section 5.3) for Omaha is the smallest. Other statistics such as the Skewedness Coefficients suggest most of the distributions are negatively skewed. The median value peaks at a larger value than the mean and implies that the mean and mean-related statistics might not be as accurate in measuring the entire distribution's behavior as other measures (like the median).

Statistics	Kansas City	Chicago	Houston	Oklahoma City	Omaha	Little Rock
Mean	3010.8696	3186.73913	4521.95652	8904.565	5020.652	4786.522
Standard Error	11.993296	112.691475	69.635371	708.1243	3.241046	25.51144
Median	3000	3120	4535	8880	5015	4750
Mode	2990	#N/A	4780	#N/A	5015	4680
Standard Dev.	57.517827	540.449326	333.959507	3396.045	15.54351	122.3486
Sample Var.	3308.3004	292085.474	111528.953	11533120	241.6008	14969.17
Kurtosis	0.4632145	–1.1215079	–1.2708889	–1.45505	6.482162	–1.04279
Skewedness	0.1407847	0.20384211	–0.0337087	0.142149	2.442156	0.571352
Range	240	1740	1045	9860	65	380
Minimum	2900	2290	4000	4030	5010	4630
Maximum	3140	4030	5045	13890	5075	5010
Sum	69250	73295	104005	204805	115475	110090
Count	23	23	23	23	23	23
Largest(1)	3140	4030	5045	13890	5075	5010
Smallest(1)	2900	2290	4000	4030	5010	4630
Confidence Level(95.0%)	24.872573	233.707814	144.41492	1468.56	6.721519	52.90748

Table 8.3 Excel Summary Statistics of Actual Monthly Customer Demand in Motors

Statistics	Kan City	Chicago	Houston	OK City	Omaha	Little Rock
N Valid	24	24	24	24	24	24
Missing	0	0	0	0	0	0
Mode	2990[a]	2130[a]	4780	4030[a]	5010[a]	4680[a]
Range	240	1900	1100	9990	65	400
Minimum	2900	2130	3945	4030	5010	4610
Maximum	3140	4030	5045	14020	5075	5010

a. Multiple modes exist. The smallest value is shown.

Table 8.4 SPSS Summary Statistics of Actual Monthly Customer Demand in Motors

To better depict the general shape of the data and to understand their behavior, line graphs (see **Chapter 5**, **Section 5.2**) of the six customer demand files are graphed using SPSS in **Figures 8.1** to **8.6**. (The Excel versions look the same and will not be displayed.) As expected based on the summary statistics and now visually from the graphs, some of the customer demand functions look fairly linear, others are clearly nonlinear, and some possess so much variation they are unrecognizable. Identifying the almost perfect linear customer demand behavior in the warehouses in Chicago (**Figure 8.2**) and Oklahoma City (**Figure 8.4**) suggests the use of a simple linear regression model for forecasting purposes. The very clear, bell-shaped, nonlinear functions for Houston (**Figure 8.3**) and Little Rock (**Figure 8.6**) suggest that a nonlinear regression model should be determined by the BA team to find the best-fitting forecasting model. Finally, the excessively random customer demand behavior for Kansas City (**Figure 8.1**) and Omaha (**Figure 8.5**) suggests that considerable effort is needed to find a model that may or may not explain the variation in the data well enough for a reliable forecast. There appear to be many time series variations (see **Appendix E**, **Section E.2**) in customer demand for the warehouses in these two cities.

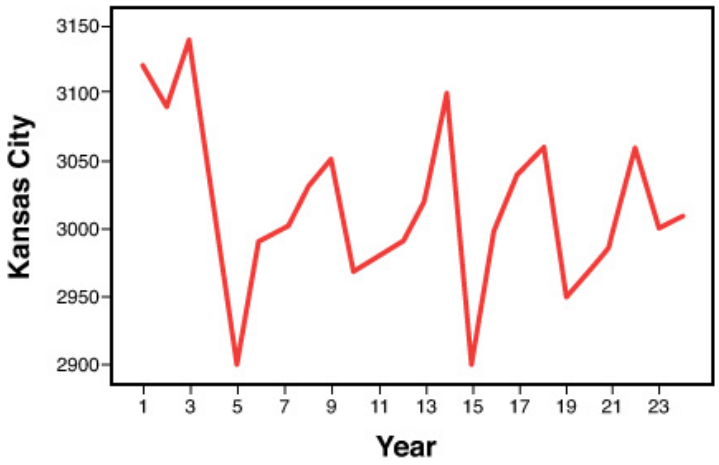


Figure 8.1 Graph of Kansas City customer demand

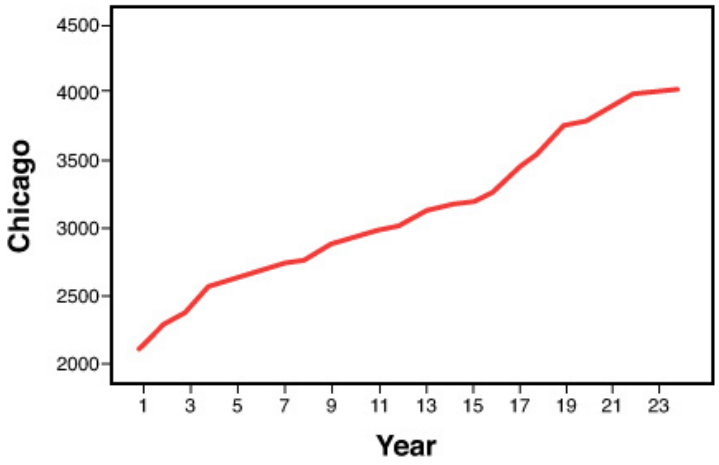


Figure 8.2 Graph of Chicago customer demand

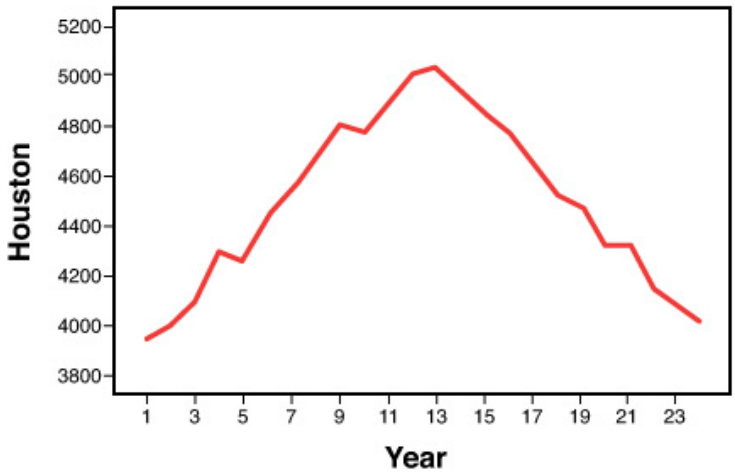


Figure 8.3 Graph of Houston customer demand

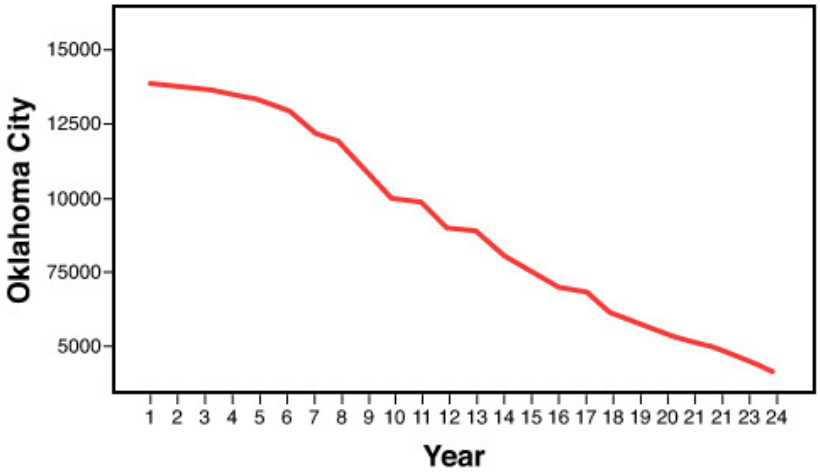


Figure 8.4 Graph of Oklahoma City customer demand

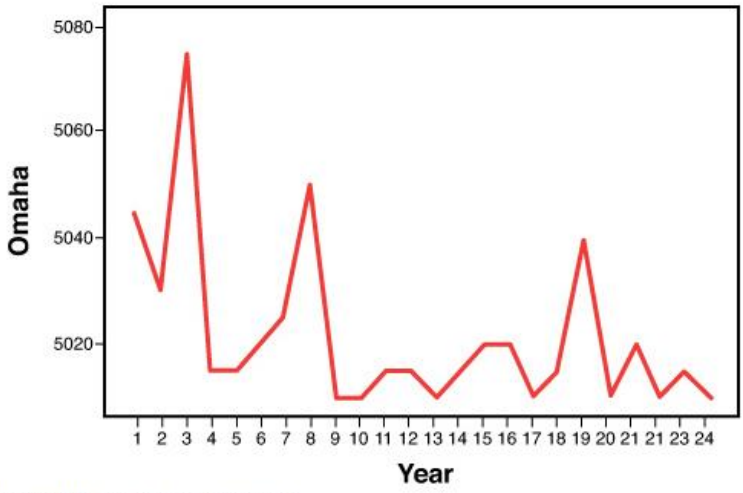


Figure 8.5 Graph of Omaha customer demand

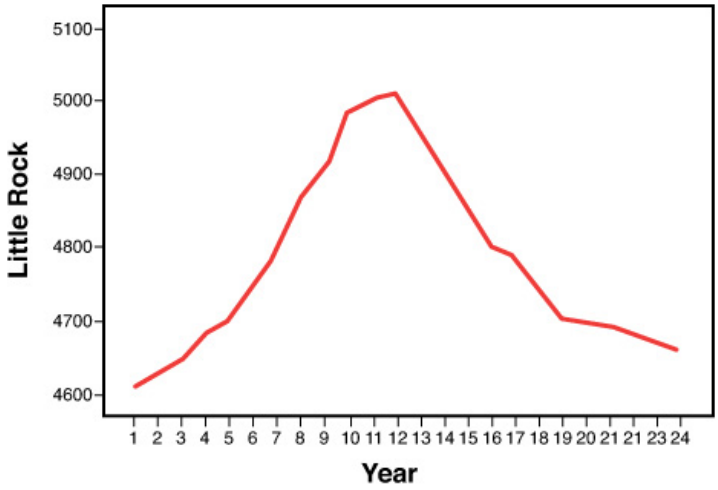


Figure 8.6 Graph of Little Rock customer demand

The fact that two of the four warehouse time series data files have more time series variations than the other four warehouse files does not prevent in this case (and in most others) a fairly accurate forecast. Because four of the six customer demand warehouses appear to have a fairly observable pattern of behavior, they will help improve the overall accuracy even with the substantial variations of the other two warehouses adding in some forecast error.

8.4. Predictive Analytics Analysis

In this section, we continue with our illustrative example. Here we use the predictive analytics analysis step that requires model development effort and then model validation for the example. To complete the predictive analytics analysis, forecasts of warehouse demand are determined.

8.4.1. Developing the Forecasting Models

The descriptive analytics analysis has suggested a course of action in identifying appropriate forecasting models in this next step of the BA process. To ensure the best possible forecasting models and confirm the descriptive analytics analysis results, the curve-fitting feature (Curve Estimation function) of SPSS will be utilized. Each of the six customer demand data files is analyzed through the SPSS program to generate potential regression models, as presented in Tables 8.5 through 8.10.

Dependent Variable: KanCity									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.073	1.727	1	22	.202	3044.275	–2.309		
Logarithmic	.174	4.637	1	22	.043	3084.805	–30.398		
Inverse	.244	7.117	1	22	.014	2992.958	142.744		
Quadratic	.190	2.463	2	21	.109	3095.667	–14.168	.474	
Cubic	**.260**	**2.346**	**3**	**20**	**.103**	**3149.071**	**–37.459**	**2.757**	**–.061**
Compound	.070	1.658	1	22	.211	3043.255	.999		
Power	.169	4.480	1	22	.046	3083.946	–.010		
S	.239	6.905	1	22	.015	8.004	.047		
Growth	.070	1.658	1	22	.211	8.021	–.001		
Exponential	.070	1.658	1	22	.211	3043.255	–.001		
Logistic	.070	1.658	1	22	.211	.000	1.001		

The independent variable is Year.

Table 8.5 SPSS Curve-Fitting Analysis for Kansas City Motor Demand Forecasting Model: Model

Summary and Parameter Estimates

Dependent Variable: Chicago									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.982	1230.679	1	22	.000	2142.409	80.024		
Logarithmic	.844	119.094	1	22	.000	1701.041	631.563		
Inverse	.480	20.270	1	22	.000	3439.516	–1886.509		
Quadratic	.984	647.977	2	21	.000	2199.355	66.883	.526	
Cubic	**.986**	**467.241**	**3**	**20**	**.000**	**2116.996**	**102.802**	**–2.994**	**.094**
Compound	.979	1046.077	1	22	.000	2239.093	1.026		
Power	.903	205.978	1	22	.000	1909.324	.211		
S	.561	28.098	1	22	.000	8.140	–.660		
Growth	.979	1046.077	1	22	.000	7.714	.026		
Exponential	.979	1046.077	1	22	.000	2239.093	.026		
Logistic	.979	1046.077	1	22	.000	.000	.974		

The independent variable is Year.

Table 8.6 SPSS Curve-Fitting Analysis for Chicago Motor Demand Forecasting Model: Model Summary and Parameter Estimates

Dependent Variable: Houston									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.003	.076	1	22	.786	4461.938	2.878		
Logarithmic	.127	3.191	1	22	.088	4158.249	148.801		
Inverse	.239	6.892	1	22	.015	4625.232	–809.218		
Quadratic	.929	138.408	2	21	.000	3632.994	194.173	–7.652	
Cubic	**.930**	**88.001**	**3**	**20**	**.000**	**3621.937**	**198.995**	**–8.124**	**.013**
Compound	.004	.088	1	22	.769	4446.289	1.001		
Power	.133	3.379	1	22	.080	4149.443	.034		
S	.252	7.415	1	22	.012	8.438	–.186		
Growth	.004	.088	1	22	.769	8.400	.001		
Exponential	.004	.088	1	22	.769	4446.289	.001		
Logistic	.004	.088	1	22	.769	.000	.999		

The independent variable is Year.

Table 8.7 SPSS Curve-Fitting Analysis for Houston Motor Demand Forecasting Model: Model Summary and Parameter Estimates

Dependent Variable: OKCity									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.986	1567.704	1	22	.000	15229.746	–488.963		
Logarithmic	.826	104.691	1	22	.000	17817.142	–3811.033		
Inverse	.411	15.382	1	22	.001	7440.961	10657.408		
Quadratic	.987	775.938	2	21	.000	15420.511	–532.986	1.761	
Cubic	**.996**	**1697.830**	**3**	**20**	**.000**	**14294.516**	**–41.911**	**–46.359**	**1.283**
Compound	.978	990.240	1	22	.000	17380.621	.944		
Power	.747	64.854	1	22	.000	22518.882	–.430		
S	.338	11.256	1	22	.003	8.860	1.147		
Growth	.978	990.240	1	22	.000	9.763	–.058		
Exponential	.978	990.240	1	22	.000	17380.621	–.058		
Logistic	.978	990.240	1	22	.000	5.754E-005	1.060		

The independent variable is Year.

Table 8.8 SPSS Curve-Fitting Analysis for Oklahoma City Motor Demand Forecasting Model: Model

Summary and Parameter Estimates

Dependent Variable: Omaha									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.220	6.205	1	22	.021	5034.928	–1.061		
Logarithmic	.300	9.424	1	22	.006	5045.741	–10.546		
Inverse	.253	7.435	1	22	.012	5015.632	38.358		
Quadratic	.299	4.477	2	21	.024	5046.077	–3.634	.103	
Cubic	**.340**	**3.432**	**3**	**20**	**.037**	**5056.845**	**–8.330**	**.563**	**–.012**
Compound	.220	6.212	1	22	.021	5034.886	1.000		
Power	.300	9.439	1	22	.006	5045.714	–.002		
S	.253	7.454	1	22	.012	8.520	.008		
Growth	.220	6.212	1	22	.021	8.524	.000		
Exponential	.220	6.212	1	22	.021	5034.886	.000		
Logistic	.220	6.212	1	22	.021	.000	1.000		

The independent variable is Year.

Table 8.9 SPSS Curve-Fitting Analysis for Omaha Motor Demand Forecasting Model: Model Summary and Parameter Estimates

Dependent Variable: LittleRock									
	Model Summary					Parameter Estimates			
Equation	R-Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.001	.019	1	22	.893	4785.580	–.513		
Logarithmic	.068	1.602	1	22	.219	4689.689	39.198		
Inverse	.167	4.399	1	22	.048	4817.464	–243.419		
Quadratic	.754	32.130	2	21	.000	4516.566	61.567	–2.483	
Cubic	**.801**	**26.824**	**3**	**20**	**.000**	**4426.153**	**100.999**	**–6.347**	**.103**
Compound	.001	.015	1	22	.903	4783.397	1.000		
Power	.070	1.659	1	22	.211	4688.108	.008		
S	.171	4.539	1	22	.045	8.480	–.051		
Growth	.001	.015	1	22	.903	8.473	–9.677E-005		
Exponential	.001	.015	1	22	.903	4783.397	–9.677E-005		
Logistic	.001	.015	1	22	.903	.000	1.000		

The independent variable is Year.

Table 8.10 SPSS Curve-Fitting Analysis for Little Rock Motor Demand Forecasting Model: Model Summary and Parameter Estimates

Reviewing the R-Square values for each of the potential curve-fitting models, it turns out that the cubic model is the best fitting for all six data files. It is not surprising that in the cases of Houston and Little Rock, where the descriptive analytics graphs clearly show typical cubic (or quadratic) function behavior, that the only significant (F-ratio, p<.000) models were cubic or quadratic (see **Chapter 6**, **Section 6.4.2**). In other cases (Chicago and Oklahoma City), it is surprising that a nonlinear cubic model does a slightly better job than the descriptive analytics step linear model. On the other hand, note that for both locations, the linear model, according to the R-Square statistics, is either the next best choice or the next to the next best choice. Indeed, in both cases, the F-ratio clearly shows that the resulting linear model can provide a statistically significant forecasting capability. Other models also have significant (p<.000) F-ratios, suggesting the possibility of accurate forecasting. Because the objective of this case study is to develop a procedure that analysts could use each month to develop an optimal supply chain schedule of shipments and to accurately forecast customer demand, the BA analyst can use the highest R-Square statistic as a means to determine the most accurate forecasting model from those fitted with the data.

In this case study the resulting cubic regression models estimated by the SPSS program based on the parameters from the curve-fitting effort are presented in Table 8.11.

Location	Constant	b1	b2	b3	R-Square
Kansas City	3149.071	–37.459	2.757	–.061	.260
Chicago	2116.996	102.802	–2.994	.094	.986
Houston	3621.937	198.995	–8.124	.013	.930
Oklahoma City	14294.516	–41.911	–46.359	1.283	.996
Omaha	5056.845	–8.330	.563	–.012	.340
Little Rock	4426.153	100.999	–6.347	.103	.801

Table 8.11 Resulting Cubic Forecasting Models from SPSS Curve-Fitting Analysis

The generalized formula for a cubic regression model follows:

$$Y_p = a + b_1 X + b_2 X^2 + b_3 X^3$$

By inserting the curve-fitted parameters for Little Rock, the resulting cubic regression model for forecasting warehouse customer demand is this:

$$Y_p = 4426.153 + 100.999 X - 6.347 X^2 + 0.801 X^3$$

where:

X = month number in the form of the time series data file (26, 27, and 28)

8.4.2. Validating the Forecasting Models

One of the fundamental requirements of a BA analysis is to show or prove the possibility of improving business performance (see Chapter 1, Section 1.1). One criterion for improving forecasting is to improve forecasting accuracy. To compare the current forecasting method with the newly devised one, each cubic model is used to forecast the respective location of customer demand. Substituting the numbered time values (25, 26, and 27) for X in each cubic model, the analyst is able to compute the three forecast values. These forecasts are then compared with the actual values in Table 8.1. The resulting comparison is expressed in the MAD statistics (see Appendix E, Section E.8), as presented in Table 8.12.

Month	Kansas City	Chicago	Houston	Oklahoma City	Omaha	Little Rock
25	2983	4285	3723	4320	5013	4594
26	2967	4419	3533	4417	5010	4572
27	2947	4561	3329	4621	5007	4554
MAD	2.33	0.66	2.33	1936	1.66	2.33

Table 8.12 Resulting Cubic Model Forecasts and MAD Statistics (Rounded Up to Next Integer Value)

The MAD statistics for all warehouse facilities except Oklahoma City are extremely small, suggesting the cubic models for these locations are very accurate. The Oklahoma City MAD statistic, on the other hand, is so great relative to the other MADs that it suggests further analysis is needed to find a better forecasting model for Oklahoma City.

To explore this forecast exception in Oklahoma City, the next two best models (based on R-Square) from the SPSS curve-fitting effort in **Table 8.8** are examined. These two include the linear regression model (R-Square 0.986):

$$Y_p = 15229.746 - 488.963X$$

and the quadratic regression model (R-Square 0.987):

$$Y_p = 15420.511 - 532.986\ X + 1.761\ X^2$$

In **Table 8.13**, the resulting warehouse forecasts of customer demand for each of the three years are presented along with their MAD statistics for both the linear and the quadratic models. Clearly, the linear regression model's small MAD suggests that it is the better model for forecasting than either the quadratic or the cubic models. This result is not surprising, given the prior descriptive analytics step, which appeared to suggest that a linear model would be the best type of forecasting model.

Year (X)			
Model	**25**	**26**	**27**
Linear Forecast	3006	2517	2028
	MAD = 2.33		
Quadratic Forecast	3197	2754	2314
	MAD = 238.33		

Table 8.13 Resulting Linear and Quadratic Forecasts with MAD Statistics for Oklahoma City (Rounded Up to the Next Integer Value)

Having found the models that provide low error rates, it is now necessary to validate them by demonstrating they can improve forecasting accuracy and, therefore, enhance business performance by minimizing costly shipping efforts.

To validate the forecasting accuracy and demonstrate forecasting improvement of the cubic and linear models, a comparison with the currently used smoothing average method is undertaken by the analyst. Utilizing a similar smoothing average formula to that mentioned in **Section E.6** in **Appendix E**, the forecast values for warehouse customer demand can be computed using the simple formula below:

$$\overline{Y}_t = (Y_{t-1} + Y_{t-2} + Y_{t-3})/3$$

where:

$\overline{Y}_t$ = the forecast value in time period *t*

Y_{t-1} = the actual value in the time period just prior to time period *t*

Y_{t-2} = the actual value of two time periods prior to time period *t*

Y_{t-3} = the actual value of three time periods prior to time period *t*

Using the formula, the resulting smooth average forecast values are presented in Table 8.14 along with their respective MAD statistics.

Month	Kansas City	Chicago	Houston	Oklahoma City	Omaha	Little Rock
25	3024	4010	4085	4370	5012	4670
26	2997	4109	3940	3785	5012	4640
27	2985	4245	3754	3184	5010	4607
MAD	21.33	111.33	158.66	590.33	1.66	31

Table 8.14 Resulting Smooth Average Forecasts and MAD Statistics (Rounded Up to Next Integer Value)

These smoothing average forecasts and their MADs can be compared with the forecasts and MADs for the cubic and linear models. Comparing the MADs in Table 8.14 with the MADs in Tables 8.12 and 8.13, several points about forecasting improvement can be made. In the case of the locations for Kansas City, Chicago, Houston, and Little Rock, the cubic regression models are the lowest; therefore, they have more accurate forecasting results. For those four locations, the cubic model is recommended. In the case of Oklahoma City, the linear regression model results in the lowest MAD value, reflecting improved forecasting accuracy over the other models. Finally, the MADs for both the cubic regression (Table 8.12) and the smoothing average methods (Table 8.14) result in the same MAD value (1.66) for Omaha, which suggests either method is accurate in forecasting this location's customer demand. Because either method can be used, the manufacturer's BA analyst selected to employ the cubic regression model for forecasting Omaha's warehouse customer demand.

8.4.3. Resulting Warehouse Customer Demand Forecasts

The selected forecasting models and their forecast values for the future 28th month (X = 28 in the models) are presented in Table 8.15. The direction of movement from the 27th to the forecast of the 28th month appears mainly to be downward for most of the warehouse locations. The resulting forecast values for the six locations are generally consistent with graphs from the descriptive analytics analysis, although some of the time series variation behavior (for example, Kansas City) can hardly be predicted up or down in movement.

Location	Forecast Model	Forecast (X = 28)	Direction
Kansas City	$Y_p = 3149.071 - 37.459X + 2.757\,X^2 - .061\,X^3$	2923	Down
Chicago	$Y_p = 2116.996 + 102.802X - 2.994\,X^2 + .094\,X^3$	4712	Up
Houston	$Y_p = 3621.937 + 198.995X - 8.124\,X^2 + .013\,X^3$	3110	Down
Oklahoma City	$Y_p = 15229.746 - 488.963X$	1539	Down
Omaha	$Y_p = 5056.845 - 8.330X + .563\,X^2 - .012\,X^3$	5002	Down
Little Rock	$Y_p = 4426.153 + 100.999X - 6.347\,X^2 + .103\,X^3$	4540	Down
Total Forecast Customer Demand for the 28th Month		21826	

Table 8.15 Resulting Forecasts for the Future 28th Month (Rounded Up to Next Integer Value)

8.5. Prescriptive Analytics Analysis

Based on the predictive analytics analysis, the total forecast demand (see Table 8.15) of 21,826 motors for

all six warehouse locations has to be balanced out by product capacity of the two production facilities. The BA team decided that the St. Louis production center would produce 10,000 motors for the 28th month, and the Dallas production center would produce the remaining 11,826 motors.

8.5.1. Selecting and Developing an Optimization Shipping Model

In terms of data, the analyst now possesses the supply, forecast demand, and cost information on which to begin selecting a modeling approach to achieve an optimal shipping schedule. Reviewing the requirements of the problem setting at this point in the analysis, the BA team is looking at a multivariable (number of motors to ship from two supply sources to three demand destinations), multidimensional (scheduling motor shipments from two supply sources to six demand markets or supply and demand), constrained (the exact number of motors required is deterministic at this point), integer (shipping whole motors, not motor parts), and optimal solution (seeking to minimize cost of shipping). The ideal BA methodology to satisfy these requirements is *integer programming* (IP). (See **Appendix D**, "**Integer Programming**.")

To conceptualize a two-dimensional problem, a *transportation method* (an operations research methodology) table that combines location cost per unit shipped and supply and demand information is presented. The table for this case study problem is shown in **Table 8.16**. The decision variables for the model are also added. So, for example, X_{11} represents the number of manufacturer motors to ship from the St. Louis production center to meet the forecast customer demand at the Kansas City warehouse. In this table, the sum of the motors produced in St. Louis and shipped to any of the six customer demand locations must add up to 10,000. Likewise, for Dallas, the shipments must equal 11,826. Also, for each column, the sum of the motors shipped must equal the forecast demand in that column. For example, the sum of X_{11} and X_{21} for the Kansas City warehouse must equal the forecast demand of 2,923 motors.

	Kansas City, Missouri	Chicago, Illinois	Houston, Texas	Oklahoma City, Oklahoma	Omaha, Nebraska	Little Rock, Arkansas	Production Center Supply
St. Louis, Missouri	$4\,X_{11}$	$6\,X_{12}$	$9\,X_{13}$	$8\,X_{14}$	$5\,X_{15}$	$6\,X_{16}$	10000
Dallas, Texas	$5\,X_{21}$	$8\,X_{22}$	$2\,X_{23}$	$5\,X_{24}$	$8\,X_{25}$	$5\,X_{26}$	11826
Forecast of Customer Demand	2923	4712	3110	1539	5002	4540	21826

Table 8.16 Transportation Method Table for Conceptualization of Supply Chain Shipping Problem

With the transportation method table as a framework, the IP model can be developed. In this type of shipping problem, there are two supply-side constraints and six demand-side constraints required to ensure the supply is allocated to meet the demand. The same formulation procedure for LP models in **Appendix B**, "**Linear Programming**," and for integer programming in **Appendix D** is applied here to generate the following integer model:

$$\text{Minimize: } Z = 4X_{11} + 6X_{12} + 9X_{13} + 8X_{14} + 5X_{15} + 6X_{16} + 5X_{21} + 8X_{22} + 2X_{23} + 5X_{24} + 8X_{25} + 5X_{26}$$

subject to:

$$X_{11} + X_{12} + X_{13} + X_{14} + X_{15} + X_{16} = 10000 \text{ (St. Louis supply requirement)}$$
$$X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{26} = 11826 \text{ (Dallas supply requirement)}$$
$$X_{11} + X_{21} = 2923 \text{ (Kansas City demand requirement)}$$
$$X_{12} + X_{22} = 4712 \text{ (Chicago demand requirement)}$$
$$X_{13} + X_{23} = 3110 \text{ (Houston demand requirement)}$$
$$X_{14} + X_{24} = 1539 \text{ (Oklahoma City demand requirement)}$$
$$X_{15} + X_{25} = 5002 \text{ (Omaha demand requirement)}$$
$$X_{16} + X_{26} = 4540 \text{ (Little Rock demand requirement)}$$
$$\text{and } X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{21}, X_{22}, X_{23}, X_{24}, X_{25}, X_{26} \geq 0 \text{ and all integer}$$

8.5.2. Determining the Optimal Shipping Schedule

To run this model, LINDO (see **Appendix B**, **Section B.5.3**) software is utilized. As it turns out, in this situation the unique formulation of the transportation method model mathematically forces an all-integer solution without the need for using the IP software algorithm. This permits the regular LP software to be used to solve this problem, although LINDO has both IP and LP solution software. The LINDO LP model input is presented in **Figure 8.7**, and the results are presented in **Figure 8.8**.

LINGO Model - LINGO1

```
Min = 4 * x11 + 6 * x12 + 9 * x13 + 8 * x14 + 5 * x15 + 6 * x16 + 5 * x21 + 8 * x22 + 2 * x23 + 5 * x24 + 8 * x25 + 5 * x26;
x11 + x12 + x13 + x14 + x15 + x16 = 10000;
x21 + x22 + x23 + x24 + x25 + x26 = 11826;
x11 + x21 = 2923;
x12 + x22 = 4712;
x13 + x23 = 3110;
x14 + x24 = 1539;
x15 + x25 = 5002;
x16 + x26 = 4540;
```

Figure 8.7 LINDO input for supply chain shipping model problem

Solution Report - LINGO1

```
  Global optimal solution found.
  Objective value:                              104226.0
  Total solver iterations:                             1


                       Variable           Value        Reduced Cost
                            X11        286.0000            0.000000
                            X12        4712.000            0.000000
                            X13        0.000000            8.000000
                            X14        0.000000            4.000000
                            X15        5002.000            0.000000
                            X16        0.000000            2.000000
                            X21        2637.000            0.000000
                            X22        0.000000            1.000000
                            X23        3110.000            0.000000
                            X24        1539.000            0.000000
                            X25        0.000000            2.000000
                            X26        4540.000            0.000000

                            Row    Slack or Surplus      Dual Price
                              1        104226.0           -1.000000
                              2        0.000000            0.000000
                              3        0.000000           -1.000000
                              4        0.000000           -4.000000
                              5        0.000000           -6.000000
                              6        0.000000           -1.000000
                              7        0.000000           -4.000000
                              8        0.000000           -5.000000
                              9        0.000000           -4.000000
```

Figure 8.8 LINDO output for supply chain shipping model problem

Extracting the shipping schedule for the supply chain problem, the number of motors to be shipped from the two supply source locations to the six demand destinations is presented in **Table 8.17** (the bold numbers in the rows and columns in the table). For example, to achieve a cost-minimized shipping schedule, the manufacturer has to ship 286 motors from St. Louis to the Kansas City warehouse in the 28th month.

Likewise, all the other eight scheduled shipments in Table 8.17 must be shipped exactly as scheduled to ensure the optimization of the total costs. Note in Table 8.17 that the allocation of motors exactly adds up to the last column supply values and the bottom row demand values.

	Kansas City, Missouri	Chicago, Illinois	Houston, Texas	Oklahoma City, Oklahoma	Omaha, Nebraska	Little Rock, Arkansas	Production Center Supply
St. Louis, Missouri	**286**	**4712**			**5002**		10,000
Dallas, Texas	**2637**		**3110**	**1539**		**4540**	11,826
Forecast of Customer Demand	2923	4712	3110	1539	5002	4540	21,826

Table 8.17 Shipping Schedule for 28th Month Supply Chain Shipping Problem

Also, the value of 104,266 in Figure 8.2 is the total optimized cost for this shipping schedule (taking the units shipped in each cell of Table 8.17 and multiplying them by the number of units in those cells). The resulting shipping schedule for the supply chain problem in month 28 is detailed in Table 8.18.

Ship From	To	Units × Per Unit Cost = Total Cost
St. Louis	Kansas City	286 × $4 = $1,144
St. Louis	Chicago	4,712 × $6 = $28,272
St. Louis	Omaha	5,002 × $5 = $25,010
Dallas	Kansas City	2,637 × $5 = $13,185
Dallas	Houston	3,110 × $2 = $6,220
Dallas	Oklahoma City	1,539 × $5 = $7,695
Dallas	Little Rock	4,540 × $5 = $22,700
Total Supply Chain Shipping Cost = $104,226		

Table 8.18 Resulting Shipping Schedule for Month 28

8.5.3. Summary of BA Procedure for the Manufacturer

The intent of this BA application is to develop a BA procedure for the manufacturer to utilize every month in planning the supply chain problem of setting up an optimal shipping schedule in the supply chain network. This BA procedure based on the BA team analysis presented here involves both data collection efforts: statistical analysis and the application of optimization software. Specifically:

1. Collect shipping cost information from the firm's cost accounting department.

2. Collect and update monthly actual demand values from warehouse customers.

3. Collect supply center supply capacity to ensure sufficient supply capacity exists to handle monthly demand.

4. Rerun curve-fitting software on new and old actual demand data to determine best forecasting model based on R-Square and other statistics as needed.

5. Forecast warehouse customer demand and affirm through analysis that the resulting estimates are truly based on the best forecasting model. Revise as needed to select the best model.

6. Incorporate the cost, supply, and forecast demand information into a linear programming model similar to what was developed in **Section 8.5.1**.

7. Run the IP or LP model and extract the shipping schedule from the model output.

8.5.4. Demonstrating Business Performance Improvement

A BA is not complete without showing that business performance can or will be improved. This case study has a basis for comparison. In comparing the MAD statistics from the present forecasting procedure and the BA proposed procedure, a potential for improvement in shipping can be observed. In **Table 8.19**, the MAD values based on the three months (months 25, 26, and 27) used for model validation are presented. The MAD statistics (see **Appendix E**, **Section E.8**) represent the average monthly overage or underage of motors that could have been avoided if the BA proposed procedure would have been in place. Such needless shipments waste effort and add to cost inefficiencies for the manufacturer. Establishing a procedure that lowers the MAD statistics would represent an opportunity for improving business performance.

MADs	Kansas City, Missouri	Chicago, Illinois	Houston, Texas	Oklahoma City, Oklahoma	Omaha, Nebraska	Little Rock, Arkansas	Total
MADs for present smoothing model (Table 8.14)	21.33	111.33	158.66	590.33	1.66	31	914.31
MADs for proposed cubic and linear* models (Tables 8.12 and 8.13)	2.33	0.66	2.33	2.33*	1.66	2.33	11.64

Table 8.19 Comparisons of MAD Statistics Between Present and BA Proposed Forecasting Procedures

As can be seen in **Table 8.19**, the current procedure of using a smoothed averaging to generate a forecast results in fairly large MAD statistics compared to the proposed BA procedure. The total of the MADs in **Table 8.19** clearly shows a significant reduction in monthly overages or underages when using the proposed BA procedure in forecasting motor customer demand. Reducing the inaccuracy in forecasting also translates into minimizing wasted costs of shipping the motors that are either not needed in the warehouses during low customer demand periods or rush-ordered when shortages occur. These results reveal that the implementation of the proposed BA procedure for the supply chain shipping schedule problem could have improved business performance for the manufacturer over what was previously used to forecast the last three months.

As a final recommendation from the BA team on the prescriptive analytics step, the analyst or BA team responsible for utilizing the new BA procedure should continuously run updates to check and confirm the benefits of using the BA procedure on a monthly basis. Continuously showing the worth of BA is recommended for the success of BA in firms (see **Table 2.2** in **Chapter 2**, "**Why Are Business Analytics Important?**").

Summary

This chapter presented a case study illustrating the use of BA to solve a supply chain shipping problem. The case study utilized the three-step BA process to develop a BA procedure that could be repeated monthly to improve a manufacturer's business performance.

The particular use of methodologies in this case study could have been different and could have highlighted the fact that BA is meant as a step-wise guide in the application of statistical, information system, and management science methodologies. Like a walk in a forest, there may be many paths, but the goal is to reach the other side using knowledge and information (from a BA analysis) to support your steps.

This chapter ends the text material of this book, but the appendixes offer readers a rich foundation of methodologies useful in BA. Some have been demonstrated in the text material, and others have not, but all can be useful for differing analyses. The more methodologies that BA analysts know, the more likely they are to utilize the right one in the right situation. The appendixes are a starting point on which to build a foundation of methodological tools to strengthen and continually augment BA knowledge.

Discussion Questions

1. Some of the graphs (for example, Chicago going up) in the descriptive analytics analysis tended to show fairly linear behavior, yet a cubic model, rather than a linear model, was used in most cases. Did it make sense to use the cubic model instead of the linear model? Why?

2. The SPSS curve-fitting process involves the development of 11 different regression models. Should all the regression models have been tested with the validation data in months 25, 26, and 27?

3. What other methodologies in this book might have been applied to analyze this case study problem?

4. Why is it necessary to show and continuously demonstrate business improvement in the analysis when it is clear that forecasting results are improved?

Problems

1. How was the Kansas City cubic regression forecast of 2,983 for month 25 computed? Show the formula with input values.

2. How was the Chicago cubic regression forecast of 4,561 for month 27 computed? Show the formula with input values.

3. How was the Oklahoma City quadratic regression forecast of 2,754 for month 26 computed? Show the formula with input values.

4. How was the MAD statistic for the Oklahoma City linear regression model computed on the three-month validation data? Show the formula with input values.

PART IV: APPENDIXES

A. Statistical Tools

A.1. Introduction

The purpose of this appendix is to provide a brief overview of some basic statistical tools, including counting, probability concepts, probability distributions, and statistical testing. Other statistical methods and testing content will be presented in chapters and other appendixes.

A.2. Counting

Counting is an important prerequisite for computing probabilities. A probability example is usually made up of the ratio of a few observed behaviors over the total of all possible behaviors:

$$\text{Probability Estimate} = \frac{\text{A Few Observed Behaviors (Numerator)}}{\text{Total of All Possible Behaviors (Denominator)}}$$

To use this probability formula, we must determine both the numerator and the denominator of the probability estimate ratio. The three methods of counting that can be used to determine probability values include permutations, combinations, and repetitions.

A.2.1. Permutations

A *permutation* is a specific ordering of elements from a collection of elements. One usually wants to determine or count the total permutations that are possible from a given collection of elements. A collection of elements, for example, might be the items that make up a data set. *Permutations* can be defined as the number of ways r elements at a time are taken from a collection of n elements, such that: (1) the ordering of the elements is important; (2) there can be no repetitions of the same element in a set of r; (3) the value of r will either equal n or will be less than n. To determine the total number of permutations possible from a collection of n elements taken r at a time, note the following permutation formula:

$${}_{n}P_{r} = \frac{n!}{(n-r)!}$$

where:

! = a factorial sign. (A *factorial* is a mathematical abbreviation indicating that the value it is attached to should be multiplied by all the values that precede it. For example, 4! is an abbreviation for $4 \times 3 \times 2 \times 1$, or 24.)

P = the possible number of permutations of elements out of a collection of elements

n = the total number of elements in a collection

r = the number of elements taken from the collection at one time

In using this formula, it is important to remember that the three characteristics included in the permutation formula (order is important, objects should not be repeated, and $r \leq n$) are strictly observed. These characteristics of the permutation formula will change for each of the other counting formulas discussed.

Question: What are the permutations of five letters A, B, C, D, and E taken two at a time?

Answer: You can use the permutations formula to calculate this:

$$_{n}P_{r} = \frac{n!}{(n-r)!} = \frac{5!}{(5-2)!} = \frac{5\times4\times3\times2\times1}{3!} = \frac{120}{6} = 20 \text{ permutations}$$

The permutations can be listed using this systematic procedure. Begin with the listing of the permutations with the first letters first. These permutations are AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE. Then reverse the letters to obtain the remaining permutations, which are BA, CA, DA, EA, CB, DB, EB, DC, EC, and ED.

A.2.2. Combinations

A combination, like a permutation, is a specific ordering of elements from a collection of elements. *Combinations* can be defined as the number of ways *r* elements at a time are taken from a collection of *n* elements, such that (1) the ordering of the elements is not important, (2) there can be no repetitions of the same element in a set of *r*, and (3) the value of *r* will either equal *n* or will be less than *n*. To determine the total number of combinations that are possible from a collection of *n* elements, taken *r* at a time, note the following formula:

$$_{n}C_{r} = \frac{n!}{r!(n-r)!}$$

where:

C = the possible number of combinations of elements out of a collection of elements

n = the total number of elements in a collection

r = the number of elements taken from the collection at one time

In using this formula, it is important to remember that the characteristic that differentiates combinations from permutations is that the order of elements is not important. The effect of ordering as not being important will mean that the possible number of combinations will always be less than the possible number of permutations.

Question: What are the combinations of five letters A, B, C, D, and E taken two at a time?

Answer: Using the permutations formula, we find the following to be true:

$$_{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{2!(5-2)!} = \frac{5\times4\times3\times2\times1}{2\times1(3\times2\times1)} = \frac{120}{12} = 10 \text{ permutations}$$

The combinations can be listed using the same systematic procedure previously discussed. List the combinations with the first letters first. The ten combinations for this problem are AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE.

A.2.3. Repetitions

When repetitions of the same element in a collection of elements are possible, a different type of counting formula is necessary. A *repetition*, like a permutation or a combination, is a specific ordering of elements from a collection of elements. *Repetitions* can be defined as the number of ways *r* elements at a time are taken from a collection of *n* elements such that: (1) the ordering of the elements is important; (2) there can be as many as *r* repetitions of the same element in a set of *r*; and (3) the value of *r* will either equal *n* or will be less than *n*. To determine the total number of repetitions possible from a collection of *n* elements, taken *r* at a time, we have the following formula:

$_{n}R_{r} = n^{r}$

where:

R = the possible number of repetitions of *r* elements out of a collection of *n* elements

n = the total number of elements in a collection

r = the number of elements taken from the collection at one time

In using this formula, it is important to remember that the characteristic that differentiates repetitions from the other two counting procedures is that repetitions of the same elements are permitted.

Question: How many different repetitions of five food products (hamburgers, French fries, hot dogs, regular drink, and large drink) can be structured into a "food-deal" package if it takes three food products to make a package of goods? Assume that the same product can be used repeatedly in a single package. (For example, three hamburgers would be considered a "food-deal" package.)

Answer: To find the number of repetitions, use this formula:

$_{n}R_{r} = n_{r} = 5_{3} = 125$ different "food-deal" packages

So a total of 125 different "food-deal" packages can be structured to sell to customers from only five different products taking three products at a time.

A.3. Probability Concepts

Understanding probability usage in business analytics requires some fundamental knowledge of proba-

bility concepts. In this section we introduce approaches used to assess probabilities and then follow the discussion up with basic rules of addition and multiplication to help understand their manipulation.

A.3.1. Approaches to Probability Assessment

Before reviewing the basic rules of probability, be aware of some of the approaches used to assess probabilities. There are two general approaches to assessing probabilities: objective and subjective. The *objective approach to probability* is based on objective methods of collecting experimental data on trials and their outcomes, tabulating the data into frequency distributions and then into probabilities. The two theories that most commonly represent the objective approach to probability are Frequency Theory and the Principle of Insufficient Reason.

Frequency Theory basically states that, through a large number of trials, the relative frequency outcome of an event, A, can be used to determine the probability of A, represented here as P(A). This probability is based on experiential observations, and it is assumed that the experiment accurately represents the possible behaviors that are being observed. The probabilities that are determined using this approach are based on past observations that are converted into probabilities to be used in the future. Because the probabilities are based on past events, we sometimes refer to these probabilities as being *posteriori probabilities*.

Question: Using Frequency Theory, what is the probability that a food server will be tipped an amount that will fall in the tip class interval of $8 to $9.99 based on the following collection of data from a food server's experiences at the restaurant?

Number of Customers

Tips	Who Tipped
$6 to $7.99	50
$8 to $9.99	120
$10 to $12.99	30
Total	200

Answer: The relative frequency of the class interval of $8–$9.99 is 120 out of 200. If we assume the 200 tip experiments recorded are sufficient to accurately describe all tip behavior of the customers for the specific server, Frequency Theory holds that the probability of a tip between $8 and $9.99 being given to that server is 0.60, or 60 percent (120 / 200 = 0.60).

The *Principle of Insufficient Reason* states that each possible outcome in an experiment is equally probable if there is no evidence to challenge the assumption. Alternatively, if there is no reason to prefer one outcome over another, each outcome is equally likely or has the same probability. To determine probabilities using this principle, we must logically abstract the decimal value probabilities by dividing the frequency of occurrence by the total number of possible outcomes. Because there is no evidence to support that any of the possible outcomes is any more probable to occur than the rest, each will be given an equal probability.

Question: A stock broker must select one stock to invest in out of a sample of four stocks. The stock broker is unfamiliar with the four sample stocks. What is the probability that the stock broker will pick the best stock out of the four?

Answer: Assuming that there is only one "best" stock in the sample of four, and because there is one chance out of four to obtain it, the probability of choosing the best one out of four is 0.25 (1 / 4 = 0.25). In this problem, each of the four is equally likely to be the best stock, because no additional information on the stocks is available.

The probabilities that are determined using this principle are based on abstract reasoning of equality of outcomes and not on experience with observing outcomes of trials. The fact that the probabilities are based on prior logic and reasoning leads one to refer to these probabilities as *priori probabilities*.

In the *subjective approach to probability,* estimation of probabilities is based on personal opinion or judgment. Under this approach, we assume that an individual's subjective judgment may be as accurate as or better than any other objective approach. BA analysts, whose subjective expertise is combined with the objective frequency information provided by a database, can combine these sources to greatly improve probability estimates for decision-making. To combine the probabilities, one must first learn some rules of how probabilities can be combined.

A.3.2. Rules of Addition

To apply one of the rules of addition, we must have a probabilistic situation that is characterized by a collectively exhaustive and mutually exclusive set of events. A *collectively exhaustive set of events* can be defined as the set of all possible events that can occur in an experiment. In a product failure test, either the product works or it fails to work. There could be two events (product works or product fails to work) that are collectively exhaustive, because each outcome of an event must fall into one of these two categories. *Mutually exclusive events* are events that are not related to one another. In other words, the probability of one event does not affect, alter, or impact in any way the probabilities of any other event. An event that is mutually exclusive cannot be counted or measured in more than one category during an experiment.

This concept is illustrated in **Figure A.1**. The Venn diagram on the left shows that P(A) and P(B) are mutually exclusive probabilities.

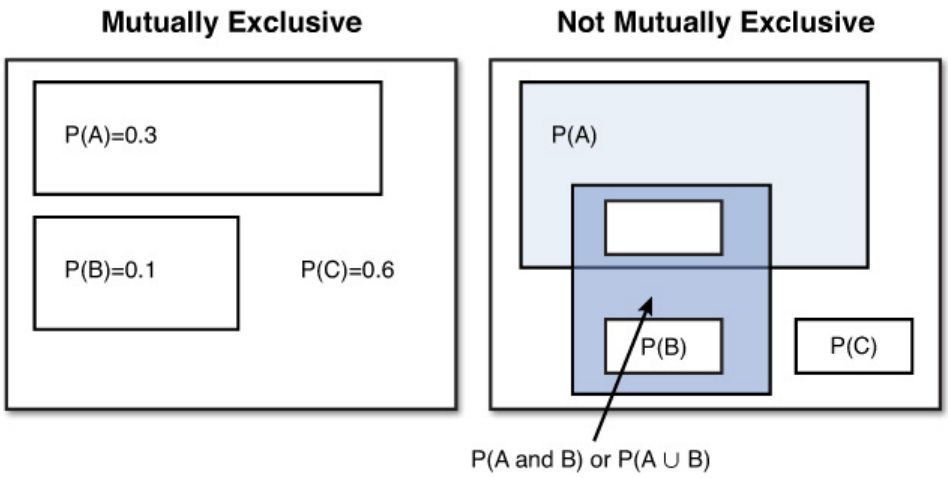


Figure A.1 Venn diagrams

The *rule of addition* of probabilities is that for a mutually exclusive and collectively exhaustive set of events, the probability of any collection of possible events is found by the summation of the probabilities of those events. In other words, to find the probability of two events, we need only add respective individual probabilities. We can express this rule of addition as follows:

Rule of Addition (for mutually exclusive events A and B): P(A or B) = P(A) + P(B)

In reading statistical literature, the expression of two probabilities being added together can be represented by any of the following: P(A) + P(B), P(A or B) (referred to as the probability of A or B), or P(A ∪ B) (referred to as the probability of A "union" B, where the ∪ represents the union of both the probabilities of A and B).

Question: If there are four mutually exclusive and collectively exhaustive events of P(A) = 0.20, P(B) = 0.40, P(C) = 0.15, and P(D) = 0.25, what are the following probabilities: P(A) + P(D), P(B ∪ C), P(A or B or C), and P(A +~A)?

Answer: P(A) + P(D) = 0.20 + 0.25 = 0.45, P(B ∪ C) = 0.40 + 0.15 = 0.55, P(A or B or C) = 0.20 + 0.40 + 0.15 = 0.75, and P(A + ~A) = P(A) + (1 − P(A)) = 0.20 + 0.80 = 1.00

Sometimes events are not mutually exclusive. In the Venn diagram on the right in **Figure A.1**, the notation where two events are possible, P(A and B), is used. This is called *a joint probability* or *compound event*. The Venn diagram in **Figure A.1** shows how the sample points of events A and B overlap. This overlapping of events means they are *not mutually exclusive*. The P(A) is jointly related to P(B). Note in the Venn diagram that if the total probabilities of P(A) and P(B) are added, the joint probability of P(A and B) would be double counted. To avoid this double counting for events that are not mutually exclusive, a revision of the rule of addition is needed, as stated earlier.
The *rule of addition* of probabilities for events that are not mutually exclusive is that the probability of any collection of possible events is found by the summation of the probabilities of those events minus the joint probability of those events. In other words, to find the probability of two events, we need to add their respective individual probabilities together and subtract their joint probabilities. We can express this rule of addition as follows:

Rule of Addition (for events A and B that are not mutually exclusive):

P(A or B) = P(A) + P(B) − P(A and B)

We can expand the rule of addition for additional events by adding the probability of the new event and subtracting all joint probabilities.

A.3.3. Rules of Multiplication

To apply one of the rules of multiplication, we must have a probabilistic situation that is characterized by independent event outcomes. *Independent event outcomes* are outcomes in which the probability of one event does not affect the probability of another event. Unlike the rules of addition, the rules of multiplication apply only to multitrial experiments. In a multitrial experiment, we have several event outcomes.

The *rule of multiplication* of probabilities is that, for independent event outcomes, the probability of any collection of possible events is found by the product of the probabilities of those events. In other words, to find the probability of two events, their respective individual probabilities should be multiplied together. This rule of addition is expressed as follows:

Rule of Multiplication (for independent events A and B):

$P(A \text{ and } B) = P(A) \times P(B)$

To illustrate this rule, assume that a purchasing manager faces a problem of selecting one of three component parts—A, B, or C—each week for the next two weeks. The probability of selecting Part A (P(A)) is 0.33. What is the probability that Part A will be selected in both Weeks 1 and 2? In this experiment, there are two events whose outcomes are independent, because the probability of Part A being selected in Week 1 will not alter the probability of Part A being selected in Week 2. The probability of Part A being selected in both weeks is P(A) × P(A), or 0.11 (0.33 × 0.33 = 0.11).

In reading statistical literature, the expression of two probabilities being multiplied together can be represented by any of the following: P(A) × P(B), P(A and B) referred to as the probability of A and B, or P(A ∩ B) (referred to as the probability of A and B, where the ∩ represents "and").

Question: If there are three independent probabilities of P(A) = 0.5, P(B) = 0.3, P(C) = 0.2, what are the values of P(A) × P(B), P(B and C), P(A ∩ C), and P(A and B and C)?

Answer: P(A) × P(B) = 0.5 × 0.3 = 0.15, P(B and C) = 0.3 × 0.2 = 0.06, P(A ∩ C) = 0.5 × 0.2 = 0.1, and P(A and B and C) = 0.5 × 0.3 × 0.2 = 0.03.

In an experiment with multiple events or trials, the outcomes of those events can be either independent or dependent. *Dependent event outcomes* occur when the probable outcome of one event affects the probable outcome of another event. For example, suppose there is a sample of five members of political Party A and five members of political Party B. Assume that the members of each party will vote for their own respective parties. Using this sample, an experiment is structured with two events. Event 1 randomly selects one of the ten members to ask whom that person will vote for, and then this party member is excluded from the sample. The process of excluding an element from a sample is called *sampling without replacement*, because one does not place the element, or party member in this example, back into the sample after its use. Event 2 in the experiment randomly selects one of the nine remaining members to ask whom that person will vote for. The probability of selecting a member of either Party A or Party B in Event 1 is 0.50 (5 / 10 = 0.50). The probability of selecting a member of either party in Event 2 is conditionally dependent on who was selected in Event 1. If, for example, a member of Party A is selected in Event 1, the probability of selecting a member of Party A in Event 2 is 0.44 (4 / 9 = 0.44), and the probability of selecting a member of Party B in Event 2 is 0.56 (5 / 9 = 0.56).

Conditional probabilities reflect the conditional dependency of probabilities. They are used when the event outcomes are not independent. In the political party example, the probability of selecting a member of Party A in Event 2 is conditional on what happened in Event 1. If we know that a member of Party A was selected without replacement in Event 1, the probability of selecting a member of Party B in Event

2 is 5 out of 9, or 0.56. So the behavior of the subsequent events needs to be given to determine later event probabilities when the events are not independent. We can express the conditional probabilities of the two-event experiment example as follows:

P(of selecting Party B member, given that we selected Party A member in Event 1) or P(B | A)

The expression of P(B | A) is read, "The probability of B given A," or "The probability of selecting B in Event 2, given that A was selected in Event 1." Conditional probabilities are used to adjust the rule of multiplication to permit the possibility of event outcomes not being independent. When the probabilities are not independent, we should use the following rule of multiplication:

Rule of Multiplication (for not independent events A and B):

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$

To illustrate this rule, consider the voter selection problem. What is the probability that a member of Party A will be selected first, without replacement, and then a member of Party B? In this experiment, there are two events whose outcomes are not independent, because the probability of Party A being selected in Event 1 will alter the probability of Party B being selected in Event 2. The probability of a member of Party A being selected in Event 1 is P(A) = 0.5, and the probability P(B | A) is 0.56 (5 / 9 = 0.56). So the probability of selecting a member of Party A and a member of Party B follows:

$$\begin{aligned} P(A \text{ and } B) &= P(A) \times P(B \mid A) \\ &= 0.5 \times 0.56 \\ &= 0.28 \end{aligned}$$

This probability means that there is a 28 percent chance that a member of Party A will be selected from the sample of ten, and then a member of Party B will be selected from the remaining sample of nine members.

Question: What is the probability of three members of Party A being selected, without replacement, in the sample of ten members mentioned earlier?

Answer:

$$\begin{aligned} P(A_1 \text{ and } A_2 \text{ and } A_3) &= P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_2) \\ &= (5 / 10) \times (4 / 9) \times (3 / 8) \\ &= 0.09 \end{aligned}$$

A.4. Probability Distributions

There are several basic terms that are used to describe probabilistic situations and distributions: experiments, trials, outcomes, and events. An *experiment* is defined as a test or a *trial,* or a set of tests or trials, in which an operation is conducted to discover unknown behavior. A consumer survey is an experiment that can be used to determine unknown consumer behavior. Each time a meal is prepared in a restaurant for a customer, it represents an experiment of the chef trying to satisfy the customer's tastes. An *outcome* is defined as the result of the experiment. The results of a consumer survey or the preparation of a meal can have a successful or unsuccessful outcome. An *event* is also defined as the outcome of an experiment. If an experiment consists of only one trial, the results of the experiment can be called an outcome or an event of the experiment. Events usually represent the different types of outcomes that are possible in an experiment when the experiment consists of multiple tests or trials. For example, in a product failure study, products are tested to determine the probabilities of a product working or failing. In this two-outcome situation, the probability of a product working is one event, whereas the probability of a product failure is another event.

We can define a *random variable* as any value resulting from a random experiment that by chance can generate different values. *Probability distributions* are described in the form of a graph, table, or formula, the probable behavior of a random variable's events in an experiment. There are two types of random variables, and two types of probability distributions to describe them. A *discrete random variable* is one whose events are integer values, starting with 0, 1, 2, and so on to some positive value that can be counted. As a trial is repeated and its events recorded into a frequency distribution, it can become a probability distribution. A *discrete probability distribution* specifies the probability associated with each possible event of the discrete random variable. All discrete probability distributions have two characteristics: the probabilities of the random variables are greater than or equal to zero, and the summation of all the random variable's probabilities is equal to one. A *continuous random variable* is one whose events can assume any integer or noninteger (real number) value. Any values that can be measured in decimals such as height, weight, or length are values that fall within an interval that can be considered continuous random variables.

Question: How many possible continuous variable weight values are there between 150 and 151 pounds?

Answer: An infinite number depending on how precise the values need to be.

The probability distribution for a continuous random variable is called a *continuous probability distribution*, a *probability density function*, because it describes how the probability is distributed within the function of the random variable, or a *frequency function*, because it originates from frequencies. The continuous probability distribution specifies the probability associated with each possible event or interval of event occurrences for the continuous random variable. Like a discrete probability distribution, a continuous probability distribution has two characteristics: the probabilities of the random variables are greater than or equal to zero, and the summation of all the random variables' probabilities is equal to one. In addition, the continuous probability distribution assumes that the range of events is infinite.

What follows will primarily present probability distributions as a means of obtaining probability statistics that describe an experiment or business decision-making situation consistent with the descriptive step of BA. It should also be mentioned that probability distributions provide the foundation for most of the inferential or predictive statistical methods discussed in other steps of the BA process. SPSS and Excel

use basically the same approach for the assessment and calculation of probability distribution values. This approach involves defining the type of probability distribution and then placing the needed parameters in the software so that probability estimation can be calculated. To avoid duplication of the printouts, the probability distributions in this section will be illustrated using only Excel.

A.4.1. Discrete Probability Distribution

This section examines a series of discrete probability distributions, including the binomial, Poisson, geometric, and hypergeometric. For each probability distribution, its formula and parameters are presented (to understand the need for the model's parameters), along with the distribution characteristics (to identify where the probability distribution should be applied), and an example is given (to show how to obtain the probabilities, know what information they provide, and observe some computer-generated solutions).

A.4.1.1. Binomial Probability Distribution

The *binomial probability distribution* is a discrete probability distribution in which the outcome of an experiment is limited to two events, like success or failure, yes or no, true or false. To simplify the two-outcome nature of this distribution, the generally accepted "success" or "failure" outcomes will be used in describing the two-outcome distribution.

The binomial probability distribution formula used to calculate the probability of a success is:

$$P(r) = \frac{n!}{r!(n-r)!}(p)^{r}(q)^{n-r}$$

where:

r = exact number of successes observed or desired (the random variable)

P(*r*) = probability of the exact number of successes occurring

n = number of trials in the experiment

p = probability of a success, *r*, on each trial

q = probability of a failure on each trial, or $q = 1 - p$, because $q + p$ must equal 1 in a two-outcome trial

The characteristics of this binomial probability distribution include the following:

- In a single trial or outcome of an experiment, there can be only two outcomes.
- The two outcomes for any of the *n* trials are mutually exclusive.
- The probability of *r* remains constant for each trial in the *n* trials of the experiment.
- The *n* number of trials is independent, so the result of one trial does not affect the outcome of any other trial.

• The data used to establish the distribution is based on counting all the n possible trial results.

• The binomial probability distribution tends to be positively skewed when the probability of success, P(r), is less than 0.5 and the number of n trials is fairly small (the number of trials is close to zero).

Question: Suppose we have just ordered 100,000 light bulbs. We want to test the quality of the bulbs to see if they will work before we accept them from the manufacturer. We have purchased the light bulbs with the agreement that up to 10 percent of the bulbs can be defective before rejecting the lot. To check out the quality of the lot, we take a sample of 10 light bulbs (i.e., n=10) to test. We will test all 10 light bulbs, and if either 0 light bulbs (r = 0) or 1 light bulb (r = 1) is defective, we will still accept the entire lot of 100,000. If 2 or more light bulbs are defective and fail to light from the sample of 10, the entire lot will be rejected. Based on past experience with hundreds of thousands of light bulbs, we know that the probability of any one bulb being defective and failing to light is only 0.05 (p = 0.05); the probability of any single bulb being nondefective and lighting on the first try is 0.95 (q = 0.95). What is the probability of 0 or 1 bulb (10 percent or less in the sample of 10 bulbs) being defective?

Answer: To successfully use the binomial probability distribution as is the case in other probability distributions, identify the distribution's variable (r) and the parameters (n, p, and q), and match the characteristics of the problem with the binomial probability distribution characteristics. Reviewing the six binomial probability distribution characteristics in this light bulb problem, one can see how this problem fits this type of distribution. Remember that finding a defective light bulb is an *r outcome*; the probability of finding a defective light bulb, p, is 0.05; each time a light bulb is tested it is a *trial*; the *experiment* consists of ten light bulb tests or trials. The characteristics follow:

• For any single light bulb test, there are only two outcomes: a defective or a nondefective light bulb.

• The defective or nondefective outcomes in any test trial are mutually exclusive.

• The probability of finding a defective light bulb r remains constant at 0.05 for each test trial in ten tests of the experiment.

• The ten test trials are independent, so the results of one test do not affect the outcome of any other test.

• The data used to establish the probability distribution is based on counting the ten test combinations taking r = 0, 1, 2, to 10 at a time.

• The resulting binomial probability distribution for the light bulb test is presented in **Figure A.2** and is positively skewed because the probability of success is p = .05 and the number of trials (n = 10) is fairly small and close to zero.

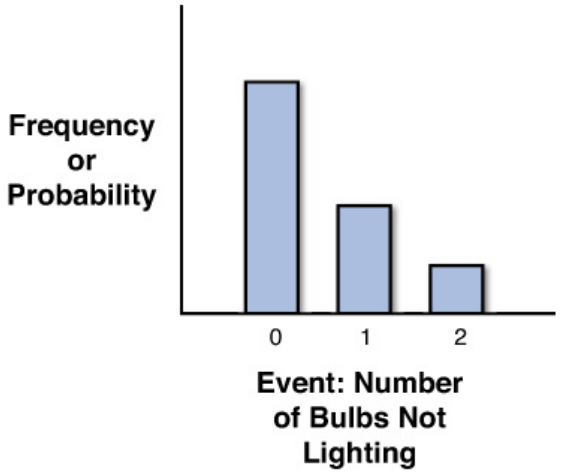


Figure A.2 Binomial probability of light bulbs not lighting

In this problem, we are trying to find out the probability of not just one event occurring, but two events ($r = 0$ and $r = 1$). Remember, when there are two mutually exclusive events occurring, add their probabilities together to determine their union of probable occurrence. This is accomplished quite easily with the binomial probability distribution software functions in SPSS and Excel. Simply compute the probability of P($r = 0$), where $p = 0.05$ and $n = 10$, and then add it to P($r = 1$), where $p = 0.05$ and $n = 10$ using the software functions. Using the Excel function BINOM.DIST, one can automatically compute the cumulative probabilities of P($r = 0$ or $r = 1$), as shown in **Figure A.3**. As we can see in the figure, the probability of 0 or 1 light bulbs being defective is P($r = 0$ or $r = 1$) = 0.9138. Cumulative probabilities are commonly used in business applications. By simply adding the values on the tabled probability distributions, these values are quickly derived, as can be seen in the preceding problem. Note, to find the exact probability of 0 or 1 or 2, type FALSE instead of TRUE into the Excel window in **Figure A.3**.

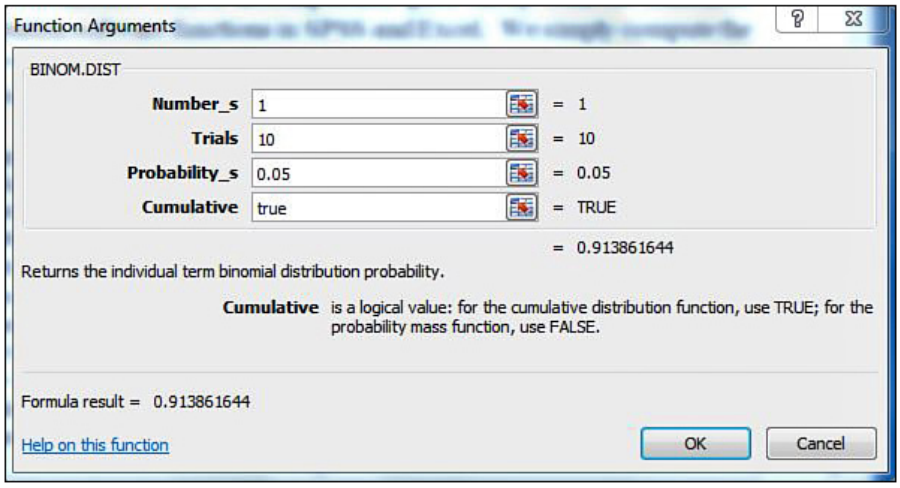


Figure A.3 Excel binomial probability of light bulbs not lighting

---

A.4.1.2. Poisson Probability Distribution

The Poisson probability distribution is like the binomial except when the number of trials is very large and the probability of success is very small. The Poisson distribution is particularly useful when one is interested in determining the probability of some number of events occurring during a specific time period (for example, the probability of exactly four people waiting in line for a service) or in a specific area (for example, the probability of one or more paint blemishes occurring on a product).

The Poisson probability distribution formula used to calculate the probability of a random event occurring follows:

$$P(X) = \frac{e^{-\mu}\mu^{x}}{X!}$$

where:

X = exact number of successes observed or desired

$P(X)$ = probability of the exact number of successes occurring

e = the constant value of 2.71826

μ = mean number of X successes occurring per unit of time, area, volume, and so on

The characteristics of this Poisson probability distribution include:

• The random variable X must be a nonnegative integer.

• The experiment consists of counting the number of times a single event occurs for a given time period, in an area, or in some volume. Within the time, area, or volume of the experiment, there is a large number of possible event occurrences, and the probability of any event is small.

• The number of events in one unit of time, area, or volume is independent of the number of any other units.

• The mean or expected number of success occurrences μ must remain constant for the same time period in an area or in some volume.

• The Poisson probability distribution tends to be positively skewed because μ must remain constant for the same time period in an area or in some volume.

As with the binomial distribution, to successfully use the Poisson probability distribution, identify the distribution's variable (X) and the parameter (μ) and match the characteristics of the problem with the Poisson probability distribution.

Question: Suppose we are planning on purchasing a printer to handle computer jobs that need to be printed. The size of the jobs and the flow to the printer is Poisson distributed with a mean arrival rate of two jobs per minute ($\mu = 2$). Model A has temporary storage areas to hold computer jobs until the printer can handle them. The temporary storage areas are called *buffers*. Unfortunately, the way the computer system is set up, if the job does not go into buffered storage, it has to be stored in the main computer at an almost prohibitive cost. The selection of the printer will be based in part on its probable ability to handle the flow of jobs from the computer without requiring costly computer storage. Model A can buffer up to five jobs at a time. What is the probability that Model A will have to store exactly three jobs in its buffer storage per minute?

Answer: Again, it was necessary to make a number of assumptions to fit the Poisson characteristics of

this problem. To the extent that the assumptions are accurate, so are the statistics based on them. We can review the first five of the Poisson probability distribution characteristics in this computer problem so we can see how this problem fits this type of distribution. Remember that the *random variable X* is the number of jobs to store in the printer in a minute; the *mean arrival rate* μ is two jobs per minute; and the *experiment* consists of counting the number of jobs that occur in a minute. The characteristics are as follows:

• The random variable *X* is an integer value of 0 jobs, 1 job, 2, jobs, and so on, to store in the printer's buffer storage.

• The experiment consists of counting the number of jobs arriving in a minute, and because there could be an infinite number of jobs arriving, the probability of any job arriving could be assumed to be infinitely small.

• The probability of any job arriving can be assumed to be equal to any other job arriving.

• The arrival of jobs in one minute is assumed to be independent of the number of jobs arriving any other minute.

• The mean number of jobs arriving μ is assumed constant for any minute.

Using the Excel function POISSON.DIST, we can automatically compute the cumulative probabilities of $P(X = 3)$, as shown in **Figure A.4**. As we can see in the figure, the probability of $P(X = 3) = 0.180447$.

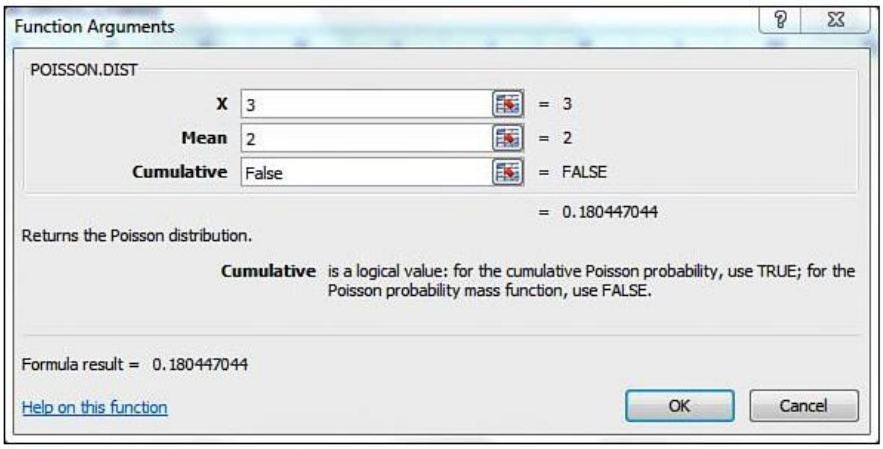


Figure A.4 Excel Poisson probability of three print jobs in the buffer

---

A.4.1.3. Other Discrete Probability Distributions

There are many other discrete probability functions. Two additional probability distributions commonly used in business analytics include those in **Table A.1**. These can also be computed using SPSS or Excel function statements.

Probability Distribution	Description	Application Examples

Geometric	Similar to the binomial in that there is a two-outcome situation of success r or failure, the p probability of success occurring, and q probability of failure of the event not occurring. Unlike the binomial, one is interested in the first success that will occur out of some unknown number of n trial periods. This distribution is particularly useful when the interest is in determining the probability of some events occurring in a specific number of discrete time periods (for example, in hours, days, minutes).	Examples of the use of this distribution can include determining the probability that a customer will wait two or more minutes in a queue for service; the probability that out of four people surveyed, one will buy a new product; and the probability that one will have to sample at least 20 tax returns before finding the first person who will need to be audited.
Hypergeometric	Similar to the binomial distribution in that it is a two-outcome situation of success r or failure, but with this distribution the trials are dependent. The dependence of the trials occurs because the results are taking place *without replacement*. In using this probability distribution, unlike all other distributions discussed so far, a total number or *population* of elements must be specified, denoted N, from which sampling takes place without replacement. This distribution is based on counting and does not require known probability parameters in its computation.	Examples of the use of this distribution include determining the probability of lot rejection based on lot size and sample size, the probability that a person has been discriminated against based on population size, and determining the proportion of voters that will vote for a specific political party's candidate based on a sample of registered voters.

Table A.1 Other Discrete Probability Distributions

A.4.2. Continuous Probability Distributions

Two of the most common continuous probability distributions are the normal and the exponential. With these distributions, we compute the area under the curve between two points or from one point out to infinity (the tails of the distribution).

A.4.2.1. Normal Probability Distribution

The most commonly used probability distribution is the *normal probability distribution*. It is a continuous probability distribution, where the outcomes of an experiment are expressed as a continuous function. This distribution can be applied in almost any problem situation in which the event in an experiment follows a continuous function and meets the distribution's characteristic requirements.

The normal probability distribution formula used to calculate the probability of the function of X for a specific measurable random variable X value follows:

$$f(X) = \frac{1}{2\pi} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

where:

X = a measurable normal random variable

f(X) = probability of exactly X occurring (usually referred to as a function of X)

σ^2 = variance of the normal random variable

μ = mean of the normal random variable

π = a constant of 3.1416

e = a constant of 2.71828

Unlike the discrete distributions, the characteristics of the normal probability distribution are more directed at describing the distribution, rather than an experiment that the distribution will be used in. The characteristics of the normal probability distribution include these:

• The mean occurrence of the measurable random variable *X* is μ, with a standard deviation of σ.

• The random variable *X* ranges over $\pm\infty$. (This characteristic is a theoretical requirement that is, of course, not observed in practical applications.)

• The shape of the curve is *bell shaped*, like that in **Figure 5.7** in **Chapter 5**.

• The distribution is evenly divided by its mean value of μ.

• The distribution is *symmetrical*, with an equal number of values falling on both sides of μ.

• The distribution is *asymptotic*, so the tails of the curve drop off toward the horizontal axis but never touch it.

With a continuous function, we are interested in probabilities and usually not the value of the normal function f(*X*). The value of f(*X*) on a continuous function would represent a very small probability. That is, the probability of any specific value of *X* occurring out of the infinite values of *X* has an infinitely small probability of occurrence. To determine the probabilities of the normal probability distribution, we convert the area under the continuous function into a probability space, as presented in **Figure A.5**.

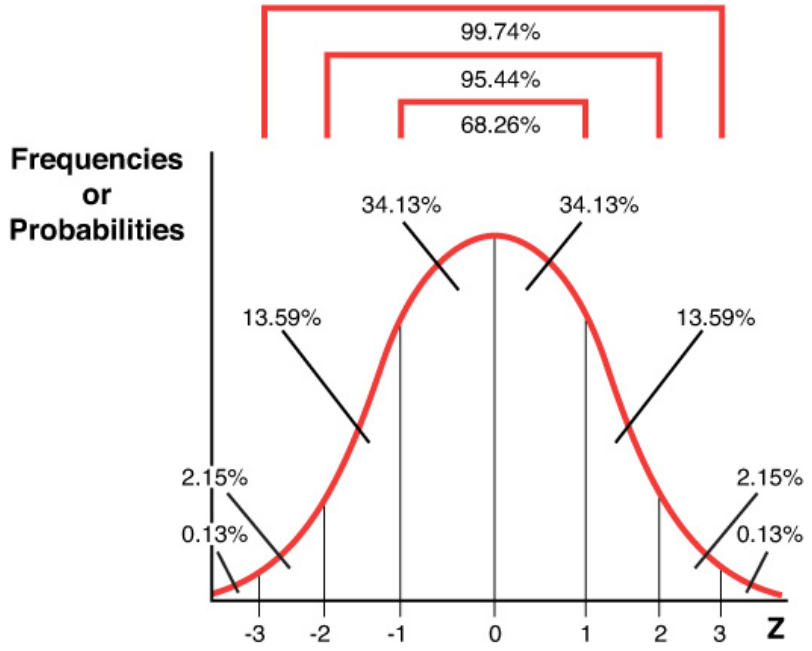


Figure A.5 Normal distribution with Z values

The actual determination of probabilities can be difficult even though a normal distribution fits the preceding characteristics. Unfortunately, the shape of the curve and the distribution's parameters impact probability calculations. Every distribution found in the real world might qualify for a normal distribution but would have differing areas under the curve (more or less skewedness). To deal with the variety (or sometimes called *family*) of normal probability distributions, mathematicians developed a single distribution called the *standard normal probability distribution* (see **Chapter 5**, **Section 5.5**). The characteristics for this distribution are the same as the normal probability distribution. The standard normal probability distribution formula used to determine the probabilities for any normal probability distribution follows:

$$Z = \frac{X-\mu}{\sigma}$$

where:

Z = the number of standard deviation units between X and μ

X = a specific random variable value

μ = mean of the normal random variable

σ = standard deviation about μ

The standard normal probability distribution has a mean of 0 and a standard deviation of 1. That is, μ is set equal to zero, and we can add and subtract units of σ to denote areas or probability spaces in the distribution. The measurable variable X is converted into units of standard deviation called Z *value*. As seen in **Figure A.5**, the area under the normal curve covering $\pm 1\sigma$ is the same as a Z value of ± 1. The area under the normal curve of a Z value ± 1 represents 68.26 percent; in other words, the probability that X will occur between $\pm 1\sigma$ is 0.6826.

Question: What is the probability that X will fall between ±2σ in a normal probability distribution?

Answer: The probability can be seen in Figure A.5 as 0.9544, or 95.44 percent.

The *Z* values provide a standardized measure of deviation units. Because all normal probability distributions have a standard deviation, the *Z* values provide a common or standardized measure by which all normal probability distributions can be linked. Once the standardized normal probability distribution values are determined, they can be used as a standard to approximate probabilities for all the other normal probability distributions. Indeed, the purpose of the standard normal probability distribution is to provide the probabilities for all normal distributions, given a *Z* value.

Question: Suppose we use an automated piece of equipment to fill metal cans with a beverage. The metal can that the machine fills is designed to hold as much as 14.5 ounces of beverage, but the plan is to put only 12 ounces in a metal can. After observing the first 1,000 cans filled by the machine (representing the population of the machines' can fill activities), we find that the mean number of ounces is 12 (μ), with a standard deviation of 0.5 ounces (σ). Based on this information, what is the probability that the amount of beverage filled in a can will be between 12 and 13 (the random variable) ounces?

Answer: We will assume that this problem fits the normal distribution. The probability distribution of this problem is presented in Figure A.6. For this question, determine the shaded area under the curve presented in the graph. To do this, use the *Z* formula to compute the *Z* value representing the area between $\mu = 12$ and a specific value of *X*, or $X = 13$. The *Z* value is 2 (13 − 12 / 0.5). So what is needed is to determine the probabilistic conversion of a *Z* value of 2 standard deviations into the area under the curve or a probabilistic occurrence of the *X* falling between 12 and 13 ounces ($P(12 \leq X \leq 13)$).

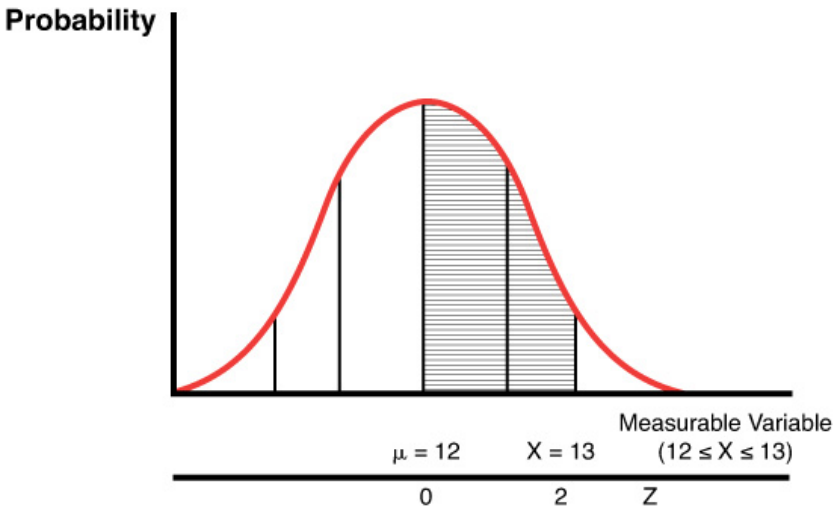


Figure A.6 Normal distribution example with Z value

Using Excel's NORM.S.DIST function, we can compute this probability. This function provides cumulative probabilities from some *X* point to the origin (all the area to the left of the distribution from random variable *X*). Using this function, the *Z* value is the only parameter needed, but it requires two steps. First, to find the cumulative probability from the *Z* value of 2 where $X = 13$, and second, to find the probability from the *Z* value of 0 where $X = 12$. The resulting probability from the *Z* value of 2 as presented in Figure

A.7 is 0.97724. This value includes all the probabilities to the left of μ, which is 0.5, or half the area under the curve. That area is where $X = 12$, so the resulting probability that a can will be filled between 12 and 13 ounces is 0.47724 (0.97724 – 0.5).

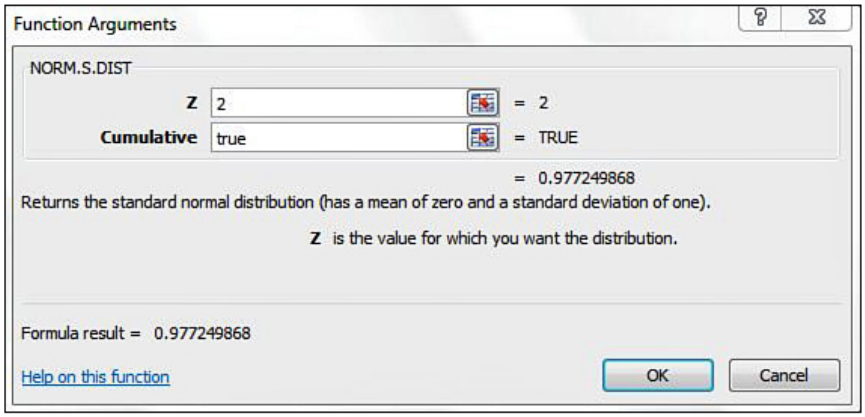


Figure A.7 Excel's computation of normal probability example

A.4.2.2. Exponential Probability Distribution

The exponential probability distribution is a continuous probability distribution, where the values of the variable *X* are measurable and expressed as a continuous function. This distribution is similar to the Poisson probability distribution in that it deals with time in the distribution. Unlike the Poisson distribution, though, time is not fixed but is the actual variable. The exponential probability distribution has been extensively used in queuing analysis to estimate the probable time required of a service facility to process a customer. In a similar way, the exponential probability distribution is used to determine computer processing time of computer software by computer hardware.

The exponential distribution formula is expressed as a continuous function and will be present in a queuing context of measuring the timing of units arriving at a service facility. The continuous function of the exponential distribution follows:

$f(X) = \lambda e^{-\lambda a}$

where:

X = a measurable random variable representing the time between successive arrivals to the service facility

f(*X*) = value of the function at *X*

a = the specific value of *X* whose probability is being sought

λ = a Greek letter, *lambda,* representing the average rate of arrivals to a service facility (like the μ of the Poisson distribution)

e = a constant of 2.71828

We calculate the area under the curve and *X* between *a* and ±∞ to determine the probability of the function. Here's the formula for determining the probability of *a* or more units being served:

$P(\geq a) = e^{-\lambda a}$

where:

$P(X \geq a)$ = probability of $X \geq a$ occurring

Following are the characteristics of the exponential probability distribution:

• The mean arrival rate of the measurable random variable X is λ.

• Both X and λ must be positive, and "a" must be within the range of X.

• The random variable X ranges from 0 to $+\infty$. (This characteristic is a theoretical requirement that is, of course, not observed in practical applications.)

• The experiment consists of determining the probability of X having a or more arrivals when the mean arrival rate of λ is known and continuous.

• The shape of the curve tends to be positively skewed (see **Figure A.8**).

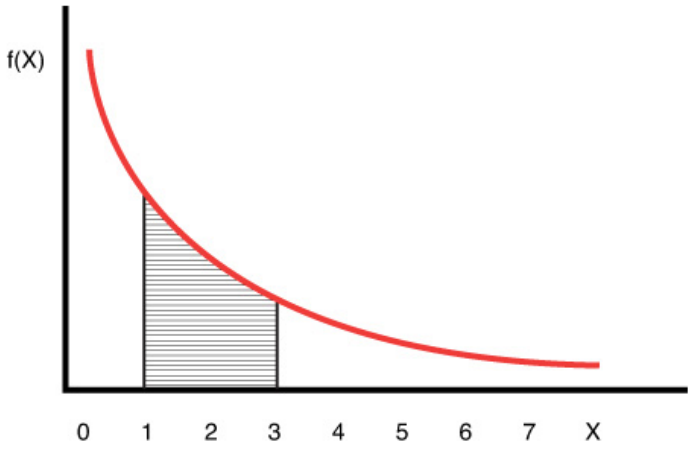


Figure A.8 Exponential probability distribution example

For this continuous exponential probability function, the focus is on the probabilities and usually not the value of the exponential function f(X). To determine the probabilities of the exponential probability distribution, convert the area under the continuous function into probabilities as with the normal distribution. To successfully use the exponential probability distribution, identify the distribution's variable (X) and the parameter (λ), and match the characteristics of the problem with the exponential probability distribution characteristics.

Question: Suppose we have been trying to find out the probable processing time for jobs on a computer. Most jobs require little or almost no processing time, but a few jobs do take more time. From internal processing records, we know that the mean processing rate is one job per microsecond (λ = 1 job per microsecond). What is the probability that a computer job will take from one to three microseconds, inclusively?

Answer: Assume that this problem meets the characteristics of the exponential probability distribution. The shaded region in **Figure A.8** is the probability area that needs to be determined. Using Excel's

EXPON.DIST function, we only need to determine the distribution's variable (X) and the parameter (λ). The random variable is an interval value $P(1 \leq X \leq 3)$ and $\lambda = 1$. The EXPON.DIST function provides cumulative probabilities (from 0 out to X in **Figure A.8**), as does the NORM.S.DIST function. This means the calculation requires first determining P(X = 1) and subtracting it from P(X = 3). The Excel exponential probability calculation for P(X = 1) is presented in **Figure A.9** as 0.63212. It turns out that P(X = 3) = 0.95021, resulting in $P(1 \leq X \leq 3) = 0.31809$. So the probability of a computer program taking from 1 to 3 microseconds is 31.809 percent.

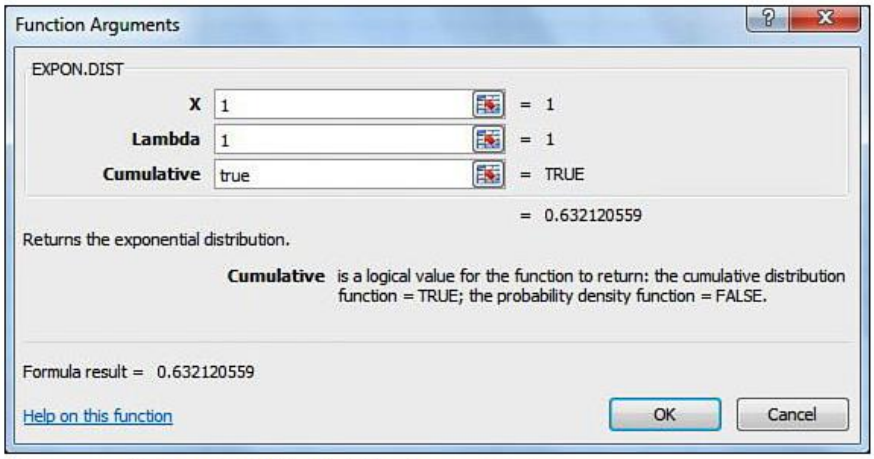
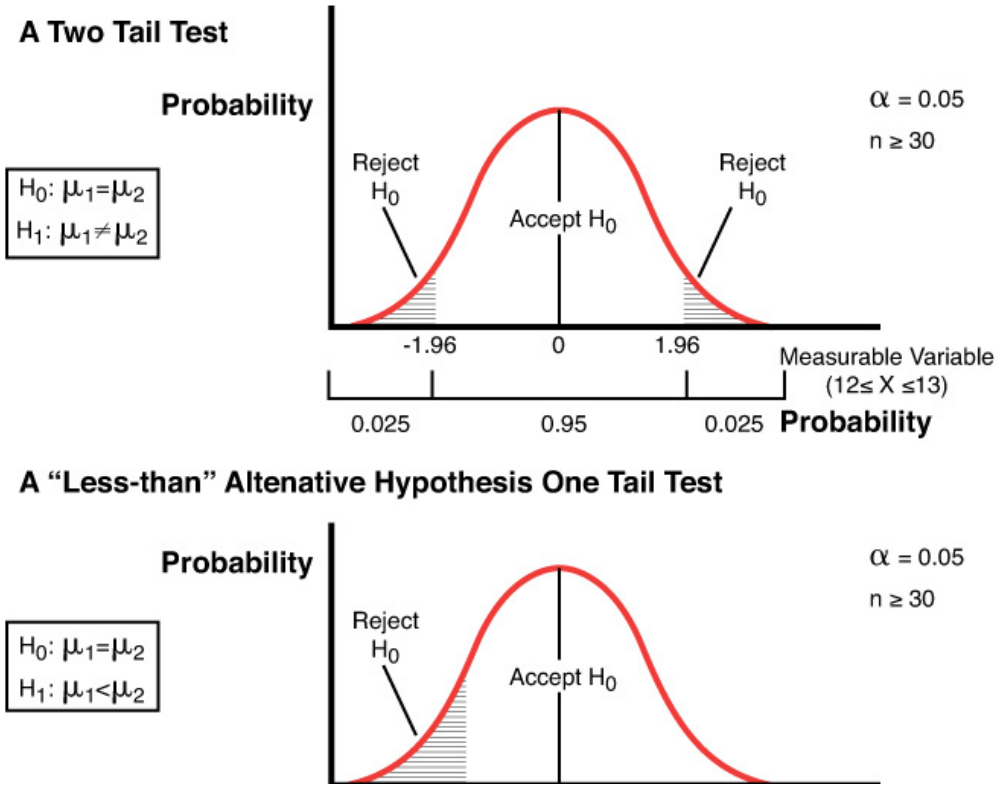


Figure A.9 Excel exponential probability distribution example

A.5. Statistical Testing

In **Chapter 5**, "**What Are Descriptive Analytics?**" when confidence intervals were discussed, the focus was on creating boundaries when a population parameter was expected to fall, given a specific confidence level (the number of plus or minus Z values and the related percentages; see **Table 5.6** in **Chapter 5**). As can be seen from the two-tailed normal distribution in **Figure A.10**, a 95 percent confidence level is computed from a Z value of 1.96.



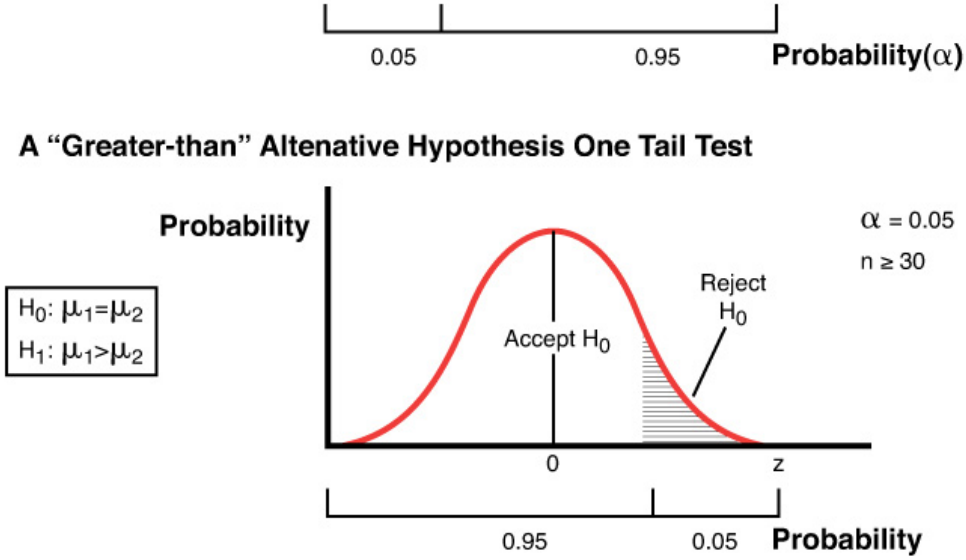


Figure A.10 Hypothesis testing, one- and two-tailed tests

We are assuming use of a confidence coefficient such that the true population parameter will fall within the confidence interval. What if it actually falls outside the interval? That is the question that hypothesis testing helps determine. In a scientific and formal manner, all hypothesis tests have a *null hypothesis* (designated as H_0). H_0 implies that there is no significant difference between the mean that we have from our sample (or some other measure of central tendency) and the population mean. The interval in **Figure A.10** also means that the 95 percent confidence allows a 5 percent chance the population mean is not in the interval. That 5 percent is called the *level of significance*, because it marks the boundary of where one designs a significant difference between the sample mean and a population mean. The level of significance is often labeled with a small alpha, α. So in **Figure A.10**, a 95 percent confidence level will have a 5 percent level of significance. If it turns out that the sample mean is significantly different from the expected population mean, we mark the occasion by rejecting the H_0 and accepting the *alternative hypothesis*, H_1. H_1 can be an inequality, a greater-than or equal-to or a less-than or equal-to expression depending on what is being tested. Note in **Figure A.10** that there can be a two-tail test that is for the inequality expression and a one-tail test that is for greater-than or equal-to expressions.

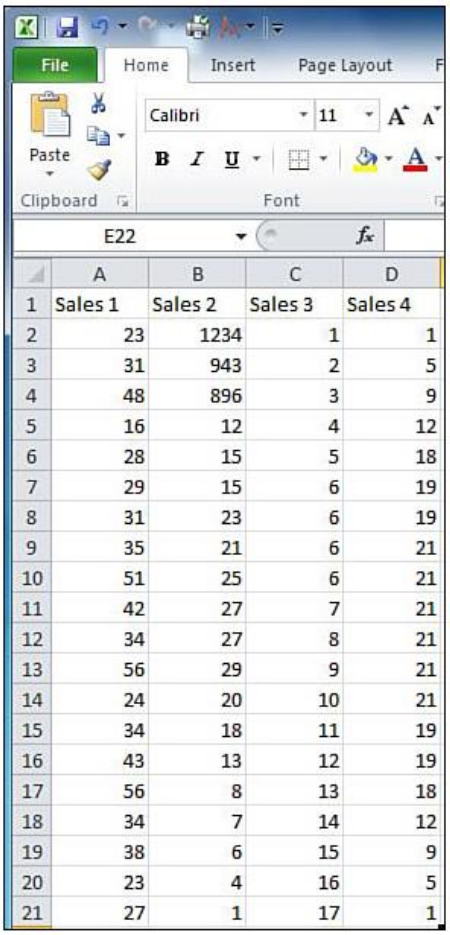
How does one determine if there is a significant difference between the random variable, which can include means, proportions, variances, or any kind of measure of central tendency computed from a sample, and the confidence interval value boundaries? These tests are accomplished through a variety of statistical tests (see **Table A.2**). When the statistic being compared is a population parameter based on a probability distribution, the hypothesis test is called a *parametric hypothesis test*. Alternatively, for hypothesis tests with measures that are not population parameters (based on counts or frequencies), use a *nonparametric hypothesis test*. There are many of both hypotheses tests in the literature. Both SPSS and Excel offer an impressive array of these tests, many of which are built into functions that can provide a level of significance on which statistics can reveal useful information for a BA analysis.

Test Statistic	Application Area	Access to SPSS Function	Access to Excel Function
F-Test Two Sample for	Compares the variances from two samples to see	Analyze > Compare Means > One-Way	Data Analysis > F-Test Two Sample for Variances

Variances	if they are from the same probability distribution	ANOVA	
t-test: Paired Two Sample Means	Compares the mean values from two samples to see if they come from the same probability distribution	Analyze > Compare Means > Paired Samples t-test	Data Analysis > t-test: Paired Two Sample Means
Z-test: Two Sample Means	Compares the mean values from two populations to see if they have the same probability distribution	For samples sizes above 30: Analyze > Compare Means > Paired Samples t-test	Data Analysis > Z-test: Two Sample Means
ANOVA: Single Factor	Compares the variance between and within two or more samples to see if the samples are drawn from the same probability distribution	Analyze > Compare Means > One-Way ANOVA	Data Analysis > ANOVA: Single Factor

Table A.2 Common Parametric Statistical Tests and Software Access Information

To illustrate the use of one of the parametric statistical tests, use the sales data presented in **Chapter 5** (**Figure 5.1**), which is repeated here in **Figure A.11**. To compare the mean sales data for Sales 1 with that of Sales 2 to determine if they are closely enough distributed, use either one for BA predictive purposes. In other words, it is important to determine if they are from the same probability distribution. Assume these values are paired with each other for this example. Because the data are samples, use the t-test rather than the Z-test because the Z-test is used to make comparisons using population-sized data.



	A	B	C	D
1	Sales 1	Sales 2	Sales 3	Sales 4
2	23	1234	1	1
3	31	943	2	5
4	48	896	3	9
5	16	12	4	12
6	28	15	5	18
7	29	15	6	19
8	31	23	6	19
9	35	21	6	21
10	51	25	6	21
11	42	27	7	21
12	34	27	8	21
13	56	29	9	21
14	24	20	10	21
15	34	18	11	19
16	43	13	12	19
17	56	8	13	18
18	34	7	14	12
19	38	6	15	9
20	23	4	16	5
21	27	1	17	1

Figure A.11 Illustrative sales data sets

Using the appropriate access to the software functions (**Table A.2**) and entering the data from **Figure A.11** for Sales 1 and 2, the software can compute the necessary test statistics. Because we are not looking specifically for one mean to be larger than the other and direction does not matter, we can use a two-tail test. The resulting SPSS (**Table A.3**) and Excel (**Figure A.12**) test statistics are presented next. Both software compute the t-test statistic as –1.573. (Note with a two-tail test, either negative or positive values can occur.) This is then compared to the related area under the curve in the probability distribution in terms of units of deviation. The result is that the software computes the resulting significance level, which is p = 0.132. The *p* symbol is often used when the exact test probabilities are computed, as opposed to the desired significance level of alpha, α. If a set desired significance level is 5 percent ($\alpha = 0.05$), it could be concluded that there is no significant difference in the paired comparison of the mean values from the same distribution of sales. So formally, we would accept H_0 and, therefore, accept H_1 if this example was formally structured as a hypothesis test. To have concluded there is a significant difference between the means, the resulting level of significance, *p*, would have had to have been less than 0.05. So for any of the test statistics, the computed level of significance is an essential test statistic result used to judge the outcome of the statistical test. The greater the computed level of significance, *p*, the less significantly different the parameters are assumed to be.

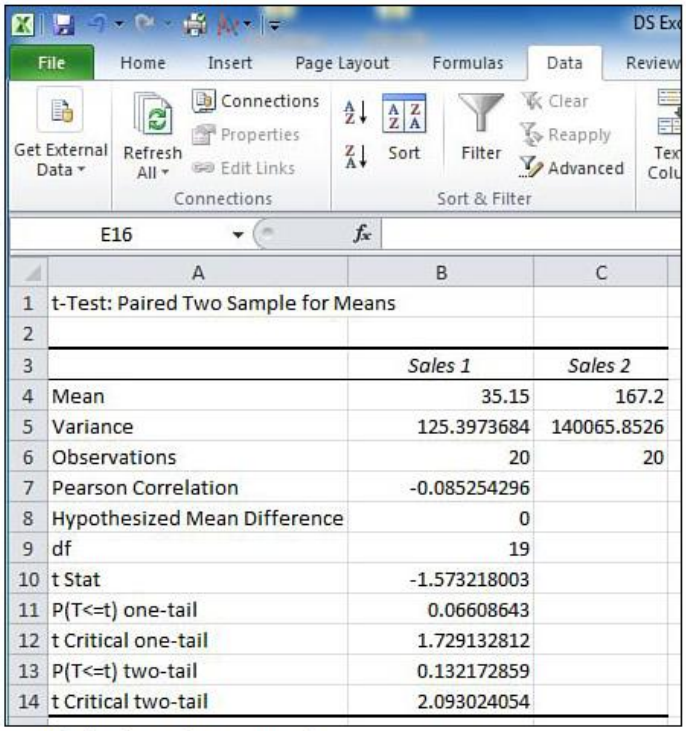
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Sales 1	35.15	20	11.198	2.504
	Sales 2	167.20	20	374.254	83.686

Paired Samples Test

		Paired Differences					t	df	Sig. (2-Tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Sales 1 – Sales 2	–132.050	375.374	83.936	–307.731	43.631	–1.573	19	.132

Table A.3 SPSS T-Test Statistics for Sales Example



	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*Sales 1*	*Sales 2*
4	Mean	35.15	167.2
5	Variance	125.3973684	140065.8526
6	Observations	20	20
7	Pearson Correlation	-0.085254296	
8	Hypothesized Mean Difference	0	
9	df	19	
10	t Stat	-1.573218003	
11	P(T<=t) one-tail	0.06608643	
12	t Critical one-tail	1.729132812	
13	P(T<=t) two-tail	0.132172859	
14	t Critical two-tail	2.093024054	

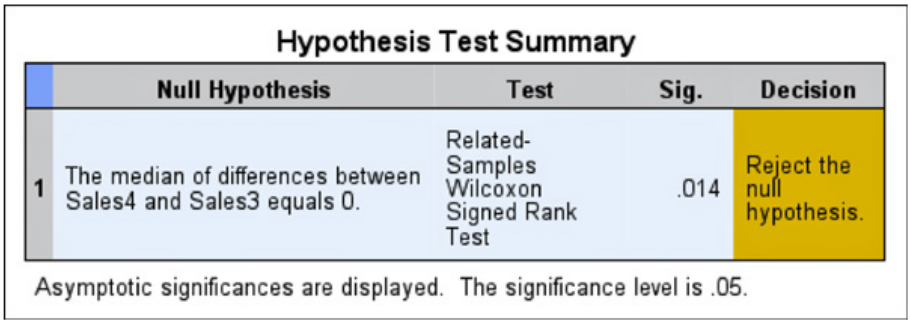
Figure A.12 Excel t-test statistics for sales example

Unlike the Z-test, t- and F-tests do not use means (a parameter) in their computed value computations. These types of hypothesis tests can be grouped into what is descriptively called *nonparametric tests*, or *distribution free tests*. Table A.4 lists some of the more common of these statistical tests.

Test Statistic	Application Area
Binomial Test	This test compares the observed frequencies of the two categories of a dichotomous variable to the frequencies that are expected under a binomial distribution with a specified probability parameter (α).
Chi-Square	This test (used often for a goodness-of-fit test) compares the observed and expected frequencies in each category of a distribution to test that all categories contain the same proportion of values or to test that each category contains a user-specified proportion of values.
Kolmogorov-Smirnov (One-Way)	Multiple versions of this test can be applied to differing comparative analyses. The one-way procedure compares the observed cumulative distribution function for a variable with any specified theoretical distribution.
Wilcoxon Signed-Rank	Multiple versions of this test can be applied to differing comparative analyses. This test is applicable when two samples (two populations) are related (not independent). The test is designed to compare some n matched pairs of ranked or ordinal data from two populations.
Run	A *run* is a sequence of like observations. This tests whether the order of occurrence of two values of a variable is random. A sample with too many or too few runs suggests that the sample is not random.

Table A.4 Common Nonparametric Statistical Tests and Software Access Information

Each software system provides a differing number of nonparametric tests. Excel's add-in function, CHISQ.TEST, provides the Chi-Square test statistic value but does not include computation of the level of significance. SPSS provides access to all the nonparametric tests in **Table A.4**, which are accessed by using Analyze > Nonparametric Tests > Related Samples as a navigation path. This particular set of functions helps the BA analyst by allowing the SPSS software to determine the best test to select for the analysis. For illustration, a comparison of the Sales 3 and 4 distributions is presented in **Figure A.13**. Note that SPSS chooses the Wilcoxon Signed Rank Test from those in **Table A.4** as the best choice for this analysis. Note also that the significance level of 0.05 is an automatic default (which can be changed), and that the computed significance level, $p = 0.014$, is less than 0.05. Thus, the decision is to reject the null hypothesis. There is a significant difference in the two distributions between Sales 3 and 4.



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Sales4 and Sales3 equals 0.	Related-Samples Wilcoxon Signed Rank Test	.014	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure A.13 Nonparametric test of Sales 3 and 4 example

B. Linear Programming

B.1. Introduction

Linear Programming (LP) is a deterministic, multivariable, constrained, single-objective, optimization methodology. It's a model with known, deterministic, and constant parameters, and it has more than one unknown or decision variable. LP has mathematical expressions that constrain the values of the decision variables, and it seeks to solve for an optimal solution with a single objective. It is a general-purpose modeling methodology, permitting application to just about every possible problem situation that fits the assumptions the model requires. (We will discuss the assumptions of the LP model in a later section of this appendix.) Specifically, LP can be used to model problems in all the functional areas of business (accounting, economics, finance, management, and marketing) and in all types of operations (industry-wide, government, agriculture, health care, and so on).

Modeling a problem using LP is called programming. As such, LP is considered one of several mathematical programming methodologies available for use in the prescriptive step of the business analytic process.

B.2. Types of Linear Programming Problems/Models

There are basically two types of LP problems/models: a maximization model and a minimization model. Business seeks to maximize profit or sales. In such cases, the single objective is maximization. Other business situations seek to minimize costs or resource utilization. In those cases, the single objective is minimization.

In addition to these two basic types of LP models, there is a group of special case models. These models are also maximization or minimization models, but they are applied to a limited set of problems. One ex-

ample is integer programming (discussed in **Appendix D**, "**Integer Programming**"), whose model solution requires integer values rather than real number solutions.

B.3. Linear Programming Problem/Model Elements

B.3.1. Introduction

All LP problem/model formulations consist of three elements: an objective function, constraints, and nonnegativity or given requirements. The generalized model (a model without actual values, only symbols) requires the three components presented in Exhibit A. Note that the applied model is also presented in Exhibit B. Both models will be discussed in this section. The exhibit used here foreshadows the formulation of models discussed in this appendix.

A. Generalized LP Model

$$\begin{aligned}
&\text{Maximize:} && Z = c_1 X_1 + c_2 X_2 + \ldots + c_n X_n && \text{(Objective Function)}\\
&\text{subject to:} && a_{11} X_1 + a_{12} X_2 + \ldots + a_{1n} X_n \le b_1 && \text{(Constraints)}\\
& && a_{21} X_1 + a_{22} X_2 + \ldots + a_{2n} X_n \le b_2 &&\\
& && \vdots &&\\
& && a_{m1} X_1 + a_{m2} X_2 + \ldots + a_{mn} X_n \le b_m &&\\
&\text{and} && X_1, X_2, \ldots, X_n \ge 0 && \text{(Nonnegativity or Given Requirements)}
\end{aligned}$$

B. An Applied LP Model (*Ford Motor Company problem/To be explained in this Chapter*)

$$\text{Maximize: } Z = 2000 X_1 + 3500 X_2$$

$$\begin{aligned}
&\text{subject to:} && 60 X_1 + 75 X_2 \le 10000\\
& && 60 X_1 + 75 X_2 \ge 3000\\
& && X_1 + X_2 = 140\\
&\text{and} && X_1, X_2 \ge 0
\end{aligned}$$