Bringing Research to Life

"Is my being early for our meeting a problem?" asked Myra, as she slid past a pile of computer printouts stacked precariously high just inside the door to Jason's office. "Might the industrious team in your outer office be studying my MindWriter Project 2 data?"

"Give me just a second," mumbled Jason, as he quickly wrote a note on a Post-it and slapped it on a pencil sketch of a graph and jotted another to a histogram. "Sammye, you want to come get these?" Jason called to one of the team members in the outer office.

Meanwhile, Myra chose an available chair and used the time to review her notes. She was here to convince Jason to take on yet another project for MindWriter. This one had a short turnaround.

Turning his attention on Myra, Jason extracted a folder lying on the credenza behind him. "Actually those worker-bees are new members of my staff, graduate students from the university. They're assigned to the City Center for Performing Arts project," shared Jason. "It's because of your recommendation that I got the job. I thought you knew."

"Of course I knew. I've been serving on CCPA's board for two years. Will you be presenting the preliminary analysis at the next meeting, this Friday?"

"Of course not! The preliminary analysis is strictly for us. While we may develop presentation charts that might be presented to the center, it is just as likely that none of the material you see stacked here will end up in the report as-is. Didn't you learn the general timetable for such projects on the first MindWriter project? We are nowhere near ready to write the client report. We just finished cleaning the data file yesterday. This morning I ran a full set of frequencies. Jill, David, and Sammye started their preliminary analysis . . . uh, 90 minutes ago."

"Anything interesting on the initial cross-tabs?"

"Well, well. You did learn something," observed Jason with a smile.

Myra smiled in return as she raised an expressive eyebrow and waited for Jason to respond.

"Just before you got here, three of the early cross-tabs appeared to show some support for the board's assumptions about the alcohol issue—on whether current patrons endorse the selling of beer and wine during intermissions. But we're not far enough into the data to say which of your board's assumptions are fully correct and which might have to be modified based on the patterns emerging within subgroups of the sample. We'll probably have to do some recoding of the age and race variables for the patterns to emerge clearly. The team is also interested in the differences between ethnic groups in future performance preferences. We've also finished coding each patron's address with its GIS (Geographic Information Systems) code. The preliminary mapping begins tomorrow; I hired a master's candidate in geography to provide the mapping. I've scheduled a conference call for . . . (Jason flips his desk calendar pages to the following week) . . . Friday of next week to talk with Jackson Murray and other members of the CCPA project team."

"When the board approved your proposed analysis plan," shared Myra, "I don't remember seeing any reference to those cute, box-like diagrams with tails I see on that graph you just handed to Sammye."

"Most of what we'll be doing the next three days involves more graphical displays than statistical ones. Right now we're just getting a sense of what the data are telling us. We'll decide what, if any, new analyses to add to the proposed plan by this Friday. It's this early work that lays the groundwork for the more sophisticated analyses that follow. There isn't anything glamorous about it, but without it we might miss some crucial findings."

Jason paused for effect, then said, "By the way, that 'cute little diagram' is called a boxplot. I actually did several during the preliminary analysis phase for MindWriter's CompleteCare study. I didn't give them to you because I would have had to explain how to interpret them and . . ."

". . . and anything you have to explain isn't clear enough," finished Myra. "I learned Jason's Rule #1 on reporting to clients very well on MindWriter-1."

Myra modified her position in the chair, leaning slightly toward Jason. Just before she spoke, Jason observed, "Oh, no! You're changing into your 'It's time to get down to business' posture. So what's the new project you want to discuss . . . and the impossible deadline you need me to meet?"

"Just hear me out, Jason. MindWriter's LT3000 product group has decided it needs to use 'superiority in custom-designed systems' as its claim in a new ad campaign, but legal says we don't have enough data to support the claim. The ad agency we have chosen has a short window of opportunity. We need supporting data within 10 days." Myra held up her hand to stop the objection she anticipated from Jason. "We know you don't have time to collect new primary data and analyze it in 10 days . . . so I brought the next best thing. I've got three boxes of miscellaneous records in my trunk . . ."

Jason groaned. "Let's go see what you brought me. Then we'll see if this project is even feasible."

Introduction

The convenience of data entry via spreadsheet, optimal mark recognition (OMR), or with the data editor of a statistical program makes it tempting to move directly to statistical analysis. Why waste time finding out if the data confirm the hypothesis that motivated the study? Why not obtain descriptive statistical summaries (based on our discussion in Chapter 15) and then test hypotheses?

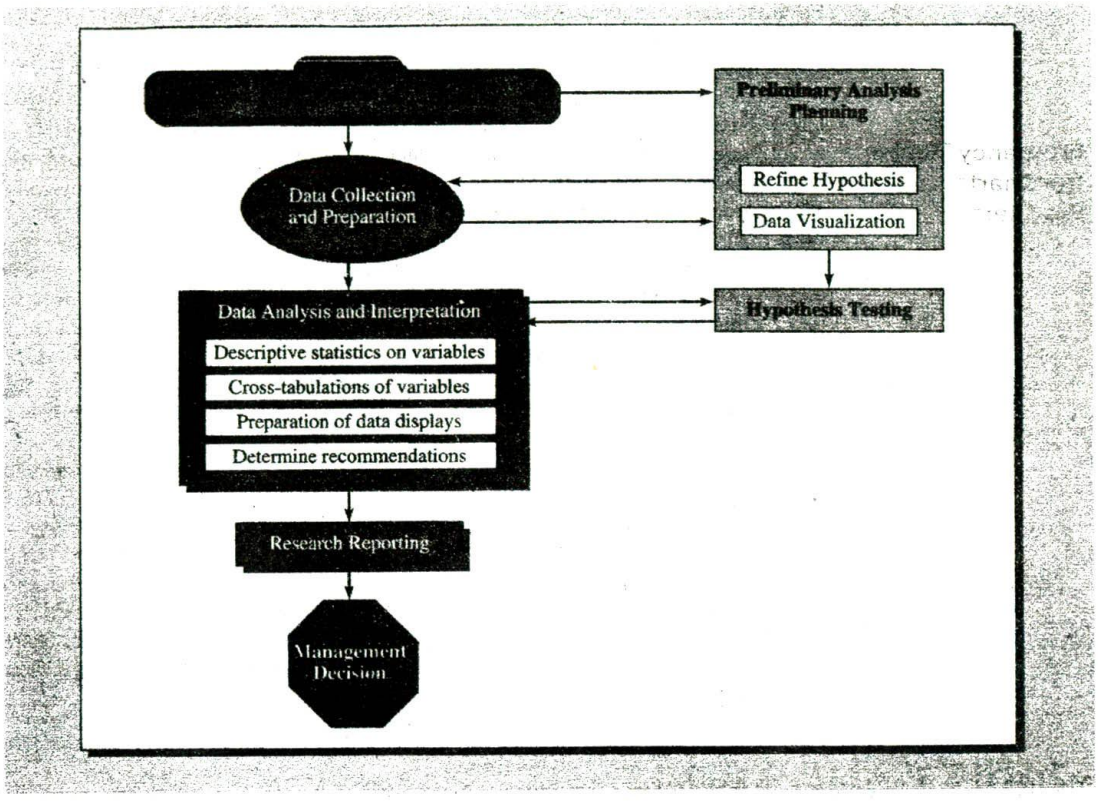
In Chapter 2, we said research conducted scientifically is a puzzle-solving activity. We also noted that an attitude of curiosity, suspicion, and imagination was essential to the discovery process. It is natural, then, that exploration of the data would be an integral part of our perspective. When the study's purpose is not the production of causal inferences, confirmatory data analysis is not required. When it is, we advocate discovering as much as possible about the data before selecting the appropriate means of confirmation. Exhibit 16–1 reminds you of the importance of data visualization as an integral element in the data analysis process and as a necessary step prior to hypothesis testing.

Depending on the type of management question, we can discover a great deal about our data through exploratory data analysis, statistical process control charting, geographical information system mapping, and cross-tabulation.

Exploratory Data Analysis

Exploratory data analysis (EDA) is both a data analysis perspective and a set of techniques.[1] In exploratory data analysis, the data guide the choice of analysis—or a revision of the planned analysis—rather than the analysis presuming to overlay its structure on the data without the benefit of the analyst's scrutiny. This is comparable to our position that research should be problem-oriented rather than tool-driven. The flexibility to respond to the patterns revealed by successive iterations in the discovery process is an

EXHIBIT 16–1 Data Analysis in the Research Process



important attribute of this approach. By comparison, **confirmatory data analysis** occupies a position closer to classical statistical inference in its use of significance and confidence. But confirmatory analysis may also differ from traditional practices by using information from a closely related data set or by validating findings through the gathering and analyzing of new data.[2]

One authority has compared exploratory data analysis to the role of police detectives and other investigators and confirmatory analysis to that of judges and the judicial system. The former are involved in the search for clues and evidence; the latter are preoccupied with evaluating the strength of what is found. Exploratory data analysis is the first step in the search for evidence, without which confirmatory analysis has nothing to evaluate.[3] Consistent with that analogy, EDA shares a commonality with exploratory designs, not formalized ones. Because it doesn't follow a rigid structure, it is free to take many paths in unraveling the mysteries in the data—to sift the unpredictable from the predictable.

A major contribution of the exploratory approach lies in the emphasis on visual representations and graphical techniques over summary statistics. Summary statistics, as you will see momentarily, may obscure, conceal, or even misrepresent the underlying structure of the data. When numerical summaries are used exclusively and accepted without visual inspection, the selection of confirmatory models may be precipitous, and

based on flawed assumptions. Consequently, it may produce erroneous conclusions.[4] For these reasons, data analysis should begin with visual inspection. After that, it is not only possible but also desirable to cycle between exploratory and confirmatory approaches.

Frequency Tables, Bar Charts, and Pie Charts[5]

Several useful techniques for displaying data are not new to EDA. They are essential to any examination of the data. For example, a **frequency table** is a simple device for arraying data. An example is presented in Exhibit 16–2. It arrays data by assigned numerical value, with columns for percent, percent adjusted for missing values, and cumulative percent. Sector, the nominal variable that describes the business classifications or markets of the sampled corporations, provides the observations for this table. Although there are 100 observations, the small number of categories makes the variable easily tabled. The same data are presented in Exhibit 16–3 using a bar chart and a pie chart. The values and percentages are more readily understood in this graphic format, and visualization of the sector categories and their relative sizes is improved.

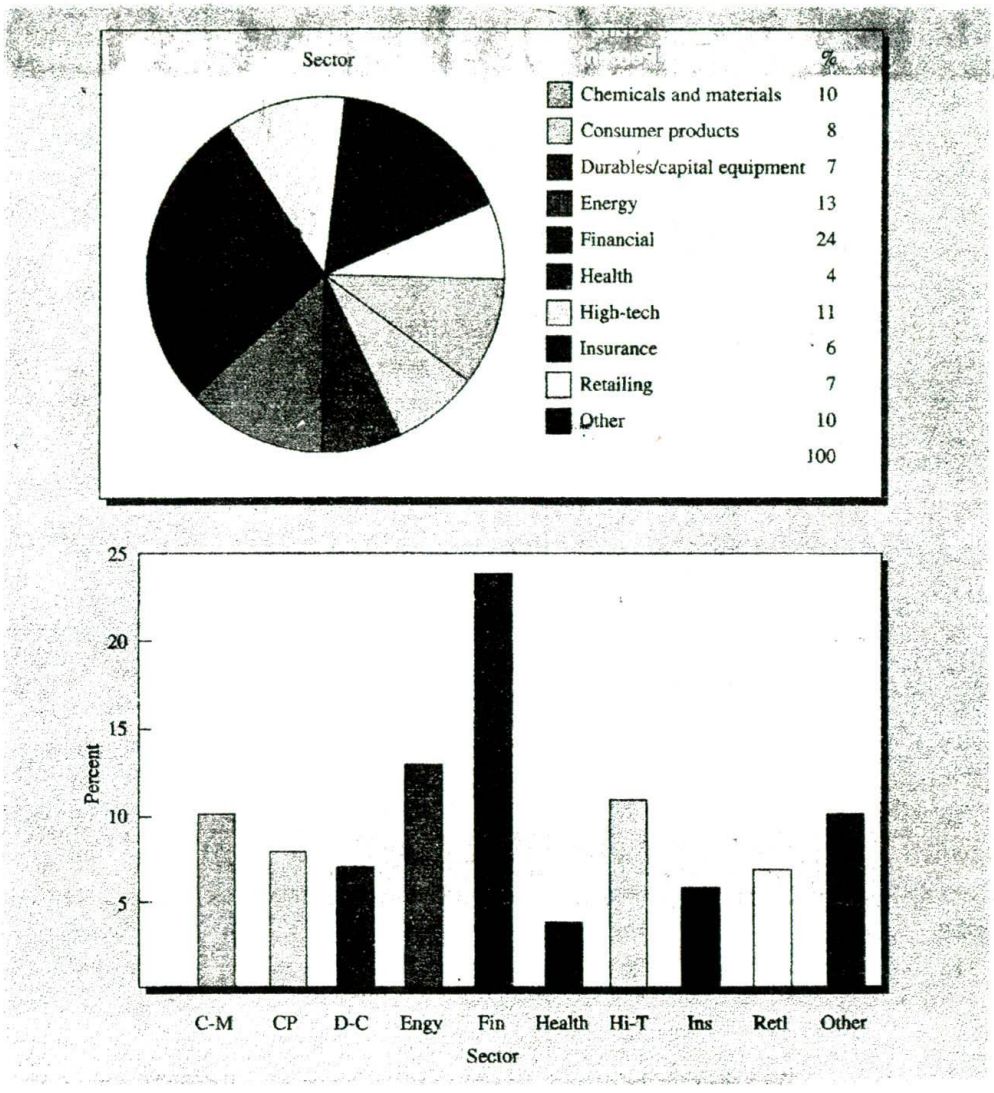
When the variable of interest is measured on an interval-ratio scale and is one with many potential values, these techniques are not particularly informative. Exhibit 16–4 is a condensed frequency table of the highest total return to investors measured in percentages of the top 48 companies in this category taken from the Fortune 500.[6] Only two values, 59.9 and 66, have a frequency greater than 1. Thus, the primary contribution of this table is an ordered list of values. If the table were converted to a bar chart, it would have 48 bars of equal length and two bars with two occurrences. And bar charts do not reserve spaces for values where no observations occur within the range. Constructing a pie chart for this variable would also be pointless.

EXHIBIT 16–2 A Frequency Table of Market Sector (Forbes Industry List)

Value Label	Value	Frequency	Percent	Valid Percent	Cumulative Percent
Chemicals and materials	1	10	10.0	10.0	10.0
Consumer products	2	8	8.0	8.0	18.0
Durables/capital equipment	3	7	7.0	7.0	25.0
Energy	4	13	13.0	13.0	38.0
Financial	5	24	24.0	24.0	62.0
Health	6	4	4.0	4.0	66.0
High-tech	7	11	11.0	11.0	77.0
Insurance	8	6	6.0	6.0	83.0
Retailing	9	7	7.0	7.0	90.0
Other	10	10	10.0	10.0	100.0
Total		100	100.0	100.0	

Valid cases 100 Missing cases 0

EXHIBIT 16–3 Nominal Variable Displays (Forbes Industry List)



Histograms

The histogram is a conventional solution for the display of interval-ratio data. **Histograms** are used when it is possible to group the variable's values into intervals. Histograms are constructed with bars (or asterisks that represent data values) where each value occupies an equal amount of area within the enclosed area. Data analysts find histograms useful for (1) displaying all intervals in a distribution, even those without observed values, and (2) examining the shape of the distribution for skewness, kurtosis, and the modal pattern. When looking at a histogram, one might ask: Is there a single hump (a mode)? Are subgroups identifiable when multiple modes are present? Are straggling data values detached from the central concentration?[7]

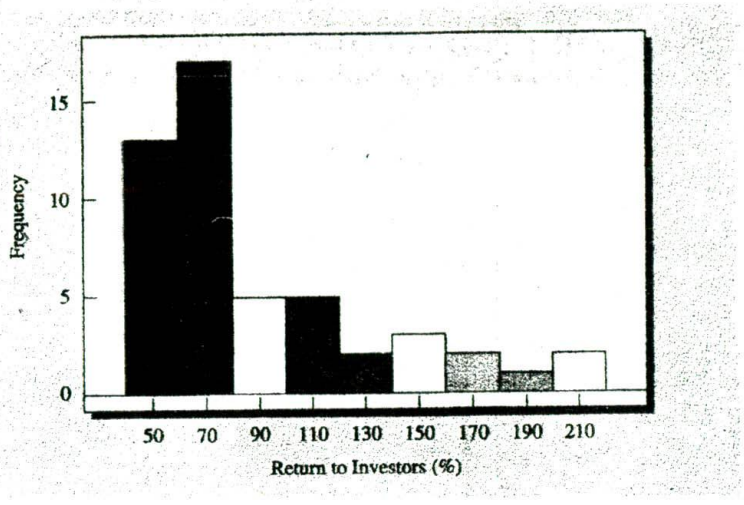
EXHIBIT 16–4 Return to Investors of the Top 48 in the Fortune 500

	Value	Frequency	Percent	Cum. Percent		Value	Frequency	Percent	Cum. Percent
1	54.9	1	2	2	25	75.6	1	2	54
2	55.4	1	2	4	26	76.4	1	2	56
3	55.6	1	2	6	27	77.5	1	2	58
4	56.4	1	2	8	28	78.9	1	2	60
5	56.8	1	2	10	29	80.9	1	2	62
6	56.9	1	2	12	30	82.2	1	2	64
7	57.8	1	2	14	31	82.5	1	2	66
8	58.1	1	2	16	32	86.4	1	2	68
9	58.2	1	2	18	33	88.3	1	2	70
10	58.3	1	2	20	34	102.5	1	2	72
11	58.5	1	2	22	35	104.1	1	2	74
12	59.9	2	4	26	36	110.4	1	2	76
13	61.5	1	2	28	37	111.9	1	2	78
14	62.6	1	2	30	38	118.6	1	2	80
15	64.8	1	2	32	39	123.8	1	2	82
16	66.0	2	4	36	40	131.2	1	2	84
17	66.3	1	2	38	41	140.9	1	2	86
18	67.6	1	2	40	42	146.2	1	2	88
19	69.1	1	2	42	43	153.2	1	2	90
20	69.2	1	2	44	44	163.2	1	2	92
21	70.5	1	2	46	45	166.7	1	2	94
22	72.7	1	2	48	46	183.2	1	2	96
23	72.9	1	2	50	47	206.9	1	2	98
24	73.5	1	2	52	48	218.2	1	2	100

The values for the return to investors variable presented in Exhibit 16–4 were measured on a ratio scale and are easily grouped. Other variables possessing an underlying order are similarly appropriate for histograms. A histogram would not be used for a nominal variable like sector (Exhibit 16–3) that has no order to its categories.

A histogram of the return to investors variable taken from the Fortune 500 ranked by performance listing is shown in Exhibit 16–5. The midpoint for each interval for the variable of interest, return to investors, is shown on the horizontal axis, the frequency or number of observations in each interval on the vertical axis. We erect a vertical bar above the midpoint of each interval on the horizontal scale. The height of the bar corresponds with the frequency of observations in the interval above which it is erected. This histogram was constructed with intervals 20 increments wide, and the last interval con-

EXHIBIT 16–5 Histogram of Fortune 500 Data



tains only two observations, 206.9 and 218.2. These values are found in the Fortune 500 frequency table (Exhibit 16–4). Intervals with 0 counts show gaps in the data and alert the analyst to look for problems with spread. When the upper tail of the distribution is compared with the frequency table, we find three extreme values (183.2, 206.9, and 218.2). Along with the peaked midpoint and reduced number of observations in the upper tail, this histogram warns us of irregularities in the data.

Stem-and-Leaf Displays[8]

The **stem-and-leaf display** is an EDA technique that is closely related to the histogram. It shares some of its features but offers several unique advantages. It is easy to construct by hand for small samples or may be produced by computer programs.

In contrast to histograms, which lose information by grouping data values into intervals, the stem-and-leaf presents actual data values that can be inspected directly without the use of enclosed bars or asterisks as the representation medium. This feature reveals the distribution of values within the interval and preserves their rank order for finding the median, quartiles, and other summary statistics. It also eases linking a specific observation back to the data file and to the subject that produced it.

Visualization is the second advantage of stem-and-leaf displays. The range of values is apparent at a glance, and both shape and spread impressions are immediate. Patterns in the data—such as gaps where no values exist, areas where values are clustered, or outlying values that differ from the main body of the data—are easily observed.

In order to develop a stem-and-leaf display for the data in Exhibit 16–4, the first digits of each data item are arranged to the left of a vertical line. Next, we pass through the return to investor percentages in the order they were recorded and place the last digit for each item (the unit position, 1.0) to the right of the vertical line. Note that the digit to the right of the decimal point is ignored. The last digit for each item is placed on the horizontal row corresponding to its first digit(s). Now it is a simple matter to rank-order the digits in each row, creating the stem-and-leaf display shown in Exhibit 16–6.

EXHIBIT 16–6 A Stem-and-Leaf Display of Fortune 500 Data

5	4 5 5 6 6 6 7 8 8 8 8 9
6	1 2 4 6 6 7 9 9
7	0 2 2 3 5 6 7 8
8	0 2 2 6 8
9	
10	2 4
11	0 1 8
12	3
13	1
14	0 6
15	3
16	3 6
17	
18	3
19	
20	6
21	8

Each line or row in this display is referred to as a *stem,* and each piece of information on the stem is called a *leaf.* The first line or row is

5 | 4 5 5 6 6 6 7 8 8 8 8 9

The meaning attached to this line or row is that there are 12 items in the data set whose first digit is five: 54, 55, 55, 56, 56, 56, 57, 58, 58, 58, 58, and 59. The second line,

6 | 1 2 4 6 6 7 9 9

shows that there are eight return to investors percentage values whose first digit is six: 61, 62, 64, 66, 66, 67, 69, and 69. The stem is the digit(s) to the left of the vertical line (6 for this example), and the leaf is the digit(s) to the right of the vertical line (1, 2, 4, 6, 6, 7, 9, 9).

When the stem-and-leaf display shown in Exhibit 16–6 is turned upright (rotated 90 degrees to the left), the shape is the same as that of the histogram shown in Exhibit 16–5.

Boxplots[9]

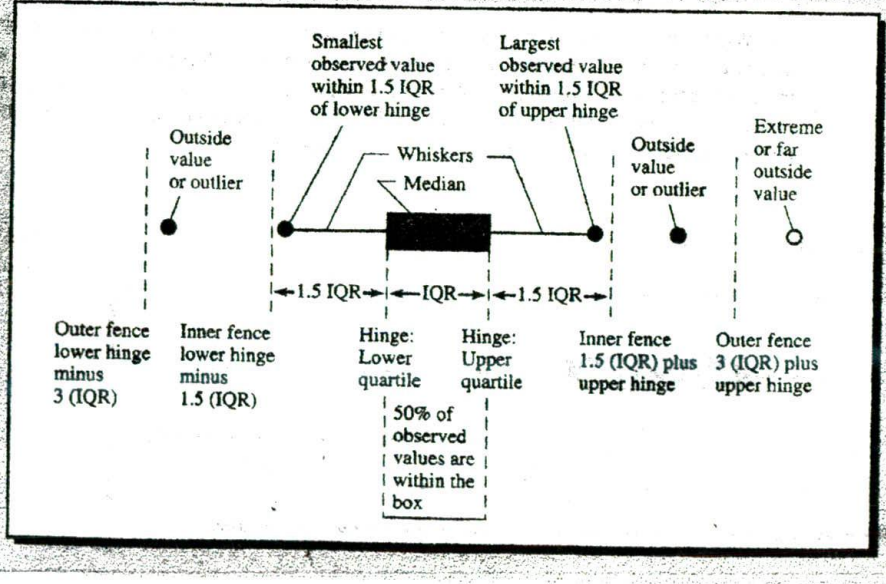
The **boxplot,** or *box-and-whisker plot,* is another technique used frequently in exploratory data analysis.[10] A boxplot reduces the detail of the stem-and-leaf display and provides a different visual image of the distribution's location, spread, shape, tail length, and outliers. Boxplots are extensions of the **five-number summary** of a distribution. This summary consists of the median, upper and lower quartiles, and the largest and smallest observations. The median and quartiles are used because they are particularly **resistant statistics.** *Resistance* is a characteristic that "provides insensitivity to localized misbehavior in data."[11] Resistant statistics are unaffected by out-

liers and change only slightly in response to the replacement of small portions of the data set.

Recall the previous discussion of the mean and standard deviation in Chapter 15. Now assume we take the data set [5,6,6,7,7,7,8,8,9]. The mean of the set is 7, the standard deviation 1.23. If the 9 is replaced with 90, the mean becomes 16 and the standard deviation increases to 27.78. The mean is now two times larger than most of the numbers in the distribution, and the standard deviation is more than 22 times its original size. Changing only one of nine values has disturbed the location and spread summaries to the point where they no longer represent the other eight values. Both the mean and the standard deviation are considered **nonresistant statistics;** they are susceptible to the effects of extreme values in the tails of the distribution and do not represent typical values well under conditions of asymmetry. The standard deviation is particularly problematic because it is computed from the squared deviations from the mean.[12] In contrast, the median and quartiles are highly resistant to change. When we changed the 9 to 90, the median remained at 7 and the lower and upper quartiles stayed at 6 and 8, respectively. Because of the nature of quartiles, up to 25 percent of the data can be made extreme without perturbing the median, the rectangular composition of the plot, or the quartiles themselves. These characteristics of resistance are incorporated into the construction of boxplots.

Boxplots may be constructed easily by hand or by computer programs. The basic ingredients of the plot are the (1) rectangular plot that encompasses 50 percent of the data values, (2) a center line (or other notation) marking the median and going through the width of the box, (3) the edges of the box, called hinges, and (4) the whiskers that extend from the right and left hinges to the largest and smallest values.[13] These values may be found within 1.5 times the **interquartile range (IQR)** from either edge of the box. These components and their relationships are shown in Exhibit 16–7.

EXHIBIT 16–7 Boxplot Components



We can create a boxplot of the return to investors variable from the information provided in Exhibit 16–4. With the five-number summary, we have the basis for a skeletal plot.

Minimum	Lower Hinge	Median	Upper Hinge	Maximum
54.90	59.90	73.20	110.78	218.20

The plot shown in Exhibit 16–8 started with these data and the following calculations. You may construct your own boxplot with the data in Exhibit 16–4 and the SPSS Explore procedure. Beginning with the box, the ends are drawn using the lower and upper quartile (hinge) data. The median is drawn in at 73.2. Then the IQR is calculated (110.78 – 59.9 = 50.88). From this we can locate the lower and upper fences. The fences are –25.44 and 187.1. Next, the smallest and largest data values from the distribution within the fences are used to determine the whisker length. These values are 54.90 and 183.20. We are now able to see the outliers in relation to the "main body" of the data. **Outliers** are data points that exceed +1.5 IQRs of a boxplot's hinges. Data values for the outliers are added, and identifiers may be provided for interesting values. The completed boxplot is shown in Exhibit 16–8.

		Fence	
IQR	Distance to	(–)	(+)
50.88	*(±1.5) = 76.32	50.88 – 76.32 = –25.44	110.78 + 76.32 = 187.1

When examining data, it is important to separate legitimate outliers from errors in measurement, editing, coding, and data entry. Outliers that reflect unusual cases are an important source of information for the study. They are displayed or given special statistical treatment, or other portions of the data set are sometimes shielded from their effects. Outliers that are mistakes should be corrected or removed.

EXHIBIT 16–8 Boxplot of Fortune 500 Data

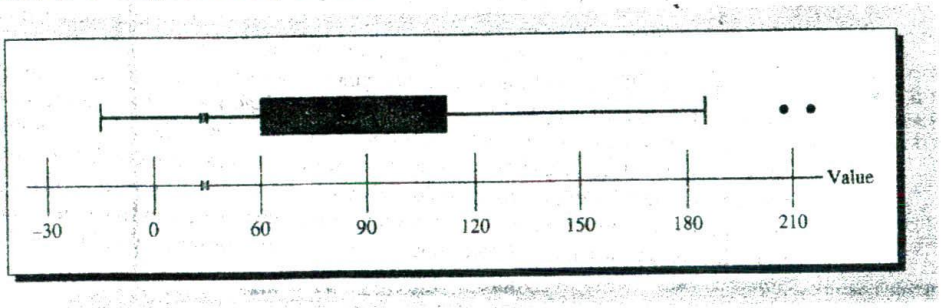


EXHIBIT 16–9 Diagnostics with Boxplots

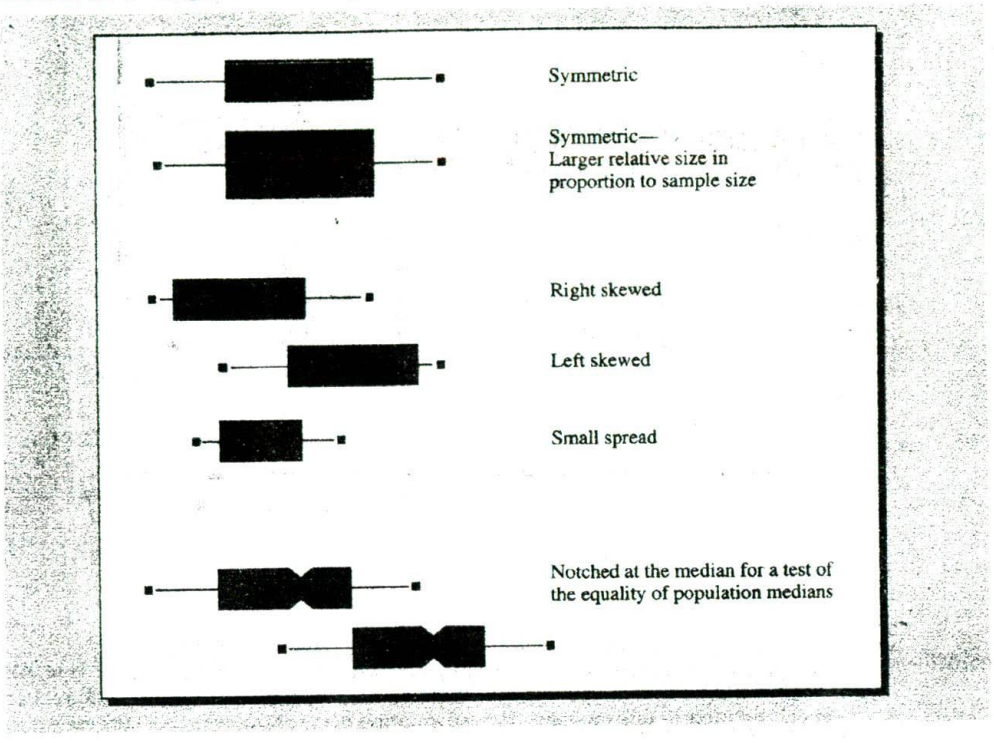
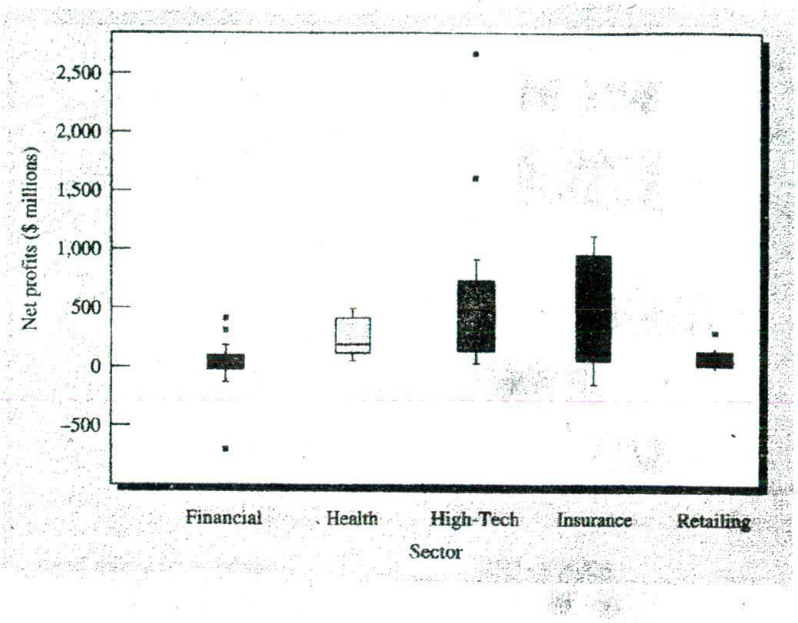


Exhibit 16–9 summarizes several comparisons that are of help to the analyst. Boxplots are an excellent diagnostic tool, especially when graphed on the same scale. The upper two plots in the exhibit are both symmetric, but one is larger than the other. Larger box widths are sometimes used when the second variable, from the same measurement scale, comes from a larger sample size. The box widths should be proportional to the square root of the sample size, but not all plotting programs account for this.[14] Right- and left-skewed distributions and those with reduced spread are also presented clearly in the plot comparison. Finally, groups may be compared by means of multiple plots. One variation, in which a notch at the median marks off a confidence interval to test the equality of group medians, takes us a step closer to hypothesis testing.[15] Here the sides of the box return to full width at the upper and lower confidence intervals. When the intervals do not overlap, we can be confident, at a specified confidence level, that the medians of the two populations are different.

In Exhibit 16–10, multiple boxplots compare five sectors on the return to investors variable. The overall impression is one of potential problems for the analyst: unequal variances, skewness, and extreme outliers. Note the similarities of the profiles of finance and retailing in contrast to the high-tech and insurance sectors. If hypothesis tests are planned, further examination of this plot for each sector would require a stem-and-leaf display and a five-number summary. From this, we could make decisions on test selection and whether the data should be transformed or reexpressed before further analysis.

EXHIBIT 16–10 Boxplot Comparison of Sectors



Transformation[16]

Some of the examples in this section have departed from normality. While this makes for good illustrations, such data pose special problems in data analysis. Transformation is one solution to this problem. **Transformation** is the reexpression of data on a new scale using a single mathematical function for each data point. Although nominal and ordinal data may be transformed, the procedures are beyond the scope of this book. We will consider only interval-ratio scale transformations here.

The fact that data collected on one scale are found to depart from the assumptions of normality and constant variance does not preclude reexpressing them on another scale. What is discovered, of course, must be linked to the original data.

We transform data for several reasons: (1) to improve interpretation and compatibility with other data sets, (2) to enhance symmetry and stabilize spread, and (3) to improve linear relationships between and among variables. We improve interpretation when we find alternate ways to understand the data and discover patterns or relationships that may not have been revealed on the original scales. A **standard score,** or *Z score,* may be calculated to improve compatibility among variables that come from different scales and require comparison. Z scores convey distance in standard deviation units with a mean of 0 and a standard deviation of 1. This is accomplished by converting the raw score, X_i, to

$$Z = \frac{X_i - \bar{X}}{s}$$

Z scores improve interpretation through their reference to the normal curve and our understanding of the areas under it.

Conversion of centimeters to inches, stones to pounds, liters to gallons, or Celsius to Fahrenheit are examples of linear conversions that change the scale but do not change symmetry or spread. Many statisticians consider these data as manipulations rather than transformations.

Nonlinear transformations are often needed to satisfy the other two reasons for reexpressing data. Normality and constancy of variance are important assumptions for many parametric statistical techniques. A transformation to reduce skewness and stabilize variance makes it possible to use various confirmatory techniques without violating their assumptions. Analysis of the relationship between variables also benefits from transformation. Improved predictions and better diagnostics of fit and residuals (as in regression analysis) are frequent payoffs.

Transformations are defined with power, p, as the reexpression of x with x^p.[17] Exhibit 16–11 shows the most frequently used power transformations.

We use *spread-and-level* plots to guide our choice of a power transformation. By plotting the log of the median against the log of the interquartile range, we can find the slope of the plot: where p, the power we are seeking, is equal to 1 – slope. Although 1/4 and 1/3 powers often result—and are sometimes preferred—many computer programs require rounding the transformation to the nearest half power.

The return to investors variable is used to illustrate this concept. The data distribution shows a right skew. The five-number summary (data in millions of dollars) reveals an extreme score as the maximum data point:

Minimum	Lower Hinge	Median	Upper Hinge	Maximum
54.9	59.9	73.2	110.78	218.2

The largest observation, 218.2, is only approximately 0.5 IQRs beyond the main body of data, and there is only one other value beyond the fence.

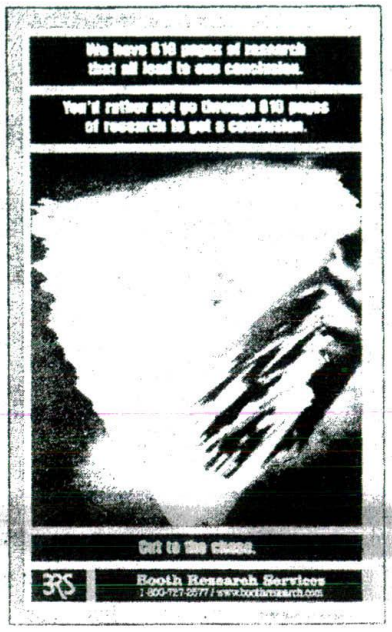
A quick calculation of the ratio of the largest observation to the smallest (218.2/54.9 = 3.97) serves as the final confirmation that transformation might not be worthwhile. It is desirable for this informal index to be greater than 20; with ratios less than 2, transformation is not practical.[18] From this information, we might conclude that the return to investors variable is not a good candidate for transformation.

When researchers communicate their findings to management, the advantages of reexpression must be balanced against pragmatism: Some transformed scales have no

EXHIBIT 16–11 Frequently Used Power Transformations

Power	Transformation
3	Cube
2	Square
1	No change: existing data
1/2	Square root
0	Logarithm (usually Lg_{10})
–1/2	Reciprocal root
–1	Reciprocal
–2	Reciprocal square
–3	Reciprocal cube

As this Booth Research Services ad suggests, the researcher's purpose is to make sense of numerous data displays and thus assist the research sponsor in making an appropriate management decision. www.boothresearch.com



familiar analogies. Logarithmic dollars can be explained, but how about reciprocal root dollars? Attitude and preference scales might be better understood transformed, but the question of interpretation remains.

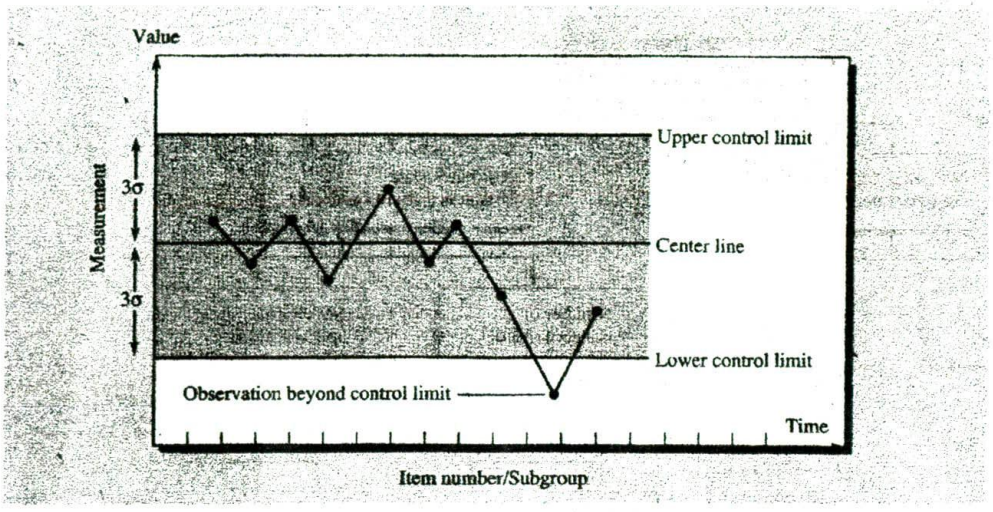
Throughout this section we have exploited the visual techniques of exploratory data analysis to look beyond numerical summaries and gain insight into the behavior of the data. Few of the approaches have stressed the need for advanced mathematics, and all have an intuitive appeal for the analyst. When the more common ways of summarizing location, spread, and shape have conveyed an inadequate picture of the data, we have used more resistant statistics to protect us from the effects of extreme scores and occasional errors. We have also emphasized the value of transforming the original scale of the data during preliminary analysis rather than at the point of hypothesis testing.

Improvement and Control Analysis

Statistical process control (SPC) uses statistical tools to analyze, monitor, and improve process performance. A process is any business system that transforms inputs to outputs. Assembling automobiles, manufacturing plastic pipe, preparing restaurant meals, and service and administrative business activities are all made up of processes. Processes that operate within control limits are stable, predictable, and responsive to improvement. Walter Shewhart at Bell Laboratories defined this concept for production processes during the 1920s; it was later expanded by W. Edwards Deming to become an integral part of **Total Quality Management (TQM).**

The element of the process control system that pertains to displaying data is its charting system. SPC charts provide easily understood and reliable visuals for evaluating point values and trends and can graphically depict variation. Control charts help

EXHIBIT 16–12 Control Chart Characteristics



managers focus on special causes of variation in a process by revealing whether a system is under control (as an early warning of change) and substantiate results from improvements (confirming results).

A **control chart** displays sequential measurements of a process together with a center line and control limits. Limit lines provide guides for evaluating performance. These lines show the dispersion of data within a statistical boundary (generally three standard deviations above and below the average or center line). If an observation falls beyond the area marked by the **upper control limit (UCL)** or the **lower control limit (LCL)**, there is evidence that the process is out of control or special causes are adversely affecting it. However, even if all data points are within the boundaries, the time plot (the sequence of points along the X-axis) can reveal that the process has sudden jumps, drifts, cycles, or the beginning of a trend. The elements that make up a control chart are shown in Exhibit 16–12.

Types of Control Charts

The selection of a control chart depends on the level of data you are measuring. Exhibit 16–13 provides a diagram for selecting an appropriate chart. In this exhibit, *variables data* refer to ratio or interval measurements (such as the diameter of a bearing, units/hour, the thickness of a plastic film, temperature, or blood pressure). Variables data are typically presented in X-bar charts and R-charts and X-bar charts and s-charts. Attributes data are nominal or categorical data that can be counted. Examples include the number of defects/part, abnormals/1,000 processed, and scratches per item. Attributes data are usually nonconforming units shown in control charts called *c-*, *u-*, *p-*, and *np-charts*. Variables data have the advantage because they reveal how close we are to a specification—just as interval data provide more information on attitudes than do nominal data.

X-bar charts and R-charts are often displayed together, as illustrated in Exhibit 16–15. The measurements describe an outboard motor manufacturer whose motors for small fishing boats require control to deliver a constant 10 horsepower. The process characteristics are reported in Exhibit 16–14 showing small subgroups of three observations per subgroup. (This type of control chart usually has constant size samples of

EXHIBIT 16–13 Control Chart Selection

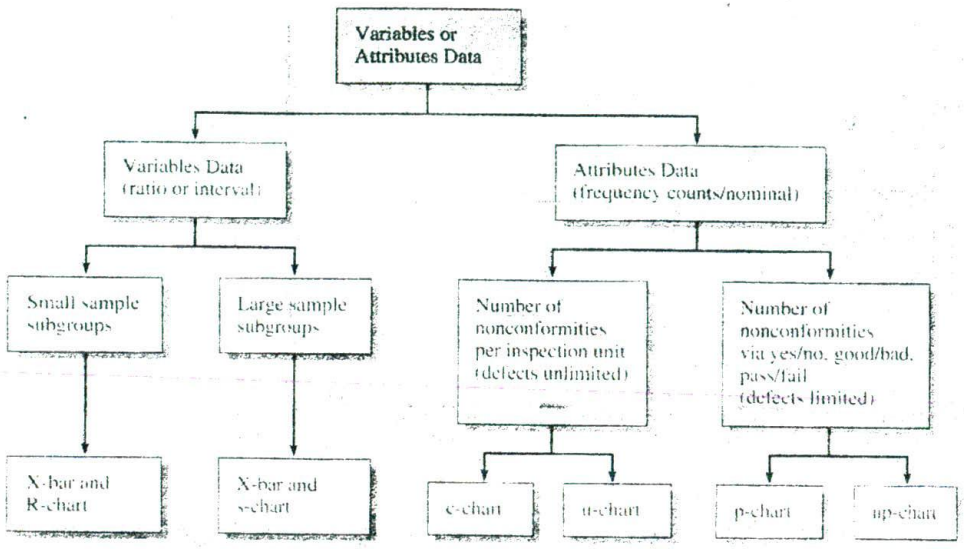


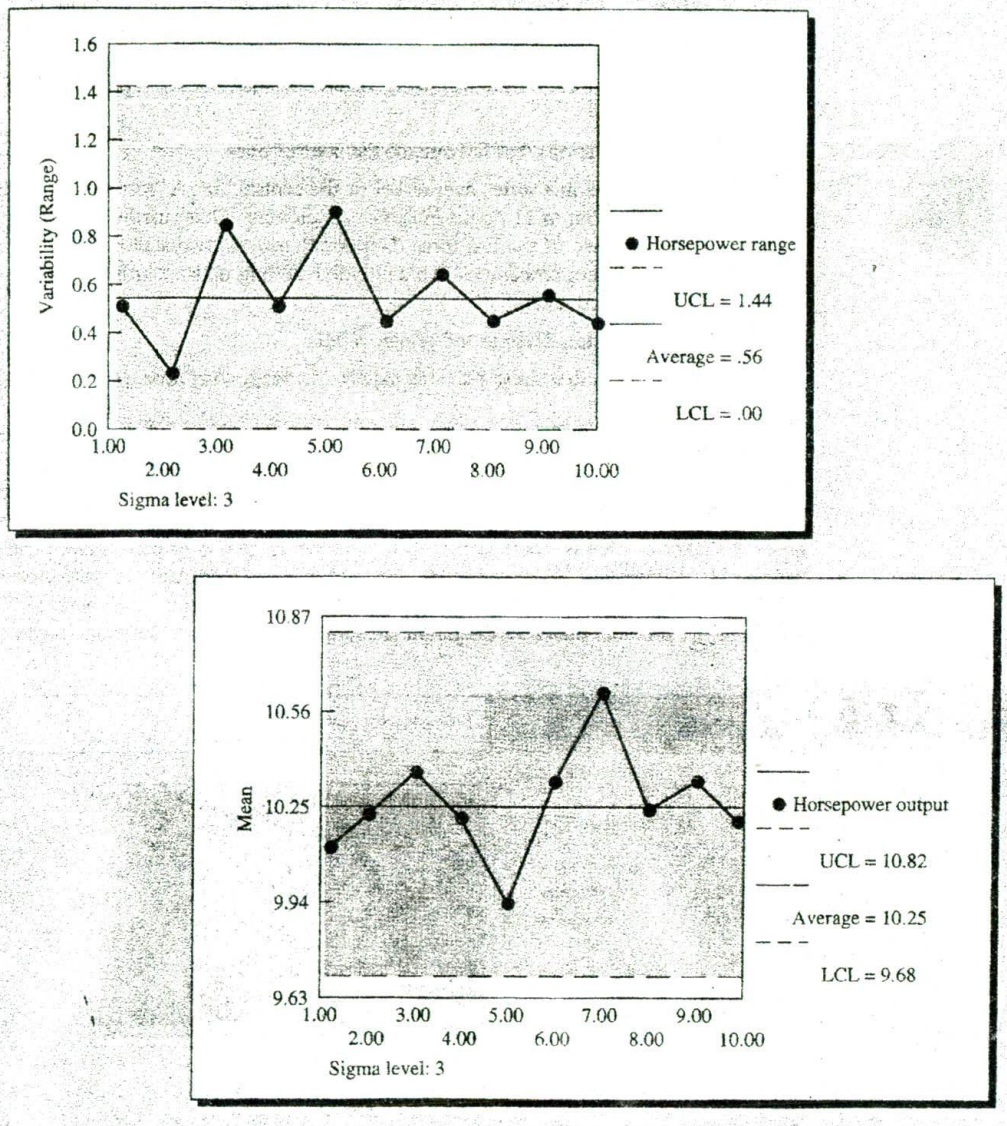
EXHIBIT 16–14 MerK Outboard Motor Data

Subgroup	Horsepower Output		
1	10.22	10.31	9.79
2	10.25	10.33	10.12
3	10.37	9.92	10.80
4	10.46	9.94	10.26
5	10.06	9.39	10.31
6	10.59	10.15	10.23
7	10.82	10.85	10.20
8	10.52	10.14	10.07
9	10.13	10.69	10.15
10	9.88	10.32	10.31

two to five measurements per subgroup.) We first evaluate the variance, as represented by the range, since it is important to keep the variability of the process low. Although the R-chart in Exhibit 16–15 suggests the process is technically in control, we can see fluctuations throughout the range.

There are no assignable causes for these discrepancies, so we proceed to analyze the X-bar chart in Exhibit 16–15. Here it is apparent that subgroups 5 and 7 under- and overproduce, respectively. One inference is that we should monitor production and secure further samples.

EXHIBIT 16–15 R-Chart and X-Bar Chart for MerK Outboard Motor Data



X-bar charts and s-charts (sample standard deviation) are also used jointly. The sample standard deviation is a more efficient indicator of process variability, especially with larger sample sizes. However, it is less sensitive in detecting special causes of variation when a single value in a subgroup is found to be unusual.

Other varieties of control charts have a layout similar to the charts in Exhibit 16–15 but differ as to purpose. The c-chart and p-chart are used primarily for attributes data.

P-charts ("proportion" charts) show the percentage of production that is defective. The light bulb illuminates or it doesn't; the aspirin bottle cap is sealed or it isn't; the diskette is preformatted or it isn't. C-charts (occasionally called "count" charts) add up defects per unit over consecutive periods of time: The stapler has two defects per unit, or the television has one defect per submodule.

Managers should look for the following visual characteristics in reports containing control charts:

MANAGEMENT Tip

1. **Outliers:** Observations that fall outside the control lines.
2. **Runs:** Data points in a series over or below the central line. A "run" of 7 consecutive points or 10 out of 11 points indicates an anomaly. Other methods reveal runs such as finding two of the last three data points beyond two standard deviations, and cumulative sum procedures, which involve adding up standardized deviations from the calculated mean.
3. **Trends:** The continual rise or fall of data points.
4. **Periodicity:** Data that show the same pattern of change over time, creating a cycle.

Pareto Diagram Pareto diagrams derive their name from a 19th-century Italian economist. In quality management, J. M. Juran first applied this concept to the industrial environment. He noted that only a vital few defects account for most problems evaluated for quality, and that the trivial many explain the rest. Historically, this has come to be known as the 80/20 rule—that is, an 80 percent improvement in quality or performance can be expected by eliminating 20 percent of the causes of unacceptable quality or performance.

The **Pareto diagram** is a bar chart whose percentages sum to 100 percent. The causes of the problem under investigation are sorted in decreasing importance with bar

SNAPSHOT

Does a Dummy Ever Lie?

Recently Ford Motor Co. has been much in the safety news, first with roll-overs of its Explorer SUV, not to mention the dissolution of its century-old alliance with Firestone. Until recently, Ford Motor Co. thought its star performer—the Ford F-150 truck—was not only popular but also safe. In laboratory offset crash tests, however, the Insurance Institute for Highway Safety concluded otherwise. It gave the Ford F-150 a "poor" rating after it performed the worst of four vehicles tested. The crash test was performed on one extended-cab vehicle of similar weight from each of four manufacturers (GM, Toyota, DaimlerChrysler, and Ford). The trucks were purchased directly from a local dealer. The crash test simulated a 40-mile-per-hour head-on collision, with each truck slamming into a stationary barrier with the left half of the front bumper. Only in the Toyota Motor Company's full-size Tundra truck did the safety cage maintain its integrity, protecting the driver "dummy" from all but minor injuries. In the Ford, the driver dummy flew uncontrolled within the cab before ending in a position beneath the steering column. Even the dummy couldn't survive the head trauma inflicted by the twisted metal carnage that had been the Ford F-150's occupant compartment. Ford said that the crash test was just one of many

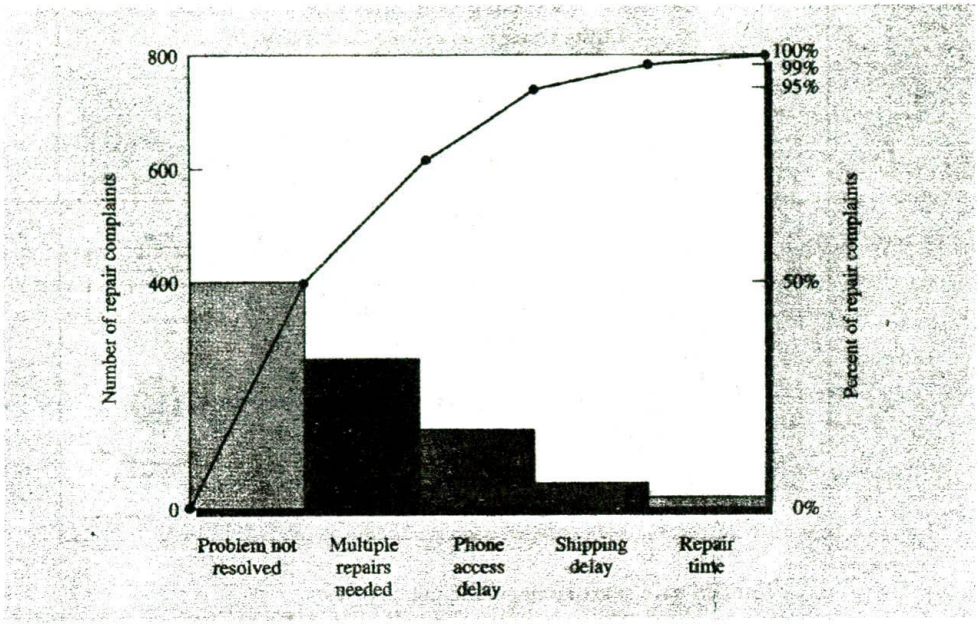
conducted on its vehicles and that, overall, the Ford F-150 was safe. For example, it showed little damage in one such test where the Ford F-150 slammed full-bumper into a stationary barrier. How would you assess the quality of the institute's experiment and its resulting evaluation of Ford?

www.hwysafety.org

www.toyota.com

www.ford.com

EXHIBIT 16–16 Pareto Diagram of MindWriter Repair Complaints



height descending from left to right. The pictorial array that results reveals the highest concentration of quality improvement potential in the fewest number of remedies. An analysis of MindWriter customer complaints is depicted as a Pareto diagram in Exhibit 16–16. The cumulative frequency line in this exhibit shows that the top two problems (the repair did not resolve the customers' problem, and the product was returned multiple times for repair) accounted for 80 percent of the perceptions of inadequate repair service.

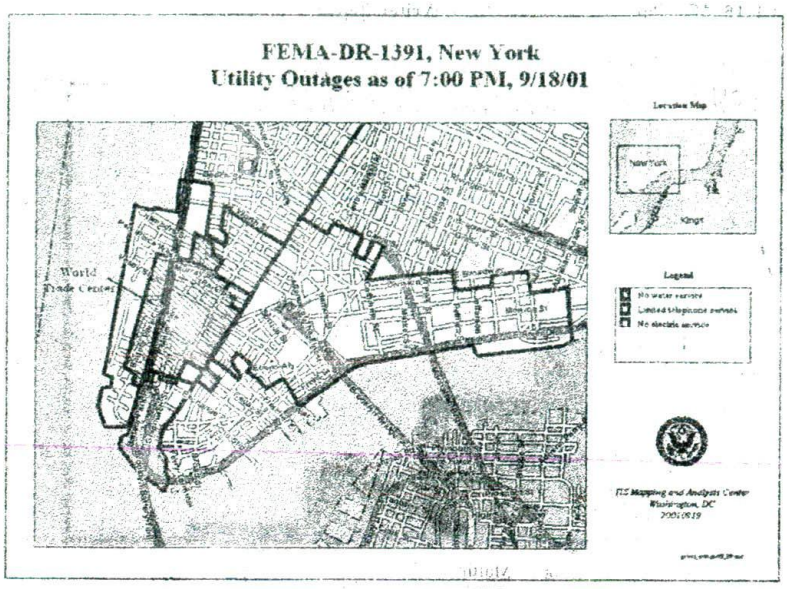
Geographic Information Systems

See an example of a GIS display in Chapter 20.

Geographic Information Systems (GIS) refers to systems of hardware and software and procedures that capture, store, manipulate, integrate, and display spatially referenced data for solving complex planning and management problems.[19]

GIS works by stacking different data sets on top of each other so that the data points from variables of interest align on a common geographical referent, allowing the user to drill through the layers and visualize the relationships among different sets of information. Its software does more, however, than draw maps on top of each other. The database system and computational buffers have specific capabilities for dealing with spatially referenced information as well as the algorithms necessary to analyze simultaneously the many layers of data.

There are numerous management applications for GIS mapping and analysis. A market analyst might seek information on sales targeting, health care or hospital network placement, real estate site selection, or customer locations. Public managers use GIS to answer questions about incorporation boundaries, emergency service delivery, school districts, redevelopment plans, natural resource preservation, and the service requirements of growing populations.



The Federal Emergency Management Agency (FEMA) uses its Mapping and Analysis Center to provide national level GIS support for the agency's mission in responding to disasters. Here we see power outages in the vicinity of New York's World Trade Center on September 18, 2001. www.gismaps.fema.gov

Most GIS have at least four process components:

- Integrating information from various sources.
- Capturing data.
- Projection and restructuring.
- Modeling.

Integrating Information

The primary requirement for GIS data is that a variable's location is known. A location may be a zip code, census boundary, survey marker, highway marker, or coordinates of longitude, latitude, and elevation. Data appropriate for GIS analysis are those that can be located spatially. By integrating variables from different sources, combinations of mapped variables may be created to simulate and analyze new ones. For example, a GIS mapping program and water company billing information could be used to simulate the impact of discharge from neighborhood septic systems on an environmentally sensitive wetland downstream.[20]

With existing information, census or zip code tabular data can be mapped to form layers of thematic information. In health care, we may want to know if the services provided match the needs of a defined geographical population. The resulting map could answer questions like: Where are certain services in higher demand? What is the best location for a new facility? Is public transportation adequate for patient access?[21]

Capturing Data

The most labor-intensive component of GIS is data capture and editing. All map objects must be specified and their spatial relationships entered. Variables related to each object are indexed. In Chapter 15 we said that editing detects errors and omissions, correcting them when possible, so quality standards may be achieved. Scanners record map imperfections as precisely as correct features. Contour lines that should not be connected or boundaries that do not exist produce false information in the spatial analysis that mislead the analyst during drill-down through the layers.

SNAPSHOT

Saving Lives with Research

Emergency services are the lifeline for millions of people every day. So when Ohio's Springfield Fire Department (SFD) noted a disturbing trend in 1997 response-time data, it needed a potentially life-saving solution. Target response time for SFD is less than eight minutes for fire emergencies and less than six minutes for medical emergencies in this community of 79,000 people. Among several options, SFD wanted to identify the possible location for a new fire station. So it turned to Dr. Olga Medvedkov, a GIS expert.

Medvedkov applied *MapInfo* and *ArcView* software to the fire department's extensive database of 911 calls. "Residents generate between 16,000 and 17,000 calls per year to our emergency system," shared Medvedkov. "Approximately 90 percent of these calls are for medical emergencies, with the remaining 10 percent for fire." Medvedkov's team geo-coded each 911 call, applying a geographic tag that described its U.S. Census block location. Then each call was plotted to a Census-block map of the city. "With its sparser population, high number of nursing homes, and older population, response times were often lengthier in the northern part of the city," explained Medvedkov. "The department was considering a new station in that location." But after seeing the spatial data display, SFD determined the sectors served by Station 8, an older and poorer part of the city, generated a more significant concentration of calls with unacceptable response times. This area was really the better location choice to reduce the largest number of calls to within the target response time."

Projection

Projection is a mathematical process for converting a three-dimensional curved surface to a two-dimensional medium such as paper or a computer screen. Since different types of maps are used, a computer program must restructure information collected from sources with different projections to a common one. Satellite images, for example, are often used for rural land use or agricultural planning but since digital data of the earth's surface are collected and stored differently, the data sources must be made compatible for mapping.

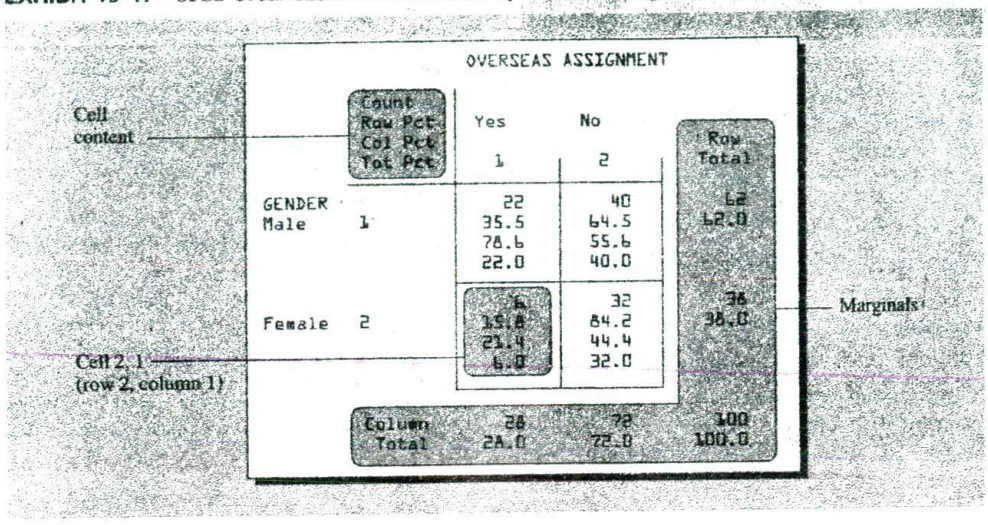
Modeling

A significant benefit of GIS is modeling. Therefore, the possible uses of the data exceed maplike outputs (roads, buildings, contours). The researcher might select two locations and ask the program to calculate the best route between them. Then factors like adjacency (what is next to what), proximity (how close one thing is to another), and containment (what is enclosed by what) come into play. Our "best route" question now extracts traffic density, construction, known bottlenecks, and hazardous intersections from different data sets to provide the best response time for an emergency vehicle.

Cross-Tabulation

Cross-tabulation is a technique for comparing two classification variables, such as gender and selection by one's company for an overseas assignment. The technique uses tables having rows and columns that correspond to the levels or values of each variable's categories. Exhibit 16–17 is an example of a computer-generated cross-tabulation. This table has two rows for gender and two columns for assignment selection. The combination of the variables with their values produces four **cells.** Each cell contains a count of the cases of the joint classification and also the row, column, and total percentages. The number of row cells and column cells is often used to designate the size of the table, as in this 2×2 table. The cells are individually identified by their row and column numbers, as illustrated. Row and column totals, called **marginals,**

EXHIBIT 16–17 SPSS Cross-Tabulation of Gender by Overseas Assignment Opportunity



appear at the bottom and right "margins" of the table. They show the counts and percentages of the separate rows and columns.

When tables are constructed for statistical testing, we call them **contingency tables,** and the test determines if the classification variables are independent. Of course, tables may be larger than 2×2.

The Use of Percentages

Percentages serve two purposes in data presentation. First, they simplify the data by reducing all numbers to a range from 0 to 100. Second, they translate the data into standard form, with a base of 100, for relative comparisons. In a sampling situation, the number of cases that fall into a category is meaningless unless it is related to some base. A count of 28 overseas assignees has little meaning unless we know it is from a sample of 100. Using the latter as a base, we conclude that 28 percent of this study's sample has an overseas assignment.

While the above is useful, it is even more useful when the research problem calls for a comparison of several distributions of data. Assume the previously reported data were collected five years ago and the present study had a sample of 1,500, of which 360 were selected for overseas assignments. By using percentages, we can see the relative relationships and shifts in the data (see Exhibit 16–18).

With two-dimension tables, the selection of a row or column will accentuate a particular distribution or comparison. This raises the question about which direction the percentages should be calculated. Most computer programs offer options for presenting percentages in both directions and interchanging the rows and columns of the table. But in situations when one variable is hypothesized to be the presumed cause, is thought to affect or predict a response, or is simply antecedent to the other variable, we label it the *independent* variable. Percentages should then be computed in the direction of this variable. Thus, if the independent variable is placed on the row, select row percentages; if it is on the column, select column percentages. In which direction should the percentages run in the previous example? If only the column percentages are

EXHIBIT 16–18 Comparison of Percentages in Cross-Tabulation Studies of Gender by Overseas Assignment

Study 1

		OVERSEAS ASSIGNMENT		
	Count Row Pct Col Pct Tot Pct	Yes 1	No 2	Row Total
GENDER Male	1	22 35.5 78.6 22.0	40 64.5 55.6 40.0	62 62.0
Female	2	6 15.8 21.4 6.0	32 84.2 44.4 32.0	38 38.0
	Column Total	28 28.0	72 72.0	100 100.0

Study 2

		OVERSEAS ASSIGNMENT		
	Count Row Pct Col Pct Tot Pct	Yes 1	No 2	Row Total
GENDER Male	1	225 25.0 62.5 15.0	675 75.0 59.2 45.0	900 60.0
Female	2	135 22.5 37.5 9.0	465 77.5 40.8 31.0	600 40.0
	Column Total	360 24.0	1140 76.0	1500 100.0

reported, we imply that assignment status has some effect on gender. This is implausible. When percentages are reported by rows, the implication is that gender influences selection for overseas assignments.

Care should be taken in interpreting percentages from tables. Consider again the data in Exhibit 16–18. From the first to the second study, it is apparent that the percentage of females selected for overseas assignment rose from 6 to 9 percent of their respective samples. This is not to be confused with the percentage of women in the samples who happen to be assignees. Among all *women eligible* for selection in the first study, 15.8 percent were assigned and 84.2 percent were not. Among all *overseas selectees* in the first study, 21.4 percent were women. Similar comparisons can be made for the other categories.

Percentages are used by virtually everyone dealing with numbers—and often incorrectly. The following guidelines, if used during analysis, will help to prevent errors in reporting.[22]

MANAGEMENT Tip

- **Averaging percentages.** Percentages cannot be averaged unless each is weighted by the size of the group from which it is derived. Thus, a simple average will not suffice; it is necessary to use a weighted average.
- **Use of too large percentages.** This often defeats the purpose of percentages—which is to simplify. A large percentage is difficult to understand and is confusing. If a 1,000 percent increase is experienced, it is better to describe this as a tenfold increase.
- **Using too small a base.** Percentages hide the base from which they have been computed. A figure of 60 percent when contrasted with 30 percent would appear to suggest a sizable difference. Yet if there are only three cases in the one category and six in the other, the differences would not be as significant as they have been made to appear with percentages.
- **Percentage decreases can never exceed 100 percent.** This is obvious, but this type of mistake occurs frequently. The higher figure should always be used as the base. For example, if a price was reduced from $1 to $.25, the decrease would be 75 percent (75/100).

SNAPSHOT

Measuring Financial Potential in Small Business

LIMRA International is a trade association for the financial services industry. It believed that one high-growth segment of its market was small businesses, as these potential customers had been growing as a segment and had the resources to afford financial services. To assist its membership in capitalizing on this potential, LIMRA undertook a massive study of the U.S. small business sector, involving telephone interviews with financial decision makers in 1,622 businesses having fewer than 100 employees. Government agencies or units were excluded, as were businesses in financial services (financial, insurance, and real estate) and nonprofit organizations. Each interviewee who agreed to participate further was mailed a follow-up questionnaire; 533 respondents returned completed surveys.

The study was designed to (1) provide a broad profile of this market segment, including a profile of the small business principal; (2) assess the potential of group life, health care and retirement plans; and (3) recommend actions for taking advantage of the market opportunities identified. The last time LIMRA had done a similar study was in 1994. The 2000 study found a significant number of new small businesses, many of which provided some group benefit (70 percent, compared to 64 percent), and sponsored retirement or pension plans (43 percent, compared to 26 percent). Most change was motivated by a tight job market, "declines in the cost of benefits relative to gains in wages/salaries, changes in tax laws, and passage of the Small Business Job Protection Act of 1996." What would you want to consider before analyzing the results of this study?

www.limra.org

EXHIBIT 16–19 SPSS Cross-Tabulation with Control and Nested Variables

	Control Variable					
	Category 1			Category 2		
	Nested Variable			Nested Variable		
	cat 1	cat 2	cat 3	cat 1	cat 2	cat 3
Stub...	Cells...					

	SEX OF EMPLOYEE			
	MALES		FEMALES	
	MINORITY CLASSIFICATION		MINORITY CLASSIFICATION	
	WHITE	NONWHITE	WHITE	NONWHITE
EMPLOYMENT CATEGORY				
CLERICAL	16%	7%	18%	7%
OFFICE TRAINEE	7%	3%	17%	2%
SECURITY OFFICER	3%	3%		
COLLEGE TRAINEE	7%	0%	1%	
EXEMPT EMPLOYEE	6%	0%	0%	
MBA TRAINEE	1%	0%	0%	
TECHNICAL	1%			

Other Table-Based Analysis

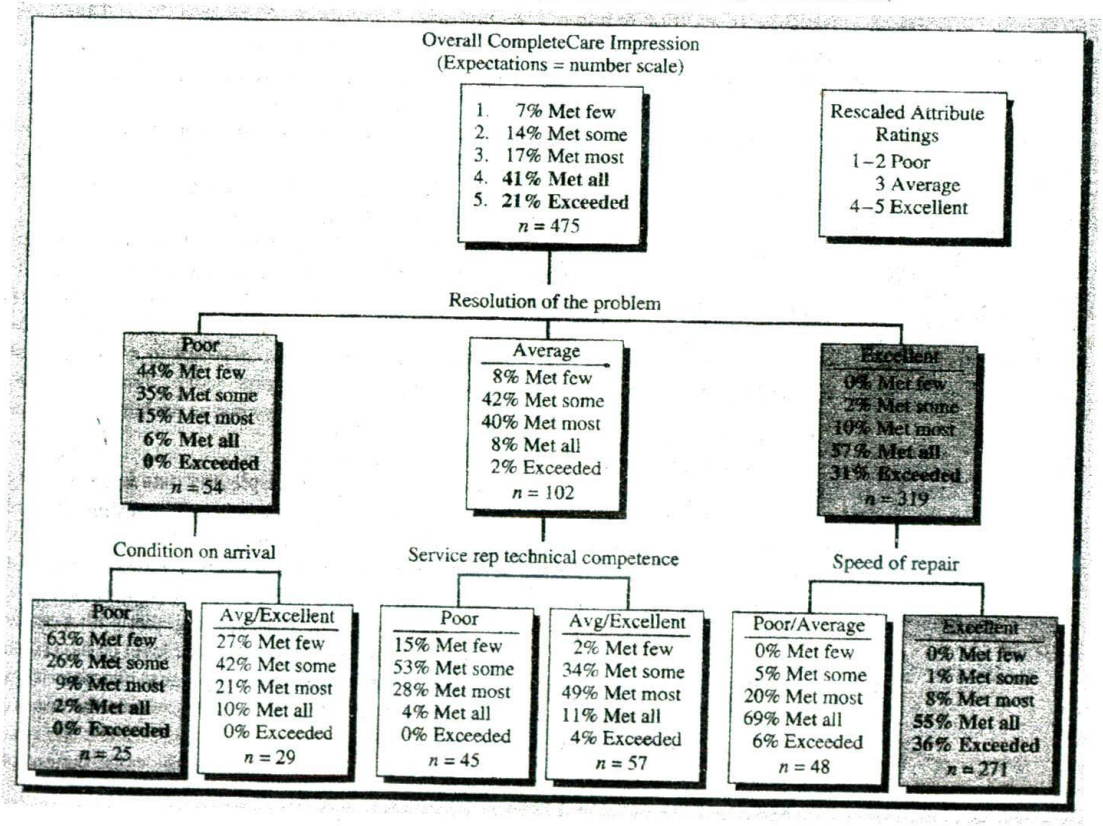
The recognition of a meaningful relationship between variables generally signals a need for further investigation. Even if one finds a statistically significant relationship, the questions of why and under what conditions remain. The introduction of a **control variable** to interpret the relationship is often necessary. Cross-tabulation tables serve as the framework.

Statistical packages like Minitab, SAS, and SPSS have among their modules many options for the construction of *n*-way tables with provision for multiple control variables. Suppose you are interested in creating a cross-tabulation of two variables with one control. Whatever the number of values in the primary variables, the control variable with five values determines the number of tables. For some applications, it is appropriate to have five separate tables; for others, it might be preferable to have adjoining tables or have the values of all the variables in one. Management reports are of the latter variety. Exhibit 16–19 presents an example in which all three variables are handled under the same banner. Programs such as this one can handle far more complex tables and statistical information.[23]

An advanced variation on *n*-way tables is **automatic interaction detection (AID).** AID is a sequential partitioning procedure that begins with a dependent variable and a set of predictors. It searches among up to 300 variables for the best single division according to each predictor variable, chooses one, and splits the sample using chi-square tests to create multiway splits. These subgroups then become separate samples for further analysis. The search procedure is repeated to find the variable that, when split into parts, makes the next largest contribution to the reduction of unexplained variation in each subsample, and so on.

Exhibit 16–20 shows the tree diagram that resulted from an AID study of MindWriter's CompleteCare repair service. The initial dependent variable is the overall

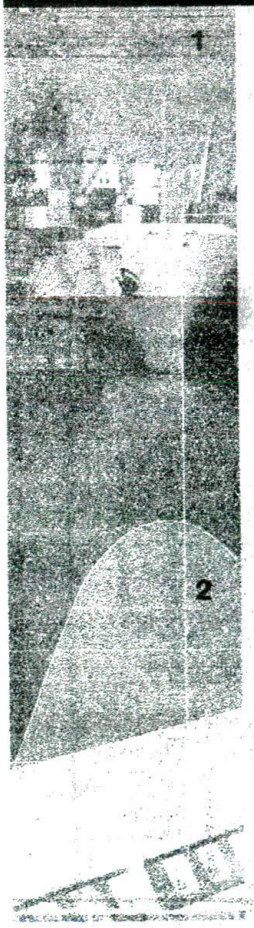
EXHIBIT 16–20 Automatic Interaction Detection Example (MindWriter's Repair Satisfaction)



impression of the repair service. This variable was measured on an interval scale. The variables that contribute to perceptions of repair effectiveness were also measured on the same scale but were rescaled to nominal data for this example (1–2 = poor, 3 = average, and 4–5 = excellent). The top box shows that 62 percent of the respondents rated the repair service as excellent (41% + 21%). The best predictor of effectiveness is "resolution of the problem." On the left side of the tree, customers who rated "resolution of the problem" as poor have fewer expectations being met or exceeded than the average (6% versus 62%). A poor rating on "condition on arrival" exacerbates this, reducing the total satisfied group to 2 percent.

On the right side of the tree, for those customers who rated "resolution of the problem" as excellent, the repair met or exceeded the average respondent's expectations (88% versus 62%). Excellent scores on "speed of repair" further improved this rating, taking the total satisfaction up to 91 percent. This analysis alerts decision makers at MindWriter to the best- and worst-case scenarios for the service, how to recover during a problematic month, and which "key drivers" of the process should receive corrective resources.

SUMMARY



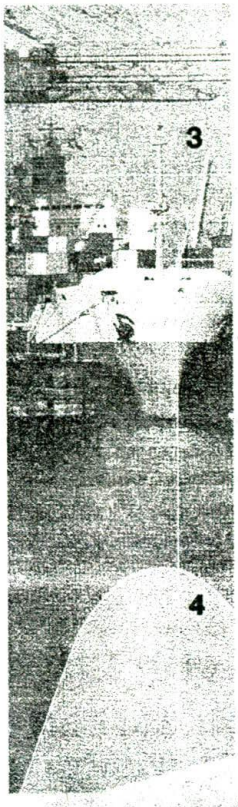
The objective of exploratory data analysis is to learn as much as possible about the data. Exploratory data analysis (EDA) simplifies this goal by providing a perspective and set of tools to search for clues and patterns. EDA augments rather than supplants traditional statistics. In addition to numerical summaries of location, spread, and shape, EDA uses visual displays to provide a complete and accurate impression of distributions and variable relationships.

Frequency tables array data from highest to lowest values with counts and percentages. They are most useful for inspecting the range of responses and their repeated occurrence. Bar charts and pie charts are appropriate for relative comparisons of nominal data. Histograms are optimally used with continuous variables where intervals group the responses.

Stem-and-leaf displays and boxplots are EDA techniques that provide visual representations of distributions. The former present actual data values using a histogram-type device that allows inspection of spread and shape. Boxplots use the five-number summary to convey a detailed picture of a distribution's main body, tails, and outliers. Both stem-and-leaf displays and boxplots rely on resistant statistics to overcome the limitations of descriptive measures that are subject to extreme scores. Transformation may be necessary to reexpress metric data so as to reduce or remove problems of asymmetry, inequality of variance, or other abnormalities.

The feature of statistical process control that pertains to displaying data is its charting system. SPC charts provide reliable visuals for evaluating point values and trends and graphically depicting variation. Control charts help managers focus on special causes of variation by revealing whether a system is under control (as an early warning of change) and substantiating results from improvements (confirming results).

A control chart displays sequential measurements of a process together with a center line and control limits. The selection of a control chart depends on the level of data (data type) you are measuring. Managers should look for the following visual characteristics in reports containing control charts: (1) outliers: observations falling outside the control lines; (2) runs: data points in a series over or below the central line; (3) trends: the continual rise or fall of data points; and (4) periodicity: data that show the same pattern of change over time, creating a cycle.



The Pareto diagram is a bar chart whose percentages sum to 100 percent. The causes of the problem under investigation are sorted in decreasing importance with bar height descending from left to right. Its pictorial array reveals the highest concentration of quality improvement potential in the fewest number of remedies.

Geographic Information Systems (GIS) refers to systems of hardware and software and procedures that capture, store, manipulate, integrate, and display spatially referenced data. GIS stacks different data sets on top of each other, so that the data points from variables of interest align on a common geographical referent, allowing the user to drill down and visualize the relationships. A GIS locator may be a zip code, census boundary, or coordinates of longitude, latitude, and elevation.

Most GIS have at least four process components: (1) integrating information from various sources, (2) capturing data, (3) projection and restructuring, and (4) modeling. By integrating variables from different sources, combinations of mapped variables may be created to simulate and analyze new ones. Data capture and editing is the most labor-intensive component of GIS because of the numerous variables and map objects. Projection converts a three-dimensional curved surface to a two-dimensional medium such as paper or a computer screen. Data collected and stored in different formats must be restructured for mapping. GIS modeling answers "what if" questions for such diverse applications as sales targeting, natural resource preservation, emergency service delivery, and service requirements of growing populations.

The evaluation of relationships involving categorical variables employs cross-tabulation. The tables used for this purpose consist of cells and marginals. The cells contain combinations of count, row, column, and total percentages. The tabular structure is the framework for later statistical testing.

Computer software for cross-classification analysis makes table-based analysis with one or more control variables an efficient tool for decision making. An advanced variation on n-way tables is automatic interaction detection (AID).

KEY TERMS

automatic interaction detection (AID) 509
boxplot 492
cells 505
confirmatory data analysis 487
contingency tables 506
control chart 499
control variable 508
cross-tabulation 505
exploratory data analysis (EDA) 486
five-number summary 492
frequency table 488
Geographic Information Systems (GIS) 503
histogram 489
interquartile range (IQR) 493
lower control limit (LCL) 499
marginals 505
nonresistant statistics 493
outliers 494
Pareto diagram 503
resistant statistics 492
standard score (Z score) 496
statistical process control (SPC) 498
stem-and-leaf display 491
Total Quality Management (TQM) 498
transformation 496
upper control limit (UCL) 499

EXAMPLES

Company	Scenario	Page
Booth Research Services	Digesting reams of data to assist research sponsors in making management decisions.	498
City Center for Performing Arts	Preliminary data analysis described.	BRTL

Federal Emergency Management Agency (FEMA)	Using GIS to provide national-level support in responding to disasters.	504
Forbes Industry List	Industries, business classifications, and market sectors displayed to illustrate frequency tables and pie and bar charts.	488
Ford Motor Co.	Challenged the Insurance Institute for Highway Safety's crash tests results and conclusions.	502
Fortune 500	Database used to demonstrate the construction of histograms, stem-and-leaf displays, and boxplot techniques!	488
Insurance Institute for Highway Safety	Conducted safety experiment involving frontal offset crash tests of competitive, full-sized trucks, including those manufactured by Ford, GM, DaimlerChrysler, and Toyota.	502
LIMRA International	Conducted a large-scale telephone survey of 1,622 small businesses for the financial services industry.	508
MerK Outboard*	Manufacturer of outboard motors for small fishing boats whose manufacturing process uses control charts to assure production quality at design specifications.	500
MindWriter*	(1) A content analysis of MindWriter customer complaints depicted as a Pareto diagram.	503
	(2) An AID tree diagram of MindWriter's CompleteCare repair service; factors driving overall customer satisfaction.	509
Springfield Fire Department	Using GIS analysis to determine the ideal location for a new fire station.	505

*Due to the confidential and proprietary nature of most research, the names of some companies have been changed.

DISCUSSION QUESTIONS

Terms in Review

1. Define or explain:
   a. Marginals.
   b. Pareto diagram.
   c. Standard scores (Z scores).
   d. Control chart.
   e. Nonresistant statistics.
   f. Lower control limit.
   g. The five-number summary.
   h. Geographic Information Systems.
   i. Spread-and-level plots.

Making Research Decisions

2. How should the researcher handle "don't know" responses?
3. How do the following detect errors in the data?
   a. Histogram.
   b. Stem-and-leaf display.
   c. Boxplot.
   d. Cross-tabulation.

EXHIBIT 16–21 Data Table for Discussion Questions 5–6

	Market Value	Sales	Sector		Market Value	Sales	Sector
1	24983.00	8966.00	2	26	9009.00	17533.00	4
2	31307.00	126932.00	3	27	7842.00	11113.00	2
3	57193.00	54574.00	7	28	5431.00	19671.00	8
4	57676.00	86656.00	4	29	5811.00	11389.00	5
5	60345.00	62710.00	7	30	16257.00	15242.00	2
6	22190.00	96146.00	3	31	16247.00	10211.00	7
7	36566.00	39011.00	2	32	18548.00	9593.00	7
8	44646.00	36112.00	7	33	13620.00	9691.00	7
9	25022.00	50220.00	4	34	10750.00	12844.00	3
10	26043.00	25099.00	1	35	12450.00	18398.00	2
11	13152.00	53794.00	2	36	16729.00	20276.00	7
12	11234.00	25047.00	5	37	16532.00	8730.00	7
13	26666.00	23966.00	4	38	5111.00	17635.00	10
14	20747.00	17424.00	7	39	9116.00	8588.00	4
15	25826.00	13996.00	7	40	26325.00	25922.00	2
16	15423.00	32416.00	4	41	8249.00	16103.00	2
17	15263.00	14150.00	8	42	8407.00	14083.00	3
18	18146.00	17600.00	1	43	18537.00	11990.00	10
19	18739.00	15351.00	4	44	23866.00	29443.00	4
20	7875.00	22605.00	2	45	6872.00	19532.00	7
21	8122.00	37970.00	5	46	4319.00	10018.00	5
22	18072.00	11557.00	5	47	9505.00	12937.00	7
23	6404.00	11449.00	7	48	3891.00	15654.00	8
24	16056.00	20054.00	8	49	8090.00	7492.00	4
25	16056.00	13211.00	7	50	11119.00	12345.00	7

From Concept to Practice

4. Use the data in Exhibit 16–4 to construct a stem-and-leaf display.

a. Where do you find the main body of the distribution?

b. How many values reside outside the inner fence(s)?

5. Select the sales variable from Exhibit 16–21.

a. Create a five-number summary.

b. Construct a boxplot.

c. Interpret the distribution and results with summary measures and descriptive statistics.

d. Transform the variable into Z scores.

e. Identify and comment on outliers, if any.

6. Select the market value variable from Exhibit 16–21 and construct a histogram with available software.
   a. What is the gain in information with 5,000-, 2,000-, or 1,000-unit intervals?
   b. Which would be the best interval to convey results to management?
   c. Why would these data need reexpression?
   d. What is the optimal power transformation for these data?
7. Suppose you were preparing two-way tables of percentages for the following pairs of variables. How would you run the percentages?
   a. Age and consumption of breakfast cereal.
   b. Family income and confidence about the family's future.
   c. Marital status and sports participation.
   d. Crime rate and unemployment rate.
8. You study the attrition between students who enter college as freshmen and those who stay to graduate. You find the following relationships between attrition, aid, and distance of home from school. What is your interpretation? Consider all variables and relationships.

	Aid		Home Near Aid		Home Far Aid	
	Yes	No	Yes	No	Yes	No
Drop out	25%	20%	5%	15%	30%	40%
Stay	75	80	95	85	70	60

9. A local health agency is experimenting with two appeal letters, A and B, with which to raise funds. It sends out 400 of the A appeal and 400 of the B appeal (divided equally among working-class and middle-class neighborhoods). The agency secures the results shown in the table below.
   a. Which appeal is the best?
   b. Which class responded better?
   c. Is appeal or social class a more powerful independent variable?

	Appeal A		Appeal B	
	Middle Class	Working Class	Middle Class	Working Class
Contribution	20%	40%	15%	30%
No contribution	80	60	85	70
	100%	100%	100%	100%

10. Assume you have collected data on employees of a large organization in a major metropolitan area. You analyze the data by type of work classification, education level, and whether the workers were raised in a rural or urban setting. The results are shown below. How would you interpret them?

Annual Employee Turnover per 100 Employees

	Part A		Part B			
			High Education		Low Education	
	Salaried	Wage	Salaried	Wage	Salaried	Wage
Rural	8	16	6	14	18	18
Urban	12	16	10	12	19	20

11. Analyze the MerK Outboard Motor data as shown and produce R- and X-bar charts. Compare your results to Exhibit 16–15. What has occurred in the manufacturing process?

Subgroup Sample	Horsepower Output		
1	10.22	10.31	8.43
2	11.25	10.33	10.12
3	10.37	9.92	11.80
4	7.53	9.94	9.94
5	10.06	9.39	10.31
6	12.01	10.15	10.23
7	7.95	8.25	8.55
8	10.13	10.69	10.16
9	10.82	11.25	11.13
10	9.88	8.95	9.25

WWW Exercises Visit our website for Internet exercises related to this chapter at www.mhhe.com/business/Cooper8

CASES

AGRICOMP

HEALTHY LIFESTYLES

INQUIRING MINDS WANT TO KNOW—NOW!

KNSD SAN DIEGO

MASTERING TEACHER LEADERSHIP

NCR: TEEING UP A NEW STRATEGIC DIRECTION

XEROX ABUSES

*All cases indicating a video icon are located on the Instructor's Videotape Supplement. All nonvideo cases are in the case section of the textbook. All cases indicating a CD icon offer a data set, which is located on the accompanying CD.

REFERENCE NOTES

1. John W. Tukey, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley, 1977).
2. David C. Hoaglin, Frederick Mosteller, and John W. Tukey, eds., *Understanding Robust and Exploratory Data Analysis* (New York: John Wiley & Sons, 1983), p. 2.
3. Tukey, *Exploratory Data Analysis*, pp. 2–3.
4. Frederick Hartwig with Brian E. Dearing, *Exploratory Data Analysis* (Beverly Hills, CA: Sage Publications, 1979), pp. 9–12.
5. The exhibits in this section were created with statistical and graphic programs particularly suited to exploratory data analysis. The authors acknowledge the following vendors for evaluation and use of their products: SPSS, Inc., 233 S. Wacker Dr., Chicago, IL, 60606; and Data Description, P.O. Box 4555, Ithaca, NY, 14852.
6. "Fortune 500 Ranked by Performance," *The Fortune 500*, April 28, 1997, p. F-30.
7. Paul F. Velleman and David C. Hoaglin, *Applications, Basics, and Computing of Exploratory Data Analysis* (Boston: Duxbury Press, 1981), p. 13.
8. John Hanke, Eastern Washington University, contributed this section. For further references to stem-and-leaf displays, see John D. Emerson and David C. Hoaglin, "Stem-and-Leaf Displays," in *Understanding Robust and Exploratory Data Analysis*, pp. 7–31; and Velleman and Hoaglin, *Applications*, pp. 1–13.
9. This section is adapted from the following excellent discussions of boxplots: Velleman and Hoaglin, *Applications*, pp. 65–76; Hartwig, *Exploratory Data Analysis*, pp. 19–25; John D. Emerson and Judith Strenio, "Boxplots and Batch Comparison," in *Understanding Robust and Exploratory Data Analysis*, pp. 59–93; and Amir D. Aczel, *Complete Business Statistics* (Homewood, IL: Irwin, 1989), pp. 723–28.
10. Tukey, *Exploratory Data Analysis*, pp. 27–55.
11. Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, p. 2.
12. Several robust estimators that are suitable replacements for the mean and standard deviation we do not discuss here—for example, the trimmed mean, trimean, the M-estimators (such as Huber's, Tukey's, Hampel's, and Andrew's estimators), and the median absolute deviation (MAD). See Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, Chapter 10; and SPSS, Inc., *SPSS Base 9.0 User's Guide* (Chicago: SPSS, Inc., 1999), Chapter 13.
13. The difference between the definition of a hinge and a quartile is based on variations in their calculation. We use Q_1, 25th percentile, and *lower hinge* synonymously; and Q_3, 75th percentile, and *upper hinge*, similarly. There are technical differences, although they are not significant in this context.
14. R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," *The American Statistician* 32 (1978), pp. 12–16.
15. See J. Chambers, W. Cleveland, B. Kleiner, and John W. Tukey, *Graphical Methods for Data Analysis* (Boston: Duxbury Press, 1983).
16. This section is based on the discussion of transformation in John D. Emerson and Michael A. Stoto, "Transforming Data," in *Understanding Robust and Exploratory Data Analysis*, pp. 97–127; and Velleman and Hoaglin, *Applications*, pp. 48–53.
17. Hoaglin et al., *Understanding Robust and Exploratory Data Analysis*, p. 71.
18. Ibid., p. 125.
19. Tor Bernhardsen, *Geographic Information Systems: An Introduction* (New York: John Wiley & Sons, 1999).
20. Example from the U.S. Geological Survey FAQ for GIS: http://www.usgs.gov (March 4, 2000).
21. See examples in "Geographical Mapping and Analysis in Health Care," SPSS White Paper, 1998, p. 5. http://www.spss.com (March 15, 2000).
22. Harper W. Boyd, Jr., and Ralph Westfall, *Marketing Research*, 3rd ed. (Homewood, IL: Irwin, 1972), p. 540.
23. SPSS, Inc., *SPSS Tables 8.0* (Chicago: SPSS, Inc. 1998), with its system file: Bank Data.

REFERENCES FOR SNAPSHOTS AND CAPTIONS

Ford F-150

"What Is Offset Crash Testing?" Insurance Institute for Highway Safety (http://www.hwysafety.org/vehicle_ratings/ce/offset.htm#).

"Crashworthiness," Insurance Institute for Highway Safety, June 4, 2001 (http://www.hwysafety.org/vehicle_ratings/ce/pdfs/large_pickups.pdf).

"New Crash Test Results: Ratings of Four Large Pickups Range from Good for Toyota Tundra to Poor for Ford F-150, Dodge Ram," Insurance Institute for Highway Safety, June 4, 2001.

LIMRA

"U.S. Small Businesses in 2000: A Dynamic Market," © 2001, LIMRA International, Inc.

New York Power Outages (September 2001)

FEMA-DR-1391, New York Utility Outages as of 7:00 P.M., 9/18/01.

The Federal Emergency Management Agency's Mapping and Analysis Center (MAC) provides national-level Geographic Information System (GIS) support and coordination to the agency. GIS mapping products are available for the New York City World Trade Center Attack, the attack on the Pentagon, and the latest disasters, along with the current year and an archive of prior-year disasters (http://www.gismaps.fema.gov/2001pages/DR1391.shtml).

Springfield Fire Department

"Hot Spots in Springfield," Olga Medvedkov, 2000.

Olga Medvedkov, project supervisor, Springfield SFD-GIS study, interview, June 2000.

CLASSIC AND CONTEMPORARY READINGS

DeMers, Michael. *Fundamentals of Geographic Information Systems*. New York: John Wiley & Sons, 2000. Methodical coverage of basic input requirements, data management, reporting concepts, and ample depth in explaining spatial analysis issues. Highly regarded for its readability by students.

Evans, James R., and William M. Lindsay. *Management and the Control of Quality*. Mason, OH: South-Western Publishing, 2002. Technically detailed coverage of quality improvement techniques.

Hoaglin, David C., Frederick Mosteller, and John W. Tukey, eds. *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons, 2000. A complete and advanced treatment by influential writers in this field. Especially well-organized topical coverage.

Velleman, Paul F., and David C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press, 1981. The basics of EDA are presented in a straightforward style with helpful examples and excellent connections to computer applications.