CHAPTER 15

Data Preparation and Description

Learning Objectives

After reading this chapter, you should understand . . .

1 **The importance of editing the collected raw data to detect errors and omissions.**

2 **How coding is used to assign numbers and other symbols to answers in order to classify responses.**

3 **The use of content analysis to interpret and summarize open questions.**

4 **Problems and solutions for "don't know" responses.**

5 **The options for data entry and manipulation.**

6 **How to select descriptive statistics to summarize the collected data and check for errors.**

Bringing Research to Life

Dorie Henry arrived for lunch, joining Jason and Myra, who were wrapping up a MindWriter meeting. The conversation naturally turned to one of Jason's early research projects.

"As a public health doctor," said Dorie, "I work with statistics—life-threatening ones. I understand Jason's perfectionism and why he carries it to such extremes. You know how difficult he can be when he is under stress."

Myra responded by arching her eyebrows and laughing mirthlessly.

"I can see this turning into a bashing Jason story," Jason groaned.

"Has Jason told you about our near blowup over the bus line survey? No? Before we moved down here and he connected with you, Jason was retained by a bus line up north to assist in a sample survey of 2,000 passengers. The line had just dismissed its research director, a horrible chap named Burbidge, who could not relate to the blue-collar riders.

"Jason had a 20-question survey of 2,000 randomly selected riders, a total of 40,000 items to be coded, some of them multiple choice, some of them numerical entry, some free response. He estimated this would take 15 seconds an item, on the average, a total of 10,000 minutes, which he rounded up to 12,000 minutes. Call it 200 person-hours, then. He decided to hire 25 temporary personnel for one eight-hour data entry shift."

"That sounds reasonable," said Myra.

"Well, of course it was quite reasonable. What wasn't 'reasonable' was Jason. Since CityBus received a county subsidy, the bus line had to follow county purchasing procedures and put the data entry out for competitive bidding.

"As board member of a nonprofit organization that provides temporary employment for former mental patients, I insisted on bidding for the data entry work, over Jason's strongest objections. He had never expressed any objections to my finding work for these people before, but when I wanted them involved in his own work, he became prickly, hostile, self-pitying, irritable, and hypercritical."

"I've seen that side of Jason," grinned Myra.

"Would you like me to leave so you can shred me without compunction?" replied Jason.

"I'm certain you have," said Dorie, ignoring Jason, "and certainly you know what motivated it—the most intense desire for accuracy and reliability."

"Which, as his client, I not only appreciate but demand," supported Myra. "See, Jason, Dorie's not really telling tales."

"Thanks to his association with you, Myra, he has learned to moderate the acting-out of his anxieties, for which I thank you. But on the inside he ceaselessly worries about threats to validity.

"I won the CityBus contract, and Jason had no choice. The contract had me bringing my crew in at 5 P.M. on a Friday evening to bus line HQ, putting each worker in a separate office, using 25 networked computers in employee offices. Each member of the crew, sitting in a separate office, was given 80 surveys to encode, 10 an hour, for eight hours.

"Jason procured a data entry program, which we ran off the system server. The program very cleverly contrived to prevent errors of data entry, so far as possible. For instance, if a survey item—a question—called for choosing between 'Ride the bus less than twice a week,' 'Ride the bus at least twice but less than 8 times,' or 'Ride the bus 8 or more times,' the data clerk would put a cursor over the correct answer and press ENTER. Making a mistake would be difficult, because the screen would closely resemble the printed page. If the item called for "Enter your age," and the clerk entered a three-digit number, the computer would beep loudly, to indicate that ages less than 10 and greater than 99 were considered suspect.

"Jason continued to face the prospect of DEN—data entry night—with thinly disguised opposition. He grumbled that my crew would ruin the results of his carefully crafted research. I reminded him that these people were not visually impaired and had no intellectual deficits. They were just folks who were not yet ready to reenter the workplace or were ready but suffered unreasonable discrimination, of exactly the sort Jason was showing.

"Anyway, while it was prudent to warn the HQ staff to secure their offices, Jason went overboard and infected the office staff with his nervousness and negativity about my crew coming in to use their computers. At 4:45 P.M., our bus arrived with me, my crew, and a supervisory social worker, to find office workers departing carrying out their potted plants. Did they think my crew were herbivores? It was a flagrant insult, and I felt awful."

"You are going to tell the part where I redeem myself, aren't you?" asked Jason plaintively.

"Jason, concerned with human error, double-checked the data entry. A 26th crew member ran up and down the halls all night retrieving randomly sampled surveys, which Jason compared with the already entered data in the server. He warned me that if he found more than a certain error rate he intended to stop the operation and send everyone home. He was working himself into a lather, looking for trouble he was not going to find.

"As he was pulling the forms generated by 25 workers, he had to work at least as fast as any one of them to keep up, slightly faster in fact, and very frankly I had serious doubts that he would not make more errors than my crew and blame the errors on them. He skipped supper at 8 P.M. and worked steadily and crankily and drank too much black coffee. No bathroom breaks or stretch breaks for super sleuth. But at 10 P.M., five hours into our eight-hour shift, he threw down his surveys, smiled broadly, murmured 'darn good,' strolled off to the bathroom, returned, ate two peanut butter and jelly sandwiches and made a few dumb jokes.

"It seems," continued Dorrie, "the discrepancy rate between his coding and my crew was less than 1 percent for the first five hours. He spent the next two hours humming 'Strangers in the Night,' over and over again and running preliminary frequency distributions and cross-tabulations."

"In the end, Jason was 'big' enough to express considerable satisfaction with your crew, right?" inquired Myra.

"Of course. After all, when HQ office workers came back Monday, toting their philodendrons, they found their offices clean, orderly, and mysteriously mint-scented. Each of my crew had cleaned the keyboards and monitor screens with little sanitized towelettes. And Jason had error-free data in the bargain."

Introduction

Once the data begin to flow in, attention turns to data analysis. If the project has been done correctly, the analysis planning is already done. Back in the research design stage or at least by completion of the proposal or the pilot test, decisions should have been made about how to analyze the data. Unfortunately, many researchers wait until the analysis stage to decide what to do. This results in the late discovery that some data will not be collected, will be collected in the wrong form, or will exhibit unanticipated characteristics.

This chapter addresses two topics. The first is *data preparation,* which includes editing, coding, and data entry. These activities ensure the accuracy of the data and their conversion from raw form to reduced and classified forms that are more appropriate for

analysis. Second, preparing a descriptive statistical summary is the preliminary step leading to an understanding of the collected data. If your research objective is reporting or basic description, this section is particularly valuable. We discuss the definitions and applications of descriptive statistics, specifically the characteristics of location, spread, and shape that give us insight into the distributions of respondent observations.

Before concluding this introduction, we must comment briefly on our assumption for this part of the book. The book was developed for and with graduate students. It may be used successfully with undergraduates if an elementary statistics course was a prerequisite. Without that grounding, some of the ideas presented may require remedial study.

Editing

The customary first step in analysis is to edit the raw data. **Editing** detects errors and omissions, corrects them when possible, and certifies that minimum data quality standards have been achieved. The editor's purpose is to guarantee that data are:

MANAGEMENT



- Accurate.
- Consistent with intent of the question and other information in the survey.
- Uniformly entered.
- Complete.
- Arranged to simplify coding and tabulation.

In the following question asked of military officers, one respondent checked two categories, indicating that he was in the reserves and currently serving on active duty.

Please indicate your current military status:

- ☑ Active duty officer
- ☐ National Guard officer
- ☑ Reserve officer
- ☐ Separated officer
- ☐ Retired officer

The editor's responsibility is to decide which of the responses is both consistent with the intent of the question or other information in the survey and most accurate for this individual respondent.

Field Editing

In large projects, field editing review is a responsibility of the field supervisor. It, too, should be done soon after the data have been gathered. During the stress of data collection, the researcher often uses ad hoc abbreviations and special symbols. Soon after the interview, experiment, or observation, the investigator should review the reporting forms. It is difficult to complete what was abbreviated or written in shorthand or noted illegibly if the entry is not caught that day. When entry gaps are present from interviews, a callback should be made rather than guessing what the respondent "probably would have said." Self-interviewing has no place in quality research.

A second important control function of the field supervisor is to validate the field results. This normally means he or she will reinterview some percentage of the respondents, at least on some questions. Many commercial research firms will recontact about 10 percent of the respondents.

Central Editing

At this point, the data should get a thorough editing. For a small study, the use of a single editor produces maximum consistency. In large studies, the tasks may be broken down so that each editor can deal with one entire section. This approach will not identify inconsistencies between answers in different sections. However, this problem can be handled by identifying points of possible inconsistency and having one editor check specifically for them.

Sometimes it is obvious that an entry is incorrect, is entered in the wrong place, or states time in months when it was requested in weeks. When replies clearly are inappropriate or missing, the editor can sometimes detect the proper answer by reviewing the other information in the schedule. This practice, however, should be limited to those few cases where it is obvious what the correct answer is. It may be better to contact the respondent for correct information, if time and budget allow. Another alternative is for the editor to strike out the answer if it is clearly inappropriate. Here an editing entry of "no answer" or "unknown" is called for.

Another editing problem concerns faking an interview that never took place. This armchair interviewing is difficult to spot, but the editor is in the best position to do so. One approach is to check responses to open-ended questions. These are most difficult to fake. Distinctive response patterns in other questions will often emerge if faking is occurring. To uncover this, the editor must analyze the instruments used by each interviewer.

Here are some useful rules to guide editors in their work:

MANAGEMENT Tip

- Be familiar with instructions given to interviewers and coders.
- Do not destroy, erase, or make illegible the original entry by the interviewer; original entries should be crossed out with a single line to remain legible.
- Make all entries on an instrument in some distinctive color and in a standardized form.
- Initial all answers changed or supplied.
- Place initials and date of editing on each instrument completed.

Coding

Coding involves assigning numbers or other symbols to answers so the responses can be grouped into a limited number of classes or categories. The classifying of data into limited categories sacrifices some data detail but is necessary for efficient analysis. Instead of requesting the word *male* or *female* in response to a question that asks for the identification of one's gender, we could use the codes "M" or "F." Normally this variable would be coded 1 for male and 2 for female or 0 and 1. If we use M and F, or other letters, in combination with numbers and symbols, the code is *alphanumeric*. When numbers are used exclusively, the code is *numeric*.

Coding helps the researcher to reduce several thousand replies to a few categories containing the critical information needed for analysis. In coding, *categories* are the partitioning of a set; and *categorization* is the process of using rules to partition a body of data.

Coding Rules

Four rules guide the establishment of category sets. The categories should be:

- Appropriate to the research problem and purpose.
- Exhaustive.

- Mutually exclusive.
- Derived from one classification principle.

Appropriateness Categories must provide the best partitioning of data for testing hypotheses and showing relationships. Year-by-year age differences may be important to the question being researched. If so, wider age classifications hamper the analysis. If specific income, attitude, or reason categories are critical to the testing relationship, then we must choose the best groupings. In particular, choose class boundaries that match those being used for comparisons. It is disheartening, late in a study, to discover that age, income, or other frequency classes do not precisely match those of the data with which we wish to make comparisons.

Exhaustiveness A large number of "other" responses suggests our classification set may be too limited. In such cases, we may not be tapping the full range of information in the data. Failure to present an adequate list of alternatives is especially damaging when multiple-choice questions are used. Any answer that is not specified in the set will surely be underrepresented in the tally.

While the exhaustiveness requirement in a single category set may be obvious, a second aspect is less apparent. Does the one set of categories fully capture all the information in the data? For example, responses to an open-ended question about family economic prospects for the next year may originally be classified only in terms of being optimistic or pessimistic. It may also be enlightening to classify responses in terms of other concepts such as the precise focus of these expectations (income or jobs) and variations in responses between family heads and others in the family.

Mutual Exclusivity Another important rule is that category components should be mutually exclusive. This standard is met when a specific answer can be placed in one and only one cell in a category set. For example, in an occupation survey, the classifications may be (1) professional, (2) managerial, (3) sales, (4) clerical, (5) crafts, (6) operatives, and (7) unemployed. Some respondents will think of themselves as being in more than one of these groups. The person who views selling as a profession and spends time supervising others may fit in three of these categories. A function of operational definitions is to provide categories that are composed of mutually exclusive elements. Here, operational definitions of the occupations to be classified in "professional," "managerial," and "sales" should clarify the situation. The problem of how to handle an unemployed salesperson brings up a fourth rule of category design.

Single Dimension The need for a category set to follow a single classificatory principle means every class in the category set is defined in terms of one concept. Returning to the occupation survey example, the person in the study might be both a salesperson and unemployed. The "salesperson" label expresses the concept *occupation type;* the response "unemployed" is another dimension concerned with *current employment status* without regard to the respondent's normal occupation. When a category set uses more than one dimension, it will normally not be mutually exclusive unless the cells in the set combine the dimensions (employed manager, unemployed manager, and so on).

Codebook Construction

A **codebook,** or *coding scheme,* contains each variable in the study and specifies the application of coding rules to the variable. It is used by the researcher or research staff as a guide to make data entry less prone to error and more efficient. It is also the definitive

source for locating the positions of variables in the data file during analysis. In many statistical programs, the coding scheme is integral to the data file. Most codebooks—computerized or not—contain the question number, variable name, location of the variable's code on the input medium, descriptors for the response options, and whether the variable is alpha or numeric. An example of a paper-based codebook is shown in Exhibit 15–1. When pilot testing has been conducted, there should be sufficient infor-

EXHIBIT 15–1 Sample Codebook of Questionnaire Items

____	1	Record number	RECNUM
____	2	Respondent number	RESID
1	3	5 digit zip code 99999 = Missing	ZIP
2	4	2 digit birth year 99 = Missing	BIRTH
3	5	Gender 1 = Male 2 = Female 9 = Missing	GENDER
4	6	Marital status 1 = Married 2 = Widow(er) 3 = Divorced 4 = Separated 5 = Never married 9 = Missing	MARITAL
5	7	Own–Rent 1 = Own 2 = Rent 3 = Provided 9 = Missing	HOUSING
6		Reason for purchase 1 = Mentioned 0 = Not mentioned	
	8	Bought home	HOME
	9	Birth of child	BIRTHCHD
	10	Death of relative or friend	DEATH
	11	Promoted	PROMO
	12	Changed job/career	CHGJOB
	13	Paid college expenses	COLLEXP
	14	Acquired assets	ASSETS
	15	Retired	RETIRED
	16	Changed marital status	CHGMAR
	17	Started business	STARTBUS
	18	Expanded business	EXPBUS
	19	Parent's influence	PARENT
	20	Contacted by agent	AGENT
	21	Other	OTHER

mation about the variables to prepare a codebook. A preliminary codebook used with pilot data may reveal coding problems that will need to be corrected before the data for the final study are collected and processed.

Coding Closed Questions

The responses to closed questions include scaled items and others for which answers can be anticipated. When codes are established early in the research process, it is possible to precode the questionnaire. **Precoding** is particularly helpful for data entry because it makes the intermediate step of completing a coding sheet unnecessary. The data are accessible directly from the questionnaire. A respondent, interviewer, field supervisor, or researcher (depending on the data collection method) is able to assign an appropriate numerical response on the instrument by checking, circling, or printing it in the proper coding location.

Exhibit 15–2 shows questions in the sample codebook. When precoding is used, editing may precede data processing. Note question 4, where the respondent may

EXHIBIT 15–2 Sample Questionnaire Items

1. What is the zip code of your residence? _ _ _ _ _
2. What is the year of your birth? 19_ _
3. Gender (1) Male
   (2) Female
   Indicate your choice → by number _
4. What is your marital status?
   (1) Married
   (2) Widow(er)
   (3) Divorced
   (4) Separated
   (5) Never married
   Indicate your choice → by number _
5. Do you own or rent your primary residence?
   (1) Own
   (2) Rent
   (3) Living quarters provided
   Indicate your choice → by number _
6. What prompted you to purchase your most recent life insurance policy?

_____

_____

_____

_____

_____

_____

choose between five characteristics of marital status and enter the number of the item best representing present status in the coding portion of the questionnaire. This code is later transferred to an input medium for analysis.

Coding Open-Ended Questions

Closed questions are favored by researchers over open-ended questions for their efficiency and specificity. They are easier to measure, record, code, and analyze. But there are situations where insufficient information or lack of a hypothesis prohibits preparing response categories in advance. Other reasons for using open-ended responses include the need to measure sensitive or disapproved behavior, discover salience, or encourage natural modes of expression.[1] However, analyzing enormous volumes of open-ended questions has always been a nightmare for researchers. The varieties of answers one may encounter is staggering. In Exhibit 15–2, question 6 illustrates the use of an open-ended question for which advance knowledge of response options was not available. The answer to "What prompted you to purchase your most recent life insurance policy?" was to be filled in by the respondent as a short-answer essay. After preliminary evaluation, response categories (shown in the codebook, Exhibit 15–1) were created for that item. Although most responses could be accounted for by the derived categories, an "other" category was established to meet the coding rule of exhaustiveness.

Using Content Analysis for Open Questions

Content analysis measures the semantic content or the *what* aspect of a message. Its breadth makes it a flexible and wide-ranging tool that may be used as a methodology or as a problem-specific technique. Trend-watching organizations like the BrainReserve, the Naisbitt Group, SRI International, and Inferential Focus use variations on content analysis for selected projects, often spotting changes from newspapers or magazine articles before they can be confirmed statistically. The Naisbitt Group's content analysis of 2 million local newspaper articles compiled over a 12-year period resulted in the publication of *Megatrends*. After a brief review of the characteristics of content analysis, we will consider one of its many problem-specific applications: handling open questions.

Content analysis has been described as "a research technique for the objective, systematic, and quantitative description of the manifest content of a communication."[2] Because this definition is sometimes confused with simply counting obvious message

SNAPSHOT

What's the Real Content of a Message?

Most potential customers who know the name associate V-8 with vegetables. They often don't think of it when choosing a thirst quencher. If they want to drink healthy, those same potential customers choose fruit juice or juice-based alternative beverages. The tomato, after all, is a fruit thought by many to be a vegetable. This split personality comes through in much of the early V-8 advertising. But Campbell wants V-8 to be more than the forgotten healthy beverage. Its ad agency FCB/Leper Katz Partners did extensive interviews with customers and noncustomers, but the initial campaign that evolved was as schizophrenic as the tomato itself. So Campbell hired PhaseOne Communications, a California firm specializing in content analysis. Its analysis revealed a typical ad contained dozens of messages, often in conflict with each other. Combining a thorough understanding of both the hidden and explicit messages revealed by this analysis with the data from in-depth interviews of customers of this venerable brand resulted in a new campaign. This new campaign imbued V-8 with testimonial "zip" so people would no longer say, "Wow! I could have had a V-8!"

www.campbellsoup.com

aspects such as words or attributes, more recent interpretations have broadened the definition to include latent as well as manifest content, the symbolic meaning of messages, and qualitative analysis. One author states:

> In any single written message, one can count letters, words, or sentences. One can categorize phrases, describe the logical structure of expressions, ascertain associations, connotations, denotations, elocutionary forces, and one can also offer psychiatric, sociological, or political interpretations. All of these may be simultaneously valid. In short, a message may convey a multitude of contents even to a single receiver.[3]

Content analysis follows a systematic process, starting with the selection of a unitization scheme. The units may be syntactical, referential, propositional, or thematic.

- *Syntactical* units are illustrated by words, which are the smallest and most reliable units.
- *Referential* units may be objects, events, persons, and so forth, to which an expression refers. An advertiser may refer to a product as a "classic," a "power performer," or "ranked first in safety"—each denoting the same object.
- *Propositional* units use several frameworks. One might show the relationships among the actor, the mode of acting, and the object—for example, "subscribers [actor] to this periodical save [mode of acting] $15 [object of the action] over the single issue rate."
- *Thematic* units are higher-level abstractions inferred from their connection to a unique structure or pattern in the content. A response to a question about working conditions may reflect a temporal theme: the past ("how good things used to be here"), the present ("the need to talk with management now before production gets worse"), or the future ("employee expectations to be involved in planning and goal setting").

Other aspects of the content analysis methodology include:

- Selection of a sampling plan.
- Development of recording and coding instructions.
- Data reduction.
- Inferences about the context.
- Statistical analysis.

Content analysis guards against selective perception of the content, provides for the rigorous application of reliability and validity criteria, and is amenable to computerization.

The data to be content-analyzed include materials of interest to management researchers: books, chapters, historical documents, speeches, interviews, advertisements, promotional brochures, group interactions, paragraphs, and words. Any recorded activity with its own syntax and semantics is subject to measurement and analysis. Thus, content analysis may be used to analyze written, audio, or video data from experiments, observations, surveys, and secondary data studies.

Let's look at an informal application of content analysis to a problematic open question. In this example, which we are processing without linguistics software technology, suppose employees in the assembly operation of a unionized manufacturing firm are asked, "How can management-employee relations be improved?" A sample of the responses yields the following:

- Management should treat the worker with more respect.
- Managers should stop trying to speed up the assembly line.

- Working conditions in the shop are terrible. Managers should correct them.
- The foreman should be fired. He is unfair in his treatment of workers.
- Managers should form management-worker councils in the department to iron out problems and improve relations.
- Management should stop trying to undermine union leadership.
- Management should accept the union's latest proposals on new work rules.

The first step in analysis requires that the units developed reflect the objectives for which the data were collected. The research question is concerned with learning what the assemblers think is the locus of responsibility for improving company-employee relations. The categories selected are keywords and referential units. The first pass through the data produces a few general categories, shown in Exhibit 15–3. These categories are mutually exclusive and contain only one concept dimension. The use of "other" makes the category set exhaustive. If the sample suggested that many respondents identified the need for action by the public, government, or regulatory bodies, then including all of them in "other" would ignore much of the richness of the data.

Since responses to this type of question often suggest specific actions, the second evaluation of the data uses propositional units. This identifies action objects and the actors previously discovered. If we used only the set of categories in Exhibit 15–3, the analysis would omit a considerable amount of information. The second analysis produces categories for action planning:

- Human relations.
- Production processes.
- Working conditions.
- Other action areas.
- No action area identified.

How can we categorize a response suggesting a combined management-production process? Exhibit 15–4 illustrates a combination of alternatives. By taking the categories of the first list with the action areas, it is possible to get an accurate frequency count of the joint classification possibilities for this question.

Using available software, the researcher can spend much less time coding open-ended responses and capturing categories. Software also eliminates the high cost of

EXHIBIT 15–3 Open Question Coding Example (before revision)
Question: "How can management-employee relations be improved?"

Locus of Responsibility	Mentioned	Not Mentioned
A. Management		
B. Union		
C. Worker (other than union)		
D. Joint management-union		
E. Joint management-workers		
F. Other		

EXHIBIT 15–4 Open Question Coding (after revision)
Question: "How can management-employee relations be improved?"

Locus of Responsibility	Frequency (n = 100)
A. Management	
1. Human relations	15
2. Production processes	45
3. Working conditions	25
4. Other action areas	10
5. No action area identified	5
B. Union	
1. Human relations	10
2. Production processes	10
3. Working conditions	65
4. Other action areas	15
5. No action area identified	0
C. Worker (other than union)	
D.	
E.	
F. Other	

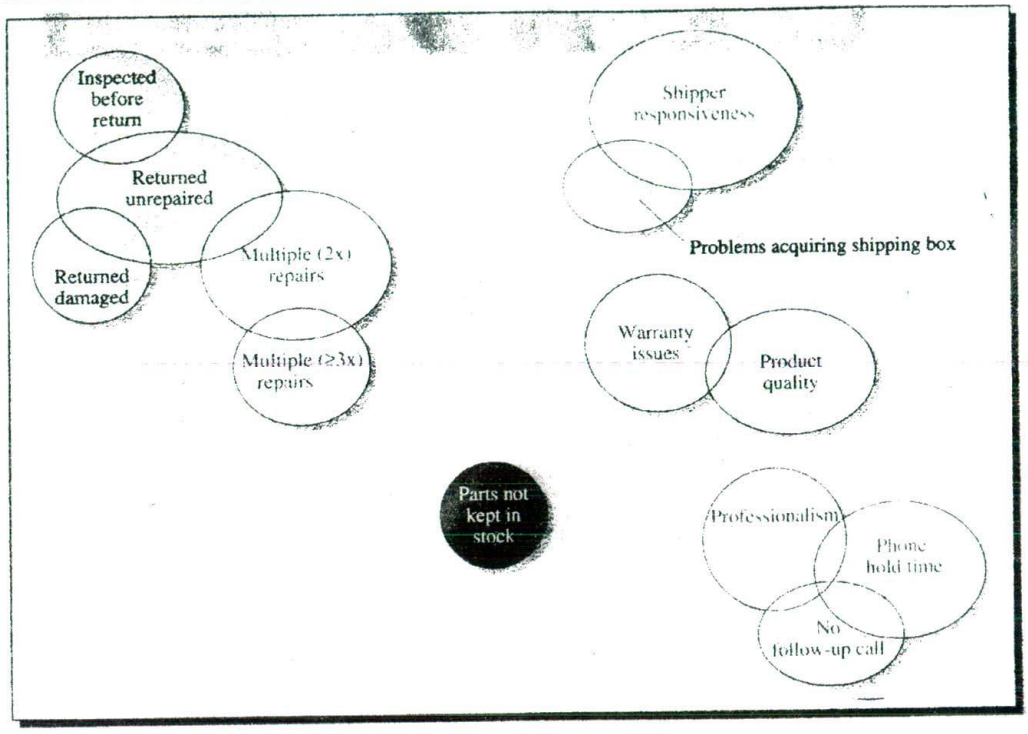
sending responses to outside coding firms. What would take a coding staff several days may now be done in a few hours.

By applying statistical algorithms to create categories from open-ended survey responses, various programs allow you to import responses in a tab-delimited (ASCII) file format and then automatically penetrate the chaos of diverse answers by stemming, aliasing, and excluding words that obscure important terms which are necessary to define meaningful categories.

- *Stemming* uses powerful linguistics technology to comb through responses, search for derivations of common root words, and combine terms to create stemmed aliases (search, searching, searches).
- *Aliasing* combines synonyms (smart, wise, intelligent) into automatic aliases.
- *Exclusion* filters for trivial words (be, is, the, of), to create a list of included terms valuable for constructing meaningful categories.[4]

When using menu-driven programs, an autocategorization option creates manageable categories by clustering terms that occur together throughout the data set. Then, with a few keystrokes, you can modify categorization parameters and refine your results. Once your categories are consistent with the research and investigative questions, you select what you want to export to a data file or in tab-delimited format. The output, in the form of tables and plots, serves as modules for your final report. Exhibit 15–5 shows a plot produced by a content analysis of MindWriter complaints data. The distances between pairs of terms reveal how likely the pair occurs together and the colors represent categories.

EXHIBIT 15–5 Proximity Plot of MindWriter Customer Complaints



"Don't Know" Responses

The **"don't know" (DK) response** presents special problems for data preparation. When the DK response group is small, it is not troublesome. But there are times when it is of major concern, and it may even be the most frequent response received. Does this mean the question that elicited this response is useless? The answer is, "It all depends." Most DK answers fall into two categories.[5] First, there is the legitimate DK response when the respondent does not know the answer. This response meets our research objectives; we expect DK responses and consider them to be useful.

In the second situation, a DK reply illustrates the researcher's failure to get the appropriate information. Consider the following illustrative questions:

1. Who developed the Managerial Grid concept?
2. Do you believe the new president's fiscal policy is sound?
3. Do you like your present job?
4. Which of the various brands of chewing gum do you believe has the best quality?
5. How often each year do you go to the movies?

It is reasonable to expect that some legitimate DK responses will be made to each of these questions. In the first question, the respondents are asked for a level of information that they often will not have. There seems to be little reason to withhold a correct answer if known. Thus, most DK answers to this question should be considered as legit-

imate. A DK response to the second question presents a different problem. It is not immediately clear whether the respondent is ignorant of the president's fiscal policy or knows the policy but has not made a judgment about it. The researchers should have asked two questions: In the first, they would have determined the respondent's level of awareness of fiscal policy. If the interviewee passed the awareness test, then a second question would have secured judgment on fiscal policy.

In the remaining three questions, DK responses are more likely to be a failure of the questioning process, although some will surely be legitimate. The respondent may be reluctant to give the information. A DK response to question 3 may be a way of saying, "I do not want to answer that question." Question 4 might also elicit a DK response in which the reply translates to, "This is too unimportant to talk about." In question 5, the respondents are being asked to do some calculation about a topic to which they may attach little importance. Now the DK may mean, "I do not want to do that work for something of so little consequence."

Dealing with Undesired DK Responses The best way to deal with undesired DK answers is to design better questions at the beginning. Researchers should identify the questions for which a DK response is unsatisfactory and design around it. Interviewers, however, often inherit this problem and must deal with it in the field. Several actions are then possible. First, good interviewer-respondent rapport will motivate respondents to provide more usable answers. When interviewers recognize an evasive DK response, they can repeat the question or probe for a more definite answer. The interviewer may also record verbatim any elaboration by the respondent and buck the problem on to the editor.

If the editor finds many undesired responses, little can be done unless the verbatim comments can be interpreted. Understanding the real meaning relies on clues from the respondent's answers to other questions. One way to do this is to estimate the allocation of DK answers from other data in the questionnaire. The pattern of responses may parallel income, education, or experience levels. Suppose a question concerning whether employees like their present jobs elicits the answers in Exhibit 15–6. The correlation between years of service and the "don't know" answers and the "no" answers suggests that most of the "don't knows" are disguised "no" answers.

There are several ways to handle "don't know" responses in the tabulations. If there are only a few, it does not make much difference how they are handled, but they will probably be kept as a separate category. If the DK response is legitimate, it should remain as a separate reply category. When we are not sure how to treat it, we should keep it as a separate reporting category and let the reader make the decision.

EXHIBIT 15–6 Handling "Don't Know" Responses

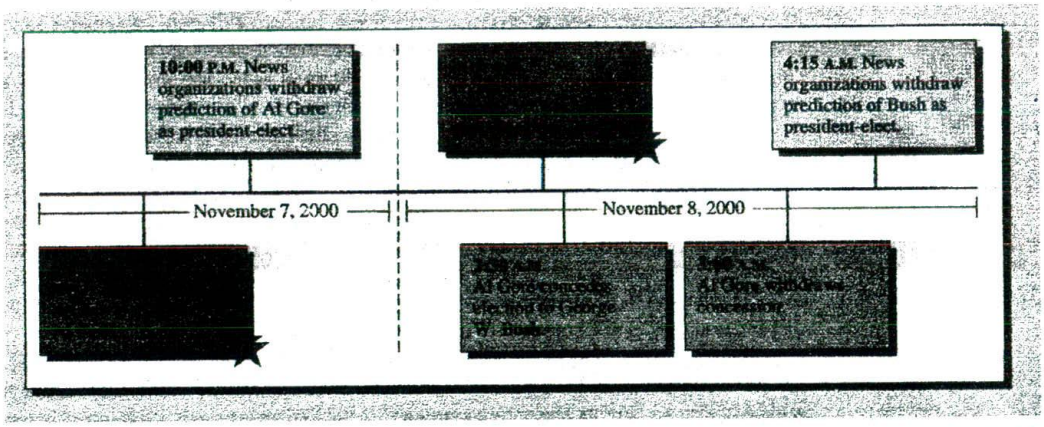
	Do You Like Your Present Job?		
Years of Service	Yes	No	Don't Know
Less than 1 year	10%	40%	38%
1–3 years	30	30	32
4 years or more	60	30	30
Total	100%	100%	100%
n =	650	150	200



SNAPSHOT VNS: A Black Eye for Research

A little before 8:00 P.M. on November 7, 2000, major news sources declared Al Gore the winner of Florida's electoral votes. Two hours later they pulled their prediction, reverting to "too close to call." At 2:15 A.M. on November 8, George W. Bush was declared the 43rd president and soon after Al Gore conceded the election. A month of Florida recounts gave the 2000 election its place in political, legal, news, and research history. Voter News Service (VNS), run by a consortium of ABC, NBC, CBA, CNN, FOX, and the Associated Press, was responsible for the exit polls. Exit polls, regarded as reliable indicators of voting results, conduct intercept interviews with voters as they leave voting booths. Voters provide who they voted for, why, and extensive demographic and psychographic information. How could an established polling organization and world-recognized news organizations' research fail so miserably? The answer can be traced to the following factors: (1) the use of a biased sample, with too many surveys completed in heavily Democratic precincts; (2) miscounted or misentered data in Duvall County (home of Jacksonville); and (3) a rush to judgment for the news value of an early "call" when statistically the margin between candidates was truly too close to decide. Antitrust advocates claim that collusion versus competition resulted in poor methodology and seek to break up VNS before another "campaign fiasco" occurs.

November 2000 Presidential Election: Timeline for Poll-Based Actions

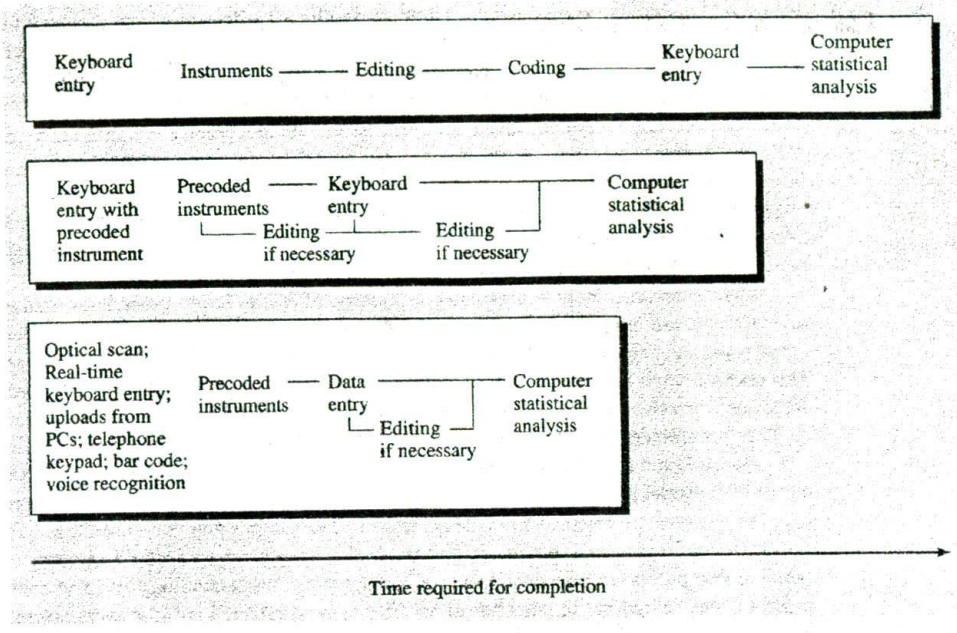


Another way to treat DK responses is to assume they occur almost randomly. Using this approach, we distribute them among the other answers in the same ratio that the other answers occur. This assumes that those who reply "don't know" are proportionally distributed among all of the groups studied. This can be achieved either by prorating the DK responses or by excluding all DK replies from the tabulation. The latter approach is better since it does not inflate the actual number of other responses.

Data Entry

Researchers have profited from new and more efficient ways of speeding up the research process (see Exhibit 15–7). **Data entry** converts information gathered by secondary or primary methods to a medium for viewing and manipulation. Keyboarding remains a mainstay for researchers who need to create a data file immediately and store it in a minimal space on a variety of media.

EXHIBIT 15–7 Methods of Data Entry



If you use a PC image scanner, you probably are familiar with **optical character recognition** programs that transfer printed text into computer files in order to edit and use it without retyping. There are other, related applications. **Optical scanning** instruments—the choice of testing services—are efficient for researchers. Examinees darken small circles, ellipses, or sets of parallel lines to choose a test answer. Optical scanners process the marked-sensed questionnaires and store the answers in a file. Depending on the scanner model (we are referring to dedicated units like NCS, Scantron, etc.), readers can process up to 8,000 forms per hour. This method, most often associated with standardized and preprinted forms, has been adopted by designers for data entry and preprocessing. It reduces the number of times data are handled, thereby reducing the number of errors that are introduced.

In a far more flexible format, **optical mark recognition (OMR)** uses a spreadsheet style interface to read and process user-created forms. The primary advantages are:

- **Speed:** About 10 times faster than manual keying.
- **Accuracy:** Error reduction from keying or stray marks.
- **Cost:** Savings on data entry, form design, and reproduction.
- **Convenience:** Makes viewing the data easy and provides charts and reports.

The researcher creates OMR forms with a word processor, page layout program, or survey design package and includes bubbles, fill-in marks, checkboxes, bar codes, and image fields. Data collected from questionnaires, reader reply cards, tests, interviews, observation forms, or checklists—virtually any paper-based medium—can be processed this way. A photocopier, laser printer, or local print shop then duplicates the forms. Special papers, drop-out inks, or the pencils associated with traditional OMR are not needed since the respondent may use markers, pens, pencils, or even crayons with the plain paper forms.

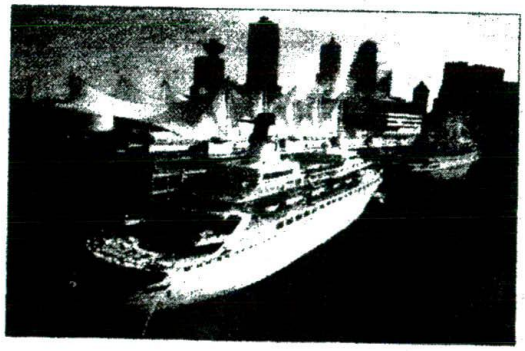
Most OMR products have "trainable" software, so your questionnaire need not conform to a particular design. Training involves creating a template by scanning in your form, displaying the image, and then, with a mouse, dragging boxes around the marking areas to define how each mark should be translated.[6] Once the template is created, a flatbed scanner can be used to read the data. For projects with hundreds or thousands of pages, a high-speed, sheet-fed scanner is essential. Scanned images of the collected data may be edited, saved, and processed with tabulation tools internal to the software. Many researchers export the results in multiple formats to statistical, database, and spreadsheet applications for more sophisticated analysis.

In addition to reading prescanned forms and questionnaire faxes, one program allows you to convert your paper-based OMR surveys into online surveys for the Internet or an intranet and then merge the results.[7]

Other techniques include direct response entry, of which voting procedures used in several states are an example. With a specially prepared punch card, citizens cast their votes by pressing a pen-shaped instrument against the card next to the preferred candidate. This opens a small hole in a specific column and row of the card. The cards are collected and placed directly into a card reader. This method also removes the coding and entry steps. Another governmental application is the 1040EZ form used by the Internal Revenue Service. It is designed for computerized number and character recognition. Similar character recognition techniques are employed for many forms of data collection. Again, both approaches move the response from the question to data analysis with little handling.

The declining cost of technology has allowed most researchers access to desktop or portable computers or networks to larger computers. This technology enables computer-assisted telephone or personal interviews to be completed with answers entered directly for processing, eliminating intermediate steps and errors.

The increase in computerized random-digit dialing has encouraged other data collection innovations. **Voice recognition** and response systems, while still far from mature, are providing some interesting alternatives for the telephone interviewer. Such systems can be used with software that is programmed to call specific three-digit pre-



By 2004 Princess Cruises will have 15 ships and 30,000 berths sailing 7 to 72 days to more than six continents on more than 150 itineraries. Princess carries more than 700,000 passengers each year and processes 245,000 customer satisfaction surveys each year—distributed to each cabin on the last day of each cruise. Princess uses scannable surveys rather than human data-entry for one reason: in the one-week to 10-day analysis lag created by human data-entry, 10 cruises could be completed with another 10 under way. For a business that prides itself on customer service, not knowing about a problem could be enormously damaging. Princess has found that scannable surveys generate more accurate data entry, while reducing processing and decision-response time—critical time in the cruise industry. www.princess.com

fixes and randomly generated four-digit numbers, reaching a random sample within a set geographical area. Upon getting a voice response, the computer branches into a questionnaire routine. Currently, the systems are programmed to record the verbal answers, but voice recognition is improving quickly enough so that these systems will soon translate voice responses into data files. Telephone keypad response is another capability made possible by computers linked to telephone lines. Using the telephone keypad (touch tone), the respondent answers questions by pressing the appropriate number. The computer captures the data by "listening," decoding the tone's electrical signal, and storing the numeric or alphabetic answer in a data file.

Field interviewers can use portable computers or electronic notebooks instead of clipboards and pencils. With a built-in communications modem or cellular link, their files can be sent directly to another computer in the field or to a remote site. This lets supervisors inspect data immediately or simplifies processing at a central facility.

From a crude bull's-eye code in 1948 to the adoption of the Universal Product Code (UPC) in 1973, the **bar code** developed from a technological curiosity to a business mainstay. The concept, originally modifying movie soundtrack technology and Morse code for bar elements and optical character recognition for scanning, was initially tested by the railroads on their gravel cars. The first actual bar code systems went into a General Motors plant to monitor the production and distribution of automobile axles and to the General Trading Company for directing shipments to the proper loading bays. After a 1970s study by McKinsey & Company that predicted unprecedented savings in the grocery industry, the Kroger grocery chain pilot-tested a production system and bar codes became ubiquitous in that industry.[8]

The bar code is split into two halves of six digits each. It is used in numerous applications: point-of-sale terminals, hospital patient ID bracelets, inventory control, product and brand tracking, promotional technique evaluation, shipping cartons, marathon runners, at rental car locations to speed the return of cars and generate invoices, and to track insects' mating habits. The military uses two-foot-long bar codes to label boats in storage. The codes appear on business documents, truck parts, and timber in lumberyards. Federal Express shipping labels use a code called Codabar. Improvements to the basic UPC include the European Article Numbering system, which has an extra pair of digits and may become the world's most widely used system. Other codes, containing letters as well as numbers, have potential for researchers.

Bar code technology is used to simplify the interviewer's role as a data recorder. Instead of writing (or typing) information about the respondents and their answers by hand, the interviewer can pass a bar code wand over the appropriate codes. The data are recorded in a small, lightweight unit for translation later. In the large-scale processing project Census 2000, the Census Data Capture Center used bar codes to identify residents.

Even with these time reductions between data collection and analysis, continuing innovations in multimedia technology are being developed by the personal computer business. The capability to integrate visual images, streaming video, audio, and data may soon replace video equipment as the preferred method for recording an experiment,



Data capture for Census 2000 used a bar code to connect the census question data with the responding household's address.

SNAPSHOT

Intelligence in Real Time

Historically, when Hewlett-Packard (HP) needed information about product performance, it waited for the product to hit the market in volume. Then, according to Beth Olson, market research manager for HP's product support division, HP would conduct time-consuming, expensive, traditional, one-on-one interviews. Recently, HP teamed with Decisive Technology, a firm that designs and implements Internet-based customer intelligence solutions for real-time customer feedback. Its product, MarketView, publishes the results immediately and continuously on a secure website. "With [Decisive's] MarketView, we were able to communicate very effectively with our 'wired' customers, obtaining information phenomenally fast, and in a format that allowed people throughout HP to access and act on the information," shared Olson.

www.hp.com

www.decisive.com

interview, or focus group. A copy of the response data could be extracted for data analysis, but the audio and visual images would remain intact for later evaluation.

Other techniques on the horizon will continue to improve research efficiency and effectiveness. Although technology will never replace researcher judgment, it can reduce data-handling errors, decrease time between data collection and analysis, and help provide more usable information.

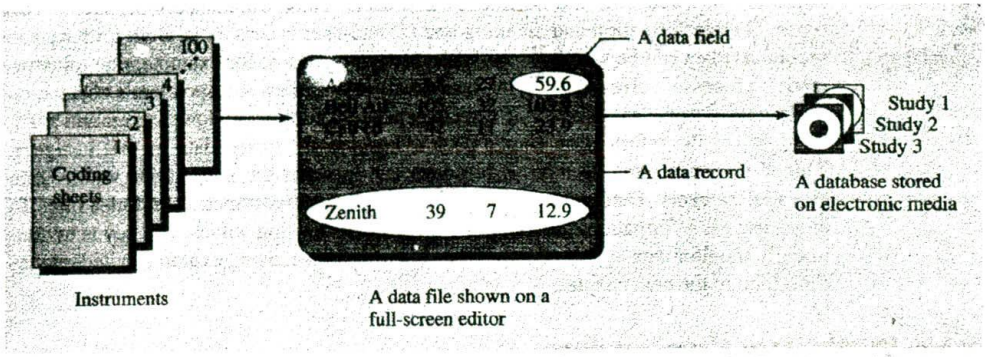
Data Entry Formats

A full-screen editor, where an entire data file can be edited or browsed, is a viable means of data entry for statistical packages like SPSS or SAS. The same software makes accessing data from databases, spreadsheets, data warehouses, or data marts effortless.

For large projects, database programs serve as valuable data entry devices. A **database** is a collection of data organized for computerized retrieval. Programs allow users to define data fields and link files so that storage, retrieval, and updating are simplified. The relationship between data fields, records, files, and databases is illustrated in Exhibit 15–8. A company's personnel records serve as an example of a database. Employee information may be kept in several files: salary and position, education, benefits, and home and family. The data are separated so that authorized people can see only those parts pertinent to their needs. However, the files may be linked so that when, say, a woman changes her name, the change is entered once and all the files are updated.

Researchers consider database entry when they have large amounts of potentially linked data that will be retrieved and tabulated in different ways over time. Another application of a database program is as a "front-end" entry mechanism. A telephone interviewer may ask the question "How many children live in your home?" The computer's software has been programmed to accept any answer between 0 and 20. If a "P" is accidentally struck, the program will not accept the answer and will return the interviewer to the question. With a precoded online instrument, much of the editing needed previously is done by the program. In addition, the program can be set for automatic conditional branching. In the example, an answer of 1 or greater causes the program to prompt the questioner to ask the ages of the children. A 0 causes the age question to be automatically skipped. Although this option is available whenever interactive computing is used, front-end processing is typically done within the database design. The database will then store the data into a set of linked files that allow the data to be easily sorted. Descriptive statistics and tables are readily generated from within the database.

EXHIBIT 15–8 Data Fields, Records, Files, and Databases

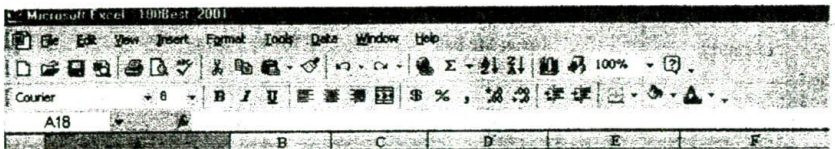


Data fields represent single elements of information (e.g., an answer to a question, a description, a number, a statement, etc.). Data fields can contain numeric, alphabetic, or symbolic information. A **record** is a set of data fields that are related. Records are the rows of a data file or spreadsheet program worksheet. **Data files** are sets of records that are grouped together for storage on diskettes, disks, tapes, CD-ROM, or optical disks. **Databases** are made up of one or more data fields that are interrelated.

SOURCES: Company data; public domain.

Spreadsheets are a specialized type of database. For data that need organizing, tabulating, and simple statistics, spreadsheets provide an easy-to-learn mechanism. They also offer some database management, graphics, and presentation capabilities. Data entry on a **spreadsheet** uses numbered rows and letter columns with a matrix of thousands of cells into which an entry may be placed. Spreadsheets allow you to type numbers, formulas, and text into appropriate cells. Many statistics programs for personal computers and also charting and graphics applications have data editors similar to the Excel™ spreadsheet format shown in Exhibit 15–9. This is a convenient and flexible means for entering and viewing the data.

EXHIBIT 15–9 Data Entry Using Spreadsheets



	A	B	C	D	E	F
1	100 Best Rank2001	Employ_US	Job Grow%	Trng_hr/yr	Ent_Sal_Prof	Revenues (Mil)
2	Edward Jones	25324	15	132	61690	2200
3	Container Store	1677	12	162	37400	225
4	SAS Institute	4309	14	32	49375	1120
5	TDIndustries	1368	7	32	45000	205
6	Synovus Financial	10995	4	52	41400	1395
7	Xilinx	2047	14	90	60000	1659
8	Plante & Moran	1240	1	60	40000	151
9	Qualcomm	5931	0	21	65000	2680
10	Alston & Bird	1338	18	40	100000	223
11	Baptist Health Care	4068	6	55	34798	647
12	Frank Russell	1068	9		51957	500
13	Hyperthera	566	5	30	45000	
14	CDW	2781	14	41	69500	3842
15	Fenwick & West	626	0	70	125000	146
16	Cisco Systems	26872	4		62000	22293

SOURCE: "The Best of 100 Companies to Work For," *Fortune*, February 4, 2002, pp. 72–80.

End-user tools for accessing data warehouses were discussed in Chapter 10.

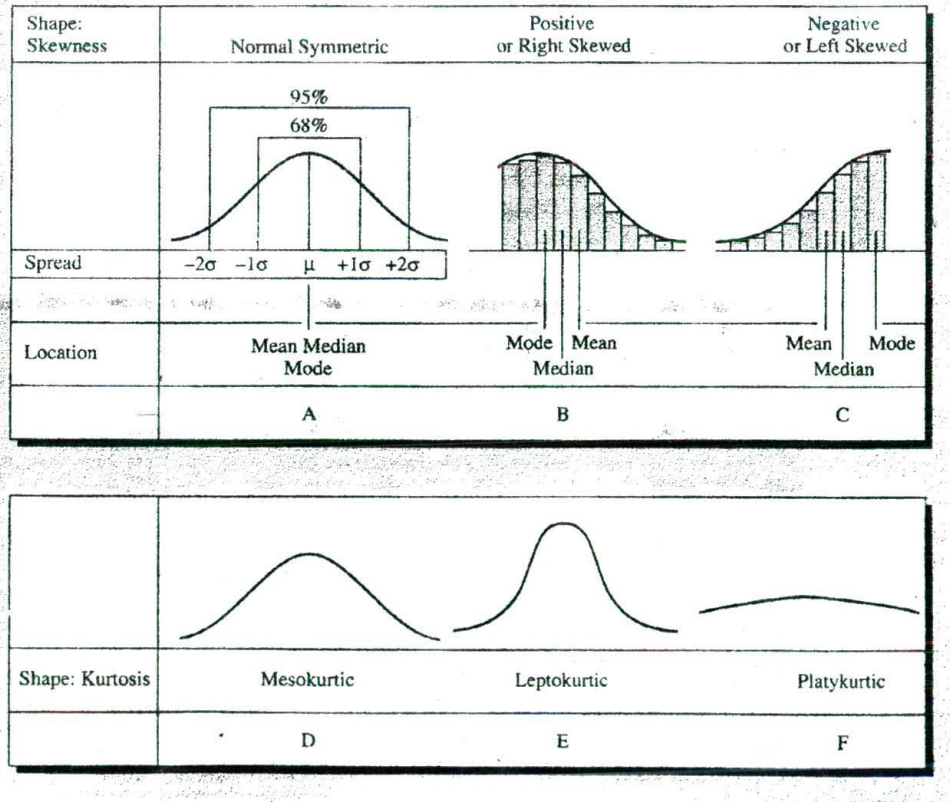
A **data warehouse** organizes large volumes of data into categories to facilitate retrieval, interpretation, and sorting by end-users. Data warehouse tools are of two types: (1) transformation and cleaning and (2) end-user access tools. Both primary and secondary data can be warehoused. With survey responses, for example, one transformation process involves *binning* the classification variables. A respondent's occupation could have a value from 1 to 20. If each occupational code was allowed to remain a discrete value, the information might be lost because of sparse data in that category. Binned responses provide more data points and increase the information within each essential category. Data quality is a problem during transformation and cleaning. To avoid useless information, the data should have few missing values. The key is to continually monitor data as it is added to the data warehouse, using exploratory techniques described in the next chapter.

Descriptive Statistical Summaries

Describing Distributions

In the first part of the chapter, we discussed how responses are coded and entered. Creating numerical summaries of this process provides valuable insights into its effectiveness. For example, **missing data,** information that is missing about a respondent or case

EXHIBIT 15–10 Characteristics of Distributions



for which other information is present, may be detected. Miscoded, out-of-range data, extreme values, and other problems also may be rectified after a preliminary look at the data set.[9] To understand each variable's characteristics, first consider the type of scale on which it was measured. With nominal measurements (e.g., company classifications like high-tech, financial, and retailing), each category is represented by a numerical code that refers back to a verbal description of the category. With ordinal data, the item's rank, reflecting a position in the range from the lowest to the highest, is entered. The same is true with interval-ratio scores. When these data are tabulated, they may be arrayed from the lowest to the highest scores on their scales. Together with the frequency of occurrence, the observations form a **distribution** of values.

Many variables of interest have distributions that approximate a **standard normal distribution.** This distribution, shown in Part A of Exhibit 15–10, is the most significant theoretical distribution in statistics. It is a standard of comparison for describing distributions of sample data and is used with inferential statistics that assume normally distributed variables.

Examine the sample distribution of variables from the MindWriter dataset shown in Exhibit 15–11. These data were collected on a five-point interval scale. There are no missing data in variable 1A, although it is apparent that a range of 6 and a maximum

EXHIBIT 15–11 MindWriter Dataset: Missing and Out-of-Range Data

Case	1A	1B	2A	2B	2C
1	5.0	5.0	5.0	5.0	9.0
2	7.0	3.0		4.0	9.0
3	5.0	5.0	5.0	5.0	5.0
4	5.0	5.0	4.0		
5	1.0			2.0	
6	5.0	5.0	5.0	5.0	9.0
7	5.0	5.0	5.0	5.0	5.0
8	4.0	3.0	3.0	3.0	3.0
9	4.0	4.0	5.0	5.0	5.0
10	4.0	5.0		4.0	5.0
11	2.0	5.0	4.0	4.0	5.0
12	6.0	4.0	3.0	3.0	4.0
13	5.0	5.0		3.0	5.0
14	5.0	5.0	5.0	5.0	5.0
15	5.0	4.0	5.0	5.0	4.0
Valid	15	14	11	14	13
Missing	0	1	4	1	2
Mean	4.53	4.50	4.45	4.14	5.61
Range	6	2	2	3	6
Minimum	1	3	3	2	3
Maximum	7	5	5	5	9

value of 7 invalidate the calculated mean or average score. Variables 1B and 2B have one case missing but values that are within range. Variable 2A is missing four cases, or 27 percent of its data points. The last variable, 2C, has a range of 6, two missing values, and three values coded as "9." A "9" is often used as a DK or missing value code when the scale has a range less than nine points. In this case both blanks and 9s are present—a coding concern. Notice that the fifth respondent answered only two of the five questions and the second respondent had two miscoded answers and one missing value. Finally, using descriptive indexes of shape, discussed later in this section, you can find three variables that depart from the symmetry of the normal distribution. They are skewed (or pulled) to the left by a disproportionately small number of 1s and 2s. And, one variable's distribution is peaked beyond normal dimensions. We have just used the minimum and maximum values, the range, and mean and have already discovered errors in coding, problems with respondent answer patterns, and missing cases.

Now let's look at some other descriptive tools for this purpose. The characteristics of location, spread, and shape are helpful initial tools for cleaning the data, discovering problems, and summarizing distributions. Their definitions, applications, and formulas fall under the heading of **descriptive statistics.** Although the definitions will be familiar to most readers, we take the following perspective on the characteristics of distributions:

We illustrate techniques for visualizing data patterns in Chapter 16.

- A distribution's shape is just as consequential as its location and spread.
- The choice of summary statistics to describe a singe variable is contingent on the appropriateness of those statistics for the shape of the distribution.
- Visual representations are ultimately superior to numerical ones for discovering a distribution's shape and should be used before selecting remedies to correct anomalies in the data.[10]

Measures of Location

Refer to Exhibit 15–9 for the values of the variables in these examples.

Summarizing the information from our collected data often requires the description of "typical" values. The data in Exhibit 15–9 are a subset of the 2002 *Fortune* magazine article "The 100 Best Companies to Work For." Suppose we want to know the typical score for companies' job growth (%). We might define *typical* as the average response (mean); the middle value, when the distribution is sorted from lowest to highest (median); or the most frequently occurring value (mode). The common **measures of location,** often called **central tendency** or *center,* include the mean, median, and mode.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

For the job growth variable in Exhibit 15–9, the distribution of responses is: 15, 12, 14, 7, 4, 14, 1, 0, 18, 6, 9, 5, 14, 0, 4. The arithmetic average, or **mean,** of the 15 values is: 15 + 12 + 14 + 7 + 4 + 14 + 1 + 0 + 18 + 6 + 9 + 5 + 14 + 0 + 4 / 15 = 8.2 (or, an average of 8.2% job growth). Note that 2 of the top 15 companies had zero job growth.

The **median** is the midpoint of the distribution. Half of the observations in the distribution fall above and the other half fall below the median. When the distribution has an even number of observations, the median is the average of the two middle scores. The median is the most appropriate locator of center for ordinal data and has resistance to extreme scores, thereby making it a preferred measure for interval-ratio data when their distributions are not normal. The median is sometimes symbolized by *M* or *mdn.*

From the sample distribution for the U.S. employees variable, the median of the 15 values is 2047. The distribution is ordered: 566, 626, 1068, 1240, 1338, 1368, 1677,

2047, 2781, 4068, 4309, 5931, 10995, 25324, and 26872; since the total number of values is 15, the midpoint is the middle value of this distribution. If the distribution had 16 values, the median would be the average of the values for the 7th and 8th cases.

The **mode** is the most frequently occurring value. When there is more than one score that has the highest yet equal frequency, the distribution is bimodal or multimodal. When every score has an equal number of observations, there is no mode. The mode is the location measure for nominal data and a point of reference along with the median and mean for examining spread and shape. In our example, the most frequently occurring value for the training hours/year variable is 32.

Measures of Spread

The common **measures of spread,** alternatively referred to as *dispersion* or **variability,** are the variance, standard deviation, range, interquartile range, and quartile deviation. They describe how scores cluster or scatter in a distribution.

The **variance** is the average of the squared deviation scores from the distribution's mean. It is a measure of score dispersion about the mean. If all the scores are identical, the variance is 0. The greater the dispersion of scores, the greater the variance. Both the variance and the standard deviation are used with interval-ratio data. The symbol for the sample variance is s^2, and the population variance is the Greek letter sigma, squared (σ^2). The variance is computed by summing the squared distance from the mean for all cases and dividing the sum by the total number of cases minus one.

$$s^2 = \sum_{i=1}^{n} \frac{(X_1 - \bar{X})^2}{n-1}$$

For the percent of job growth variable, we would compute the variance as:

$$s^2 = (15 - 8.2)^2 + (12 - 8.2)^2 + \ldots + (4 - 8.2)^2/14 = 5.96$$

The **standard deviation** summarizes how far away from the average the data values typically are. It is perhaps the most frequently used measure of spread because it improves interpretability by removing the variance's square and expressing deviations in their original units (e.g., revenues in dollars, not dollars squared). It is also an important concept for descriptive statistics because it reveals the amount of variability of individuals within the dataset. Like the mean, the standard deviation is affected by extreme scores. The symbol for the sample standard deviation is *s*, and a population standard deviation is σ. Alternatively, it is labeled *std. dev.*

$$s = \sqrt{s^2}$$

The standard deviation for the percent of job growth variable in our example is 5.96.

The **range** is the difference between the largest and smallest score in the distribution. The revenues (millions) variable has a range of 22147 (22293 – 146). Unlike the standard deviation, it is computed from only the minimum and maximum scores; thus, it is a very rough measure of spread. With the range as a point of comparison, it is possible to get an idea of the homogeneity (small std. dev.) or heterogeneity (large std. dev.) of the distribution. For homogeneous distribution, the ratio of the range to the standard deviation should be between 2 and 6. In the revenues example, the ratio is 22147 / 5759.7 = 3.84. A number above 6 would indicate a high degree of heterogeneity. The range provides useful but limited information for all data. It is mandatory for ordinal data.

Quartiles are used to build boxplots of the data. Their construction is described in Chapter 16.

The **interquartile range (IQR)** is the difference between the first and third quartiles of the distribution. It is also called the *midspread.* Ordinal or ranked data use this

measure in conjunction with the median. It is also used with interval-ratio data when asymmetrical distributions are suspected or for exploratory analysis. Recall the following relationships. The minimum value of the distribution is the 0th percentile; the maximum, the 100th percentile. The first quartile (Q_1) is the 25th percentile; the median, or Q_2, is the 50th percentile. The third quartile (Q_3) is the 75th percentile.

The **quartile deviation,** or semi-interquartile range, is expressed as

$$Q = \frac{Q_1 - Q_3}{2}$$

The quartile deviation is always used with the median for ordinal data. It is helpful for interval-ratio data of a skewed nature. In a normal distribution, the median plus one quartile deviation (Q) on either side encompasses 50 percent of the observations. Eight Qs cover approximately the range. Q's relationship with the standard deviation is constant ($Q = .6745s$) when scores are normally distributed. For the percent of job growth variable, Q is: 14 – 4 / 7 = 1.43.

Measures of Shape

The **measures of shape,** *skewness* and *kurtosis,* describe departures from the symmetry of a distribution and its relative flatness (or peakedness), respectively. They are related to statistics known as *moments,* which use deviation scores ($X - \bar{X}$). The variance, for example, is a second power moment. The measures of shape use third and fourth power deviations for their computations and are often difficult to interpret when extreme scores are in the distribution. Generally, shape is best communicated through visual displays. From a practical standpoint, the calculation of skewness and kurtosis is best with spreadsheet or statistics software.

Skewness is a measure of a distribution's deviation from symmetry. In a symmetrical distribution, the mean, median, and mode are in the same location. A distribution that has cases stretching toward one tail or the other is called *skewed.* As shown in Exhibit 15–10, when the tail stretches to the left, to smaller values, it is negatively skewed. Scores stretching toward the right, toward larger values, skew the distribution positively. Note the relationship between the mean, median, and mode in asymmetrical distributions. The mean and standard deviation are called *dimensional* measures. That is, they are expressed in the same units as the measured quantities. In contrast, skewness is considered a nondimensional measure because it is an index which only characterizes the shape of the distribution. The symbol for skewness is *sk.*

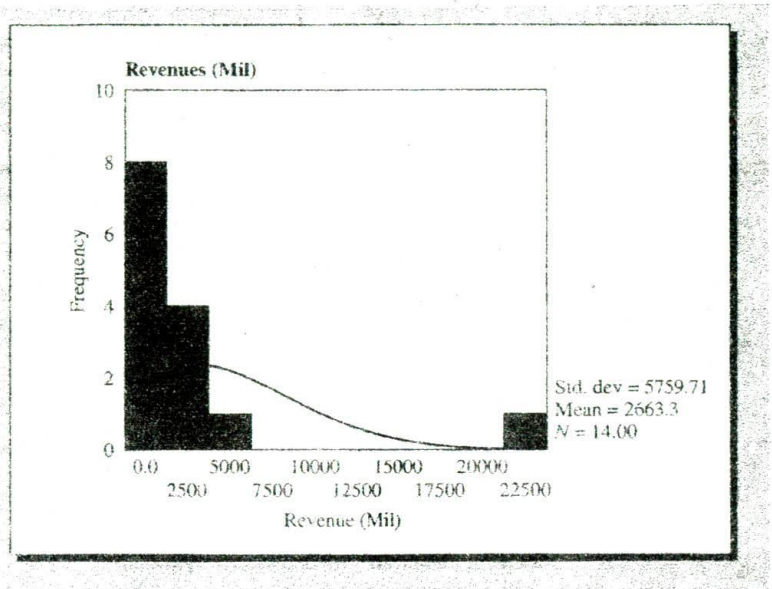
$$sk = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where s is the sample standard deviation (the unbiased estimate of sigma).

When a distribution approaches symmetry, *sk* is approximately 0. With a positive skew, *sk* will be a positive number; with negative skew, *sk* will be a negative number. The calculation of skewness for our sample revenue data produces an index of 3.5 and reveals a positive skew (see Exhibit 15–12).

As illustrated in the lower portion of Exhibit 15–10, **kurtosis** is a measure of a distribution's peakedness (or flatness). It is also a nondimensional index. Distributions that have scores which cluster heavily or pile up in the center (along with more observations than normal in the extreme tails) are peaked or *leptokurtic.* Flat distributions, with scores more evenly distributed and tails fatter than a normal distribution, are called *platykurtic.* Intermediate or *mesokurtic* distributions are neither too peaked nor too flat. The symbol for kurtosis is *ku.*

EXHIBIT 15–12 Shape Characteristics in Spreadsheet Variable Revenue



$$ku = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

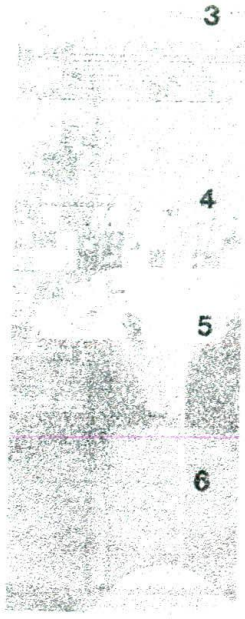
where s is the sample standard deviation (the unbiased estimate of sigma).

The value of *ku* for a normal or mesokurtic distribution is close to 0. A leptokurtic distribution will have a positive value, and the platykurtic distribution will be negative. As with skewness, the larger the absolute value of the index, the more extreme is the characteristic. In the revenues example, the kurtosis is calculated as +12.7, which suggests a strong deviation from a normally shaped curve with peaking contributed by greater frequency of values in the range of 146 to 647 million (see Exhibit 15–12).

SUMMARY

1 The first step in data preparation is to edit the collected raw data to detect errors and omissions that would compromise quality standards. The editor is responsible for making sure the data are accurate, consistent with other data, uniformly entered, and ready for coding. In survey work, it is common to use both field and central editing.

2 Coding is the process of assigning numbers and other symbols to answers so we can classify the responses into categories. Categories should be appropriate to the research problem, exhaustive of the data, mutually exclusive, and unidimensional. The reduction of information through coding requires the researcher to design category sets carefully, using as much of the data as possible. Codebooks are guides to reduce data entry error and serve as a compendium of variable locations and other information for the analysis stage.



3 Closed questions include scaled items and other items for which answers are anticipated. Precoding of closed items avoids tedious completion of coding sheets for each response. Open questions are more difficult to code since answers are not prepared in advance, but they do encourage disclosure of complete information. A systematic method for analyzing open questions is called content analysis. It uses preselected sampling units to produce frequency counts and other insights into data patterns.

4 "Don't know" replies are evaluated in light of the question's nature and the respondent. While many DKs are legitimate, some result from questions that are ambiguous or from an interviewing situation that is not motivating. It is better to report DKs as a separate category unless there are compelling reasons to treat them otherwise.

5 Data entry is accomplished by keyboard entry from precoded instruments, optical scanning, real-time keyboarding, telephone pad data entry, bar codes, voice recognition, OMR, and data transfers from electronic notebooks and laptop computers. Database programs, spreadsheets, and editors in statistical software programs offer flexibility for entering, manipulating, and transferring data for analysis, warehousing, and mining.

6 The objective of descriptive statistical analysis is to develop sufficient knowledge to describe a body of data. This is accomplished by understanding the data levels for the measurements we choose, their distributions, and characteristics of location, spread, and shape. The discovery of miscoded values, missing data, and other problems in the dataset is enhanced with descriptive statistics.

KEY TERMS

bar code 469
central tendency 474
codebook 457
coding 456
content analysis 460
database 470
data entry 466
data fields 471
data files 471
data warehouse 472
descriptive statistics 474
distribution 473
"don't know" (DK) response 464
editing 455
interquartile range (IQR) 475
kurtosis 476
mean 474
measures of location 474
measures of shape 476
measures of spread 475
median 474
missing data 472
mode 475
optical character recognition 467
optimal mark recognition (OMR) 467
optical scanning 467
precoding 459
quartile deviation 476
range 475
record 471
skewness 476
spreadsheet 471
standard deviation 475
standard normal distribution 473
variability 475
variance 475
voice recognition 486

EXAMPLES

Company	Scenario	Page
BrainReserve	A trend-spotting organization that uses content analysis.	460
Campbell Soup Company	Content analysis is used to structure the promotional message for V-8.	460
CityBus	Using competitive bidding to acquire data-entry personnel.	453
Decisive Technology	Developed MarketView software that allows Hewlett-Packard to quickly and less expensively evaluate new products online.	470

Federal Express	Use of bar codes (called Codabar) on shipping labels.	469
Fortune	Evaluates the 100 best companies to work for in a tough year; what makes a great workplace.	474
General Motors	Development of bar coding for inventory analysis.	469
General Trading Company	Use of bar coding to direct shipments to appropriate loading bays.	469
Hewlett-Packard	How online software from Decisive Technology provides quicker, less costly evaluation of new products.	470
Inferential Focus	A trend-spotting organization that uses content analysis.	460
Kroger	Pioneering use of bar code data in the grocery industry.	469
MindWriter*	Content analysis of customer complaints; evaluation of missing data with statistics.	463
Naisbitt Group	A trend-spotting organization that uses content analysis.	460
PhaseOne Communications	Conducted content analysis of V-8 ads for Campbell Soup, resulting in a more effective promotional campaign.	460
Princess Cruises P&O Princess Cruises plc	Quick turnaround for a customer satisfaction study using optical scanning for data entry.	468
SAS	A statistical package for data entry and analysis.	470
SPSS, Inc.	A statistical package for data entry and analysis.	470
Stanford Research Institute (SRI) International	A research organization that uses content analysis.	460
U.S. Census Bureau	Using bar-coding to provide tracking of participation and faster data analysis.	469
U.S. Internal Revenue Service	Use of computerized number and character recognition to process 1040EZ forms.	468
Voter News Service (VNS)	The election polling consortium of ABC, NBC, CBA, CNN, FOX, and the Associated Press, responsible for the early mispredictions in the 2000 presidential election.	466

*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. Coding rules.
   b. Spreadsheet data entry.
   c. Bar codes.
   d. Precoded instruments.
   e. Measures of shape.
   f. Content analysis.
   g. Missing values.

h. Optical mark recognition.

i. Measures of spread.

2. How should the researcher handle "don't know" responses?

3. Why is the standard deviation a more useful statistic than the variance?

Making Research Decisions

4. A problem facing shoe store managers is that many shoes eventually must be sold at markdown prices. This prompts us to conduct a mail survey of shoe store managers in which we ask, "What methods have you found most successful for reducing the problem of high markdowns?" We are interested in extracting as much information as possible from these answers to better understand the full range of strategies that store managers use. Establish what you think are category sets to code 500 responses similar to the 14 given below. Try to develop an integrated set of categories that reflects your theory of markdown management. After developing the set, use it to code the 14 responses.

a. Have not found the answer. As long as we buy style shoes, we will have markdowns. We use PMs on slow merchandise, but it does not eliminate markdowns. (PM stands for "push-money"—special item bonuses for selling a particular style of shoe.)

b. Using PMs before too old. Also reducing price during season. Holding meetings with salespeople indicating which shoes to push.

c. By putting PMs on any slow-selling items and promoting same. More careful check of shoes purchased.

d. Keep a close watch on your stock, and mark down when you have to—that is, rather than wait, take a small markdown on a shoe that is not moving at the time.

e. Using the PM method.

f. Less advance buying—more dependence on in-stock shoes.

g. Sales—catch bad guys before it's too late and close out.

h. Buy as much good merchandise as you can at special prices to help make up some markdowns.

i. Reducing opening buys and depending on fill-in service. PMs for salespeople.

j. Buy more frequently, better buying, PMs on slow-moving merchandise.

k. Careful buying at lowest prices. Cash on the buying line. Buying closeouts, FDs, overstock, "cancellations." (FD stands for "factory-discontinued" style.)

l. By buying less "chanceable" shoes. Buy only what you need, watch sizes, don't go overboard on new fads.

m. Buying more staple merchandise. Buying more from fewer lines. Sticking with better nationally advertised merchandise.

n. No successful method with the current style situation. Manufacturers are experimenting, the retailer takes the markdowns—cuts gross profit by about 3 percent—keep your stock at lowest level without losing sales.

5. Define a small sample of class members, work associates, or friends and ask them to answer the following in a paragraph or two: "What are your career aspirations for the next five years?" Use one of the four basic units of content analysis to analyze their responses. Describe your findings as frequencies for the unit of analysis selected.

6. What is the median of the distribution 123, 154, 160, 187?

7. What happens to the mean and median in a set of five scores when the largest one is increased by several points? Set A: 12, 13, 23, 32, 43; Set A altered: 12, 13, 23, 32, 143.

From Concept to Practice

8. Select the variables in Exhibit 15–9 and enter their values into a spreadsheet.
   a. Compute their means, medians, and modes. Which variables have similar means and medians?
   b. Compute the standard deviations. Can you find any variables with proportionately larger standard deviations? What can you infer about the distributions' shapes?
   c. Now compute skewness and kurtosis. Which variables have the least skewness? The most kurtosis?

WWW Exercises

Visit our website for Internet exercises related to this chapter at www.mhhe.com/business/cooper8

CASES*

AGRICOMP

AIDS RATES FOR FEMALES

HEALTHY LIFESTYLES

INQUIRING MINDS WANT TO KNOW—NOW!

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. Jean M. Converse and Stanley Presser, *Survey Questions: Handcrafting the Standardized Questionnaire* (Beverly Hills, CA: Sage Publications, 1986), pp. 34–35.
2. B. Berelson, *Content Analysis in Communication Research* (New York: Free Press, 1952), p. 18.
3. Klaus Krippendorff, *Content Analysis: An Introduction to Its Methodology* (Beverly Hills, CA: Sage Publications, 1980), p. 22.
4. Based on the operation of the SPSS, Inc., product TextSmart.
5. Hans Zeisel, *Say It with Figures*, 6th ed. (New York: Harper & Row, 1985), pp. 48–49.
6. See the description of Remark Office OMR® found at: http://www.principiaproducts.com.
7. Remark Web Survey® is a product of Principia Products, Inc., 16 Industrial Blvd., Suite 102, Paoli, PA, 19301.
8. Adapted from a history of bar code development: http://www.lascofittings.com/BarCode-EDI/bc-history.htm.
9. For a thorough discussion of missing data analysis and remedies in multivariate data, see Joseph F. Hair, Jr., et al., *Multivariate Data Analysis*, 5th ed. (Upper Saddle River, NJ: Prentice Hall, 1998), pp. 46–64.
10. Frederick Hartwig, with Brian E. Dearing, *Exploratory Data Analysis* (Beverly Hills, CA: Sage Publications, 1979), p. 15.

REFERENCES FOR SNAPSHOTS AND CAPTIONS

Decisive/Hewlett-Packard

"Hewlett-Packard Uses Decisive Technology's MarketView℠ to Analyze and Enhance Customers' After-Purchase Experiences," press release, Decisive Technology, April 6, 1999 (http://www.decisive.com/html/news/hp.htm).

Princess Cruises

"About Princess," P&O Princess Cruises plc, 02/09/2002 (http://www.princess.com/about).

J. Rydholm, "Scanning the Seas," *Quirk's Market Research Review*, May 1993 (http://www.quirks.com/articles/article_print.asp?arg_articleid=207).

V-8

"Good-Bye Guesswork," The Advertising Education Foundation, 1995 (http://www.campbellsoup.com/OurBrands/v8juice.cfm).

VNS

"A Bad Day at the Exit Polls," Savanna Morning News, November 9, 2000 (http://www.savannamorningnews.com/sma/stories/110900/locsurvey.shtml).

"Antitrust Group Urges U.S. to Break Up Voter News Service," Anti-Trust Institute, November 27, 2000 (http://www.antitrustinstitute.org/recent/90.cfm).

"Internet Exit Poll Leaks Election Day," ABC News, November 7, 2000 (http://www.abcnews.go...News/ELECTION_exitpolls001107.html).

Poll results at (www.cnn.com/election/2000/epolls/us/p000.html) or (www.abcnews.go.com/sections...2000vote/general/exitpoll_hub.html).

CLASSIC AND CONTEMPORARY READINGS

Aczel, Amir D., and Jayauel Sounderpandian. *Complete Business Statistics*. 5th ed. Chicago: Irwin/McGraw-Hill, 2001. See Chapter 1 on descriptive statistics and Chapter 4 on the normal distribution.

Bux, William E., and Kenneth L. Gorman. *DataEntry Activities for Windows*. Cincinnati, OH: Thomson Learning, 2000. Fundamentals of data entry in a self-paced learning package suitable for novice computer users.

Foster, Jeremy. *Data Analysis Using SPSS® for Windows® Version 8 to 10: A Beginner's Guide*. Thousand Oaks, CA: Sage Publishing, 2001. The first 15 chapters deal with data management: merging, sorting, coding, grouping, and fundamentals of descriptive analysis.

Strauss, Anselm, and Juliet Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: Sage Publishing, 1998. Part II (10 chapters) illustrates several approaches to coding procedures for qualitative data.

Zeisel, Hans. *Say It with Figures*. 6th ed. New York: Harper & Row, 1985. The entire book is worth reading for its excellent discussion of numerical presentation.

CHAPTER 16

Exploring, Displaying, and Examining Data

Learning Objectives

After reading this chapter, you should understand . . .

1 **That exploratory data analysis techniques provide better diagnostics than traditional summary statistics by emphasizing visual representations of the data.**

2 **How statistical process control charts offer understandable and reliable displays for evaluating point values, trends, and special causes of variation in a process.**

3 **The uses of Geographic Information Systems for mapping, detecting patterns in the data, and discovering relationships between different sets of spatially referenced information.**

4 **How cross-tabulation is used to evaluate relationships involving categorical variables, serves as a framework for later statistical testing, and makes table-based analysis using one or more control variables an efficient tool for decision making.**