




## DATA PROCESSING AND FUNDAMENTAL DATA ANALYSIS

### LEARNING OBJECTIVES

- 1. To get an overview of the data analysis procedure.
- 2. To develop an understanding of the importance and nature of quality control checks.
- 3. To understand the data entry process and data entry alternatives.
- 4. To learn how surveys are tabulated.
- 5. To learn how to set up and interpret crosstabulations.
- 6. To comprehend the basic techniques of statistical analysis.



Stephanie Benson, of Technology Decisions, is the firm's account executive for Dell Computer. She recently submitted a proposal to Dell for a project involving the processing of 20,000 to 25,000 questionnaires to be collected by Dell personnel from attendees at a series of high-tech trade shows over the next 6 months. On this project, she will be working directly with the manager from the sales group responsible for Dell's trade show activities, Jill Jackson. Benson did not take this into account when she wrote her proposal. Normally, she worked with marketing research department staff, who would interface between her and the managers for whom the research was being done. Knowing that these marketing researchers were well acquainted with editing, coding, data entry, and tabulation procedures, she did not cover those topics in any depth in her proposal.

She has just received a lengthy e-mail from Jackson who says she likes the price quoted and the sample report included in the proposal. However, Benson can see that Jackson is a process-oriented person who wants lots of details regarding how various things will be done. Jill Jackson's questions, taken from her e-mail, follow:

- ❑ Will the questionnaires be checked for logical consistency, accuracy, and completeness before they are entered into electronic files? How will this be done? What quality checks are built into this process?
- ❑ I'm assuming that no data entry will be done until questionnaires have been checked as suggested above. Is that assumption correct?
- ❑ As you know from the sample questionnaire, the survey has seven open-ended questions. This information is very important to us. We intend to use feedback from the trade show attendees—they are either customers or people we would like to have as customers—to guide us in developing a number of sales and marketing initiatives. Therefore, it is important that we get an accurate and complete summarization of these comments. Obviously, there are far too many questionnaires for us to read and somehow summarize ourselves. In your proposal, you refer to the "coding" of open-ended questions. I have only a very general idea of what that means. What does it mean to "code" open-ended questions? How do you go about it? Outline the process. What quality control checks are built into the process? Finally, can we [management] have some input in shaping how comments are coded?
- ❑ In your proposal, you say you will enter data from the paper questionnaires after completion of the coding process. I'm assuming that you're talking about transferring the data from the paper questionnaire to an electronic file. How will this be done? What quality control procedures are built into this process so that I can be assured that the data in the electronic file accurately reflect the original responses on the paper questionnaires?
- ❑ You refer to cross tabulations in your proposal. What are cross tabulations, and how are they produced? I know they are tables of some sort. Can we have input into the design of those tables?
- ❑ Finally, is there some way that we could have access to our data over the Internet? Having access to the tables would be okay, but we really would like to be able to have access to the data and a tool that would permit us to generate any tables that we might want to produce. Is this possible?

This chapter will offer answers to Jackson's questions by providing all of the background and tools needed to perform these important tasks. The seemingly mechanical data processing activities are a critical bridge between the data-collection and data analysis phases of a project. ■



You should be able to give her the answers after you read this chapter.

## Overview of the Data Analysis Procedure

Once data collection has been completed and questionnaires have been returned, the researcher may be facing anywhere from a few hundred to several thousand interviews, each ranging from a few pages to 20 or more pages. We recently completed a study involving 1,300 questionnaires of 10 pages each. The 13,000 pages amounted to a stack of paper nearly 3 feet high. How should a researcher transform all the information contained on 13,000 pages of completed questionnaires into a format that will permit the summarization necessary for detailed analysis? At one extreme, the researcher could read all the interviews, make notes while reading them, and draw some conclusions from this review of the questionnaires. The folly of this approach is fairly obvious. Instead of this haphazard and inefficient approach, professional researchers follow a five-step procedure for data analysis:

- Step One. Validation and editing (quality control)
- Step Two. Coding
- Step Three. Data entry
- Step Four. Machine cleaning of data
- Step Five. Tabulation and statistical analysis

## Step One: Validation and Editing


The purpose of the first step is twofold. The researcher wants to make sure that all the interviews actually were conducted as specified (validation) and that the questionnaires have been filled out properly and completely (editing).

### Validation

First, the researcher must determine, to the extent possible, that each of the questionnaires to be processed represents a valid interview. Here, we are using the term *valid* in a different sense than in Chapter 9. In Chapter 9, *validity* was defined as the extent to which what was being measured was actually measured. In this chapter, **validation** is defined as the process of ascertaining that interviews were conducted as specified. In this context, no assessment is made regarding the validity of the measurement. The goal of validation is solely to detect interviewer fraud or failure to follow key instructions. You may have noticed that the various questionnaires presented throughout the text almost always have a place to record the respondent's name, address, and telephone number. This information is seldom used in any way in the analysis of the data; it is collected only to provide a basis for validation.

Professional researchers know that interviewer cheating does happen. Various studies have documented the existence and prevalence of interviewer falsification of several types. For this reason, validation is an integral and necessary step in the data processing stage of a marketing research project. See the Practicing Marketing Research feature on page 435 for some ways to address this problem.

After all the interviews have been completed, the research firm recontacts a certain percentage of the respondents surveyed by each interviewer. Typically, this percentage ranges from 10 to 20 percent. If a particular interviewer surveyed 50 people and the research firm

 **validation**  
Process of ascertaining that interviews actually were conducted as specified.

# PRACTICING MARKETING RESEARCH



## ***New Data Quality Procedures to Identify Interviewer Falsification***

Interviewer falsification or cheating is a serious concern when it comes to data quality control. It introduces bias into survey responses, and it tends to increase when interviewers are working under terms of monetary incentives. The National Survey on Drug Use and Health (NSDUH), a federally sponsored annual survey on substance use and abuse that polls 150,000 households and 67,500 people, came up with innovative and effective new measures to catch interviewer falsification-generated responses before they enter the final data.

For example, in their reviews of interview data, NSDUH researchers focused on four interviewers whose work included 760 screenings and 464 interviews. NSDUH targeted these four because a preliminary inspection of their responses showed they had entered their own or another interviewer's telephone number for initial verification. Inspectors determined that only 38 percent (287) of screening cases and 29 percent (134) of interview cases were valid; 473 screenings and 330 interviews were falsified and thus rejected.

The NSDUH's solution was to rigorously examine timing data in the context of interview response and question level. Given an interviewer's caseload, the responses collected by an interviewer could be evaluated in terms of whether

they seemed likely in the given time frame, which was generally set at anything less than 30 minutes or longer than 60. Timestamp data was collected from the interviewers' computer terminals each day; any cases falling outside this margin were entered on a weekly Interview Length Report. Then technical staff examined the timing data against any explanation the interviewers offered.

"If a problematic pattern emerges, the interviewer's cases may be forced into verification and examined for shortcutting or fraudulent behavior," explains Joe Murphy (and colleagues) of RTI International, who reported on NSDUH's methods at the American Association for Public Opinion Research's 59th Annual Conference in May 2004. Murphy notes that virtually the only way to beat this new falsification detection system would be for interviewers to have advanced detailed understanding of the prevalence and correlates of substance use so that they could concoct likely responses; this level of subject sophistication is unlikely.

Other methods in NSDUH's data quality monitoring system include response deviation score (falsification is suspected if the interviewer's response deviation score, based on prevalence rates, is five times higher than the average); rare response combinations (falsification is suspected if the interviewer's responses show at least two examples or 5 percent of the total); and total interview seconds per question (examiners look for shorter or longer than average interview duration times).<sup>1</sup>

normally validates at a 10 percent rate, 5 respondents surveyed by that interviewer would be recontacted by telephone. Telephone validation typically answers four questions:

1. Was the person actually interviewed?
2. Did the person who was interviewed qualify to be interviewed according to the screening questions on the survey? For example, the interview may have required that the person being interviewed come from a family with an annual household income of \$25,000 or more. On validation, the respondent would again be asked whether the annual household income for his or her family was \$25,000 or more per year.



3. Was the interview conducted in the required manner? For example, a mall survey should have been conducted in the designated mall. Was this particular respondent interviewed in the mall, or was she or he interviewed at some other place, such as a restaurant or someone's home?
4. Did the interviewer cover the entire survey? Sometimes interviewers recognize that a potential respondent is in a hurry and may not have time to complete the entire survey. If respondents for that particular survey are difficult to find, the interviewer may be motivated to ask the respondent a few questions at the beginning and a few questions at the end and then fill out the rest of the survey without the respondent's input. Validation for this particular problem would involve asking respondents whether they were asked various questions from different points in the interview.

Validation also usually involves checking for other problems. For example: Was the interviewer courteous? Did the interviewer speculate about the client's identity or the purpose of the survey? Does the respondent have any other comments about the interviewer or the interview experience?

The purpose of the validation process, as noted earlier, is to ensure that interviews were administered properly and completely. Researchers must be sure that the research results on which they are basing their recommendations reflect the legitimate responses of target individuals.

A mall survey should be conducted in the designated mall. An important part of data analysis is validating that the data were gathered as specified.

#### ➤ editing

Process of ascertaining that questionnaires were filled out properly and completely.

#### ➤ skip pattern

Sequence in which later questions are asked, based on a respondent's answer to an earlier question or questions.

## Editing

Whereas validation involves checking for interviewer cheating and failure to follow instructions, **editing** involves checking for interviewer and respondent mistakes. Paper questionnaires usually are edited at least twice before being submitted for data entry. First, they may be edited by the field service firm that conducted the interviews, and then they are edited by the marketing research firm that hired the field service firm to do the interviewing. CATI, Internet, and other software-driven surveys have built-in logical checking. The editing process for paper surveys involves manual checking for a number of problems, including the following:

1. *Whether the interviewer failed to ask certain questions or record answers for certain questions.* In the questionnaire shown in Exhibit 14.1, no answer was recorded for question 19. According to the structure of the questionnaire, this question should have been asked of all respondents. Also, the respondent's name does not give a clear indication of gender. The purpose of the first edit—the field edit—is to identify these types of problems when there is still time to recontact the respondent and determine the appropriate answer to questions that were not asked. This may also be done at the second edit (by the marketing research firm), but in many instances there is not time to recontact the respondent and the interview has to be discarded.
2. *Whether skip patterns were followed.* According to the **skip pattern** in question 2 in Exhibit 14.1, if the answer to this question is “Very unlikely” or “Don't know,” the interviewer should skip to question 16. The editor needs to make sure that the interviewer followed instructions. Sometimes, particularly during the first few interviews in a particular study, interviewers get mixed up and skip when they should not or fail to skip when they should.
3. *Whether the interviewer paraphrased respondents' answers to open-ended questions.* Marketing researchers and their clients usually are very interested in the responses to

Exhibit 14.1

Sample Questionnaire

Consumer Survey  
Cellular Telephone Survey Questionnaire

Long Branch—Asbury, N.J.  
(01-03) 001

Date 1-05-01

Respondent Telephone Number 201-555-2322

Hello. My name is Sally with POST Research. May I please speak with the male or female head of the household?

(IF INDIVIDUAL NOT AVAILABLE, RECORD NAME AND CALLBACK INFORMATION ON SAMPLING FORM.)

(WHEN MALE/FEMALE HEAD OF HOUSEHOLD COMES TO PHONE): Hello, my name is \_\_\_\_\_, with POST Research. Your number was randomly selected, and I am not trying to sell you anything. I simply want to ask you a few questions about a new type of telephone service.

1. First, how many telephone calls do you make during a typical day?

- (04)
- 0-2 .....1
- 3-5 .....2
- 6-10 .....③
- 11-15 .....4
- 16-20 .....5
- More than 20 .....6
- Don't know .....7

Now, let me tell you about a new service called cellular mobile telephone service, which is completely wireless. You can get either a portable model that may be carried in your coat pocket or a model mounted in any vehicle. You will be able to receive calls and make calls, no matter where you are. Although cellular phones are wireless, the voice quality is similar to your present phone service. This is expected to be a time-saving convenience for household use.

This new cellular mobile phone service may soon be widely available in your area.

2. Now, let me explain to you the cost of this wireless service. Calls will cost 26 cents a minute plus normal toll charges. In addition, the monthly minimum charge for using the service will be \$7.50 and rental of a cellular phone will be about \$40. Of course, you can buy the equipment instead of leasing it. At this price, do you think you would be very likely, somewhat likely, somewhat unlikely, or very unlikely to subscribe to the new phone service?

- (05)
- Very likely .....1
- Somewhat likely .....②
- Somewhat unlikely .....3
- Very unlikely .....(GO TO QUESTION 16) .....4
- Don't know .....(GO TO QUESTION 16) .....5

INTERVIEWER—IF “VERY UNLIKELY” OR “DON’T KNOW,” GO TO QUESTION 16.

3. Do you think it is likely that your employer would furnish you with one of these phones for your job?

- (06)
- No .....(GO TO QUESTION 5) .....1
- Don't know .....(GO TO QUESTION 5) .....2
- Yes .....(CONTINUE) .....③

INTERVIEWER—IF “NO” OR “DON’T KNOW,” GO TO QUESTION 5; OTHERWISE CONTINUE.

4. If your employer did furnish you with a wireless phone, would you also purchase one for household use?

- (07)
- Yes .....(CONTINUE) .....①
- No .....(GO TO QUESTION 16) .....2
- Don't know .....(GO TO QUESTION 16) .....3

5. Please give me your best estimate of the number of mobile phones your household would use (write in “DK” for “Don’t know”).

Number of Units \_\_\_\_\_ 01 \_\_\_\_\_ (08-09)

**Exhibit 14.1** (continued)

6. Given that cellular calls made or received will cost 26 cents a minute plus normal toll charges during weekdays, how many calls on the average would you expect to make in a typical weekday?

RECORD NUMBER \_\_\_\_\_ 06 \_\_\_\_\_ (10–11)

7. About how many minutes would your average cellular call last during the week?

RECORD NUMBER \_\_\_\_\_ 05 \_\_\_\_\_ (12–13)

8. Weekend cellular calls made or received will cost 8 cents per minute plus normal toll charges. Given this, about how many cellular calls on the average would you expect to make in a typical Saturday or Sunday?

RECORD NUMBER \_\_\_\_\_ 00 \_\_\_\_\_ (14–15)

9. About how many minutes would your average cellular call last on Saturday or Sunday?

RECORD NUMBER \_\_\_\_\_ (16–17)

10. You may recall from my previous description that two types of cellular phone units will be available. The vehicle phone may be installed in any vehicle. The portable phone will be totally portable—it can be carried in a briefcase, purse, or coat pocket. The totally portable phones may cost about 25 percent more and may have a more limited transmitting range in some areas than the vehicle phone. Do you think you would prefer portable or vehicle phones if you were able to subscribe to this service?

(18)

- Portable .....1
- Vehicle .....②
- Both .....3
- Don't know .....4

11. Would you please tell me whether you, on the average, would use a mobile phone about once a week, less than once a week, or more than once a week from the following geographic locations.

	Less Than Once a Week	Once a Week	More Than Once a Week	Never	
Monmouth County (IF "NEVER," SKIP TO QUESTION 16)	1	2	③	4	(19)
Sandy Hook	1	2	3	④	(20)
Keansburg	1	2	3	④	(21)
Atlantic Highlands	1	2	③	4	(22)
Matawan-Middletown	①	2	3	4	(23)
Red Bank	①	2	3	4	(24)
Holmdel	1	2	③	4	(25)
Eatontown	1	②	3	4	(26)
Long Branch	1	2	3	④	(27)
Freehold	1	2	3	④	(28)
Manalapan	1	2	3	④	(29)
Cream Ridge	1	2	3	④	(30)
Belmar	1	2	3	④	(31)
Point Pleasant	1	2	③	4	(32)

I'm going to describe to you a list of possible extra features of the proposed cellular service. Each option I'm going to describe will cost not more than \$3.00 a month per phone. Would you please tell me if you would be very interested, interested, or uninterested in each feature:

	Very Interested	Interested	Uninterested	
12. Call forwarding (the ability to transfer any call coming in to your mobile phone to any other phone).	①	2	3	(33)
13. No answer transfer (service that redirects calls to another number if your phone is unanswered).	1	2	③	(34)

**Exhibit 14.1** (continued)

	Very Interested	Interested	Uninterested	
14. Call waiting (a signal that another person is trying to call you while you are using your phone).	1	②	3	(35)
15. Voice mailbox (a recording machine that will take the caller's message and relay it to you at a later time. This service will be provided at \$5.00 per month).	1	2	③	(36)
16. What is your age group? (READ BELOW)				(37)
Under 25				.1
25-44				②
45-64				.3
65 and over				.4
Refused, no answer, or don't know				.5
17. What is your occupation?				(38)
Manager, official, or proprietor				①
Professional (doctors, lawyers, etc.)				.2
Technical (engineers, computer programmers, draftsmen, etc.)				.3
Office worker/clerical				.4
Sales				.5
Skilled worker or foreman				.6
Unskilled worker				.7
Teacher				.8
Homemaker, student, retired				.9
Not now employed				.X
Refused				.Y
18. Into which category did your total family income fall in 2002? (READ BELOW)				(39)
Under \$15,000				.1
\$15,000-\$24,999				.2
\$25,000-\$49,999				.3
\$50,000-\$74,999				.4
\$75,000 and over				⑤
Refused, no answer, don't know				.6
19. (INTERVIEWER—RECORD SEX OF RESPONDENT):				(40)
Male				.1
Female				.2
20. May I have your name? My office calls about 10 percent of the people I talk with to verify that I have conducted the interview.				
Gave name				①
Refused				.2

Jordan Beasley  
Name

Thank you for your time. Have a good day.

open-ended questions. The quality of the responses, or at least what was recorded, is an excellent indicator of the competence of the interviewer who recorded them. Interviewers are trained to record responses verbatim and not to paraphrase or insert their own language. They also are instructed to probe the initial response. The first part of Exhibit 14.2 shows an example of an interviewer's paraphrasing and interpretation of a response to an open-ended question. The second part of Exhibit 14.2



**Exhibit 14.2****Recording of Open-Ended Questions****A. Example of Improper Interviewer Recording of Response to an Open-Ended Question**

Question: Why do you go to Burger King most often among fast food/quick service restaurants? (PROBE)

*Response recorded:*

The consumer seemed to think Burger King had better tasting food and better quality ingredients.

**B. Example of Interviewer Failure to Probe a Response**

Question: Same as Part A.

*Only response recorded:*

Because I like it.

**C. Example of Proper Recording and Probing**

Question: Same as Part A.

*Response recorded:*

Because I like it. (P)\* I like it, and I go there most often because it is the closest place to where I work. (AE)\*\* No.

\*(P) is an interviewer mark indicating he or she has probed a response.

\*\* (AE) is interviewer shorthand for “Anything else?” This gives the respondent an opportunity to expand on the original answer.

shows the result of interviewer failure to probe a response. The response is useless from a decision-making perspective. It comes as no surprise that the respondent goes to Burger King most often because he likes it. The third part of Exhibit 14.2 shows how an initial meaningless response can be expanded to a useful response by means of proper probing. A proper probe to the answer “Because I like it” would be “Why do you like it?” or “What do you like about it?” The respondent then indicates that he goes there most often because it is the fast-food restaurant most convenient to his place of work.

The person doing the editing must make judgment calls in regard to substandard responses to open-ended questions. She or he must decide at what point particular answers are so limited as to be useless and whether respondents should be recontacted.

The editing process is extremely tedious and time-consuming. (Imagine for a moment reading through 13,000 pages of interviews!) However, it is a very important step in the processing of survey responses.

## Step Two: Coding

### coding

Process of grouping and assigning numeric codes to the various responses to a question.

As discussed in Chapter 11, **coding** refers to the process of grouping and assigning numeric codes to the various responses to a particular question. Most questions on surveys are closed-ended and precoded, meaning that numeric codes have been assigned to the various responses on the questionnaire. All answers to closed-ended questions should be precoded, as they are in question 1 on the questionnaire in Exhibit 14.1. Note that each answer has a numeric code to its right; the answer “0–2” has the code 1, the answer “3–5” has the code 2, and so on. The interviewer can record the response by circling the numeric code next to the answer given by the respondent. In this case, the respondent’s answer was seven calls per day. The code 3 next to the category “6–10” (calls per day) is circled.

Open-ended questions create a coding dilemma. They were phrased in an open-ended manner because the researcher either had no idea what answers to expect or wanted a richer response than is possible with a closed-ended question. As with editing, the process of coding responses to open-ended questions is tedious and time-consuming. In addition, the procedure is to some degree subjective.<sup>2</sup> For these reasons, researchers tend to avoid open-ended questions unless they are absolutely necessary.

## Coding Process

The process of coding responses to open-ended questions includes the following steps:

1. *List responses.* Coders at the research firm prepare lists of the actual responses given to each open-ended question on a particular survey. In studies of a few hundred respondents, all responses may be listed. With larger samples, responses given by a sample of respondents are listed. The listing of responses may be done as part of the editing process or as a separate step, often by the same individuals who edited the questionnaires.
2. *Consolidate responses.* A sample list of responses to an open-ended question is provided in Exhibit 14.3. Examination of this list indicates that a number of the responses can be interpreted to mean essentially the same thing; therefore, they can be appropriately consolidated into a single category. This process of consolidation might yield the list shown in Exhibit 14.4. Consolidating requires a number of subjective decisions—for example, does response number 4 in Exhibit 14.3 belong in category 1 or should it have its own category? These decisions typically are made by a qualified research analyst and may involve client input.
3. *Set codes.* A numeric code is assigned to each of the categories on the final consolidated list of responses. Code assignments for the sample beer study question are shown in Exhibit 14.4.

### EXHIBIT 14.3

### Sample of Responses to Open-Ended Question

**Question: Why do you drink that brand of beer? (BRAND MENTIONED IN ANSWER TO PREVIOUS QUESTION)**

#### Sample responses:

1. Because it tastes better.
2. It has the best taste.
3. I like the way it tastes.
4. I don't like the heavy taste of other beers.
5. It is the cheapest.
6. I buy whatever beer is on sale. It is on sale most of the time.
7. It doesn't upset my stomach the way other brands do.
8. Other brands give me headaches. This one doesn't.
9. It has always been my brand.
10. I have been drinking it for over 20 years.
11. It is the brand that most of the guys at work drink.
12. All my friends drink it.
13. It is the brand my wife buys at the grocery store.
14. It is my wife's/husband's favorite brand.
15. I have no idea.
16. Don't know.
17. No particular reason.

EXHIBIT 14.4		Consolidated Response Categories and Codes for Open-Ended Responses from Beer Study	
Response Category Descriptor	Response Items from Exhibit 14.1 Included	Assigned Numeric Code	
Tastes better/like taste/tastes better than others	1, 2, 3, 4	1	
Low/lower price	5, 6	2	
Does not cause headache, stomach problems	7, 8	3	
Long-term use, habit	9, 10	4	
Friends drink it/influence of friends	11, 12	5	
Wife/husband drinks/buys it	13, 14	6	

EXHIBIT 14.5		Example Questionnaire Setup for Open-Ended Questions	
37.	Why do you drink that brand of beer? (BRAND MENTIONED IN PREVIOUS QUESTION)?	(48)	<u>  2  </u>
	<i>Because it's cheaper. (P) Nothing. (AE) Nothing.</i>		

4. *Enter codes.* After responses have been listed and consolidated and codes set, the last step is the actual entry of codes. This involves several substeps:
- Read responses to individual open-ended questions on questionnaires.
  - Match individual responses with the consolidated list of response categories, and determine the appropriate numeric code for each response.
  - Write the numeric code in the appropriate place on the questionnaire for the response to the particular question (see Exhibit 14.5) or enter the appropriate code in the database electronically.<sup>3</sup>

Here's an example of the process, using the listing of responses shown in Exhibit 14.3 and the consolidation and setting of codes shown in Exhibit 14.4.

- You turn to the first questionnaire and read this response to the question "Why do you drink that brand of beer?": "Because it's cheaper."
- You compare this response with the consolidated response categories and decide that it best fits into the "Low/lower price" category. The numeric code associated with this category is 2 (see Exhibit 14.4).
- You enter the code in the appropriate place on the questionnaire (see Exhibit 14.5).

## Automated Coding Systems

With CATI and Internet surveys, data entry and coding are completely eliminated for closed-ended questions. However, when the text of open-ended questions is electronically captured, a coding process is still required. A number of developments are making it likely that the tedious coding process for open-ended questions will soon be replaced with computer-based systems requiring limited high-level human intervention and decision making.<sup>4</sup> The Practicing Marketing Research feature on page 443 provides an example of an automated coding system.

# PRACTICING MARKETING RESEARCH



## Text Analytics Software Streamlines Coding Open-Ended Responses

As market researchers well know, the major shortcoming of open-ended questions in a survey is the postinterview coding. Opinion-based answers do not easily lend themselves to simple numerical coding. It is very expensive, in terms of time and effort, to categorize individualized responses, and it tends to limit survey size so as to avoid this complication during data processing.

Keyword-based search software helps the human analysis of open-ended responses, but typically this kind of software cannot deal with the variety of unstructured responses. Each answer must be interpreted by an analyst at an average cost of \$2.00 to \$5.00 question.

A new Text Analytics software program called Content Analyst™ now offers an automated way to code open-ended responses and cut costs (and processing time) by about 50 percent. This

now makes it feasible to conduct in-depth surveys of 100,000, for example, asking many open-ended questions without concern for the prohibitive postinterview cost of coding. The software automates the laborious analysis of conceptual information and the processing and interpretation of large volumes of open-ended data. The product is made by Content Analyst Company, LLC, of Reston, Virginia ([www.contentanalyst.com](http://www.contentanalyst.com)).

"It's taking us in the direction of concept-based coding rather than keyword coding, and that's a significant advance," comments Justin Greeves of Worthlin Worldwide, an opinion research company based in MacLean, Virginia, that uses the software. The approach, Greeves adds, takes us one step closer to the "automation of human-level analysis." The software also shifts the focus from coding answers to a higher-value interpretation of results. "The open-ends reveal the voice of the customer," so that their different and unstructured answers are no longer a problem but a source of greater, more valuable market information.<sup>5</sup>

The TextSmart module of SPSS is one example of the new breed of automated coding systems. Algorithms based on semiotics<sup>4</sup> are at the heart of these systems and show great promise for speeding up the coding process, reducing its cost, and increasing its objectivity. Basically, these algorithms use the power of computers to search for patterns in open-ended responses and in group responses, based on certain keywords and phrases.

## Step Three: Data Entry

Once the questionnaires have been validated, edited, and coded, it's time for the next step in the process—data entry. We use the term **data entry** here to refer to the process of converting information to a form that can be read by a computer. This process requires a data entry device, such as a computer terminal or a personal computer, and a storage medium, such as magnetic tape, a floppy disk, or a hard (magnetic) disk.

### Intelligent Entry Systems

Most data entry is done by means of intelligent entry systems. With **intelligent data entry**, the information entered is checked for internal logic. Intelligent entry systems can

#### ▶ data entry

Process of converting information to an electronic format.

#### ▶ intelligent data entry

Form of data entry in which the information being entered into the data entry device is checked for internal logic.

be programmed to avoid certain types of errors at the point of data entry, such as invalid or wild codes and violation of skip patterns.

Consider question 2 on the questionnaire in Exhibit 14.1. The five valid answers have the associated numeric codes 1 through 5. An intelligent data entry system programmed for valid codes would permit the data entry operator to enter only one of these codes in the field reserved for the response to this question. If the operator attempts to enter a code other than those defined as valid, the device will inform the data entry operator in some manner that there is a problem. The data entry device, for example, might beep and display a message on the screen that the entered code is invalid. It will not advance to the next appropriate field until the code has been corrected. Of course, it is still possible to incorrectly enter a 3 rather than the correct answer 2. Referring again to question 2, note that if the answer to the question is “Very unlikely” or “Don’t know,” then the data entry operator should skip to question 16. An intelligent data entry device will make this skip automatically.

## The Data Entry Process

The validated, edited, and coded questionnaires have been given to a data entry operator seated in front of a personal computer. The data entry software system has been programmed for intelligent entry. The actual data entry process is ready to begin. Usually, the data are entered directly from the questionnaires, because experience has shown that a large number of errors are introduced when questionnaire data are transposed manually to coding sheets. Going directly from the questionnaire to the data entry device and associated storage medium is much more accurate and efficient. To better understand the mechanics of the process, look again at Exhibit 14.1.

- In the upper right-hand corner of the questionnaire, the number 001 is written. This number uniquely identifies the particular questionnaire, which should be the first questionnaire in the stack that the data entry operator is preparing to enter. This number is an important point of reference because it permits the data entry staff to refer back to the original document if any errors are identified in connection with the data input.
- To the left of the handwritten number 001 is (01–03). This tells the data entry operator that 001 should be entered into fields 01–03 of the data record. Throughout the questionnaire, the numbers in parentheses indicate the proper location on the data record for the circled code for the answer to each question. Question 1 has (04) associated with the codes for the answers to the question. Thus, the answer to this question would be entered in field 04 of the data record. Now, take a look at the open-ended question in Exhibit 14.5. As with closed-ended questions, the number in parentheses refers to the field on the data record where the code or codes for the response to this question should be entered. Note the number 2 written in to the right of (48); a 2 should be entered in field 48 of the data record associated with this questionnaire.

Exhibit 14.1 clearly illustrates the relationship between the layout of the questionnaire, in terms of codes (numbers associated with different answers to questions) and fields (places on the data record where the code is entered), and the layout of the data record.

## Scanning

As all students know, the scanning of documents (test scoring sheets) has been around for decades. It has been widely used in schools and universities as an efficient way to capture and score responses to multiple-choice questions. However, until more recently, its use in marketing research has been limited. This limited use can be attributed to two factors: setup costs and the need to record all responses with a No. 2 pencil. Setup costs include the cost of special paper, special ink in the printing process, and very precise placement of the bubbles for recording responses. The break-even point, at which the

savings in data entry costs exceeded the setup costs, was in the 10,000 to 12,000 survey range. Therefore, for most surveys, scanning was not feasible.

However, changes in **scanning technology** and the advent of personal computers have changed this equation. Today, questionnaires prepared with any one of a number of Windows word-processing software packages and printed on a laser printer or by a standard printing process can be readily scanned, using the appropriate software and a scanner attached to a personal computer. In addition, the latest technology permits respondents to fill out the survey using almost any type of writing implement (any type of pencil, ballpoint pen, or ink pen). This eliminates the need to provide respondents with a No. 2 pencil and greatly simplifies the process of mailing surveys. Finally, the latest technology does not require respondents to carefully shade the entire circle or square next to their response choices; they can put shading, a check mark, an X, or any other type of mark in the circle or square provided for the response choice.<sup>6</sup>

As a result of these developments, the use of scannable surveys is growing dramatically. An analyst who expects more than 400 to 500 surveys to be completed will find scannable surveys to be cost-effective.

Though no reliable volume figures are available, it is an accepted fact that the amount of survey data being captured electronically is increasing. For example, electronic data capture is used in computer-assisted telephone interviewing, Internet surveys, disks-by-mail surveys, and TouchScreen kiosk surveys.


## Step Four: Machine Cleaning of Data


At this point, the data from all questionnaires have been entered and stored in the computer that will be used to process them. It is time to do final error checking before proceeding to the tabulation and statistical analysis of the survey results. Many colleges have one or more statistical software packages available for the tabulation and statistical analysis of data, including SAS (Statistical Analysis System) and SPSS (Statistical Package for the Social Sciences), which have proven to be the most popular mainframe computer statistical packages. Most colleges have personal computer versions of SPSS and SAS, in addition to other PC statistical packages. The number of other PC packages is large and growing.


Regardless of which computer package is used, it is important to do a final computerized error check of the data, or what is sometimes referred to as **machine cleaning of data**. This may be done through error checking routines and/or marginal reports.


Some computer programs permit the user to write **error checking routines**. These routines include a number of statements to check for various conditions. For example, if a particular field on the data records for a study should be coded with only a 1 or a 2, a logical statement can be written to check for the presence of any other code in that field. Some of the more sophisticated packages generate reports indicating how many times a particular condition was violated and the data records on which it was violated. With this list, the user can refer to the original questionnaires and determine the appropriate values.

Exhibit 14.6 illustrates the **marginal report**, another approach to machine cleaning often used for error checking. The first row of this report lists the fields of the data record. The columns show the frequency with which each possible value was encountered in each field. For example, the second row in Exhibit 14.6 shows that in field 111 of the data records for this study there are 100 “1” punches, 100 “2” punches, 1 “3” punch, and 99 “10” punches. This report permits the user to determine whether inappropriate codes were entered and whether skip patterns were properly followed. If all the numbers are consistent, there is no need for further cleaning. However, if logical errors (violated skip patterns and impossible codes) are detected, then the appropriate original questionnaires must be located and the corrections made in the computer data file. Note that these

 **scanning technology**  
Form of data entry in which responses on questionnaires are read in automatically by the data entry device.

 **machine cleaning of data**  
Final computerized error check of data.

 **error checking routines**  
Computer programs that accept instructions from the user to check for logical errors in the data.

 **marginal report**  
Computer-generated table of the frequencies of the responses to each question, used to monitor entry of valid codes and correct use of skip patterns.

**EXHIBIT 14.6** Sample Marginal Report (Marginal Counts of 300 Records)


FIELD	1	2	3	4	5	6	7	8	9	10	11	12	BL	TOT
111	100	100	1	0	0	0	0	0	0	99	0	0	0	300
112	30	30	30	30	30	30	30	30	30	0	0	0	0	300
113	30	30	30	30	30	30	30	30	30	30	0	0	0	300
114	67	233	0	0	0	0	0	0	0	0	0	0	0	300
115	192	108	0	0	0	0	0	0	0	0	0	0	0	300
116	108	190	0	0	0	0	0	0	0	0	0	2	0	300
117	13	35	8	0	2	136	95	7	2	0	0	0	2	298
118	0	0	0	0	0	0	0	0	0	0	0	2	298	2
119	29	43	12	1	2	48	50	6	4	1	0	0	104	196
1111	6	16	6	1	1	10	18	4	2	0	0	0	236	64
1113	3	4	1	1	0	1	2	0	1	0	0	0	288	12
1115	0	0	0	1	1	0	0	2	0	0	0	0	296	4
1117	24	2	22	0	1	239	9	2	0	0	0	0	1	299
1118	0	0	0	0	0	0	0	0	0	0	0	0	299	1
1119	4	49	6	0	0	81	117	5	2	0	0	0	36	264
1120	0	0	0	0	0	0	0	0	0	0	0	36	264	36
1121	5	60	6	0	0	84	116	4	3	1	0	0	21	279
1122	0	0	0	0	0	0	0	0	0	0	0	21	279	21
1123	118	182	0	0	0	0	0	0	0	0	0	0	0	300
1124	112	187	0	0	0	0	0	0	0	0	0	0	1	299
1125	47	252	0	0	0	0	0	0	0	0	0	1	0	300
1126	102	198	0	0	0	0	0	0	0	0	0	0	0	300
1127	5	31	5	1	0	33	31	9	1	0	0	0	184	116
1128	0	0	0	0	0	0	0	0	0	0	0	2	298	2
1129	0	3	1	0	0	4	8	2	1	0	0	0	281	19
1131	7	16	3	0	2	60	21	3	0	0	0	0	188	112
1133	1	3	1	0	0	2	3	1	0	0	0	0	289	11

procedures cannot identify situations in which an interviewer or data entry operator incorrectly entered a 2 for a “no” response instead of a 1 for a “yes” response.

This is the final error check in the process. When this step is completed, the computer data file should be ready for tabulation and statistical analysis. Exhibit 14.7 shows the data for the first 50 respondents (out of a total of 400) for the study associated with the questionnaire shown in Exhibit 14.1. Note that the apparent gaps in the data are a result of the skip called for in question 4. Also note that the gender data (noted as missing earlier) for respondent 001 has been filled in with a 2 for female based on information obtained by recontacting the respondent.

## Step Five: Tabulation and Statistical Analysis

The survey results have been stored in a computer file and are free of logical data entry and interviewer recording errors. The next step is to tabulate the survey results.

 **one-way frequency table**  
Table showing the number of respondents choosing each answer to a survey question.

### One-Way Frequency Tables

The most basic tabulation is the **one-way frequency table**, which shows the number of respondents who gave each possible answer to each question. An example of this type of table appears in Exhibit 14.8. This table shows that 144 consumers (48 percent) said they

**EXHIBIT 14.7****Printout of Data for the First 50 Respondents  
for Cellular Telephone Survey**

001323101060500	234431132444443132321521
00224	23412
00334	49622
00414	36221
00524	33312
00634	22612
00714	21321
008221	0204050310334232444434444222229321
00925	36311
01044	23311
01161310240050330134234444434443322330321	
012622	01400720073344444444444132330511
013221	0106030603231312333323322123216211
01424	29321
01514	40121
01624	22612
01774	20622
01854	34621
01924	25212
02024	23622
02114	16611
02214	36211
02314	15611
024131	0010100410221334444444442229611
02524	26621
026131	0101030203124221422244414223222611
02724	10122
02814	59622
02924	39622
03024	49611
03134	53621
03234	32622
03321	01 1244444444444211220211
03424	32622
035311	0410300430133131131113131211220121
03623230301050201334414433344244232320622	
03724	37622
03814	40121
03934	30121
04024	16121
04124	26311
04264	26411
04324	20321
04414	26311
04524	19321
04634	19222
04724	29621
04824	31422
04924	33121
05014	21311

---



**EXHIBIT 14.8** One-Way Frequency Table

**Q.30** If you or a member of your family were to require hospitalization in the future, and the procedure could be performed in Minneapolis or St. Paul, where would you choose to go?

	Total
Total	300 100%
To a hospital in St. Paul	144 48.0%
To a hospital in Minneapolis	146 48.7%
Don't know/no response	10 3.3%

would choose a hospital in St. Paul, 146 (48.7 percent) said they would choose a hospital in Minneapolis, and 10 (3.3 percent) said they didn't know which location they would choose. A printout is generated with a one-way frequency table for every question on the survey. In most instances, a one-way frequency table is the first summary of survey results seen by the research analyst. In addition to frequencies, these tables typically indicate the percentage of those responding who gave each possible response to a question.

An issue that must be dealt with when one-way frequency tables are generated is what base to use for the percentages for each table. There are three options for a base:

- 1. Total respondents.** If 300 people are interviewed in a particular study and the decision is to use total respondents as the base for calculating percentages, then the percentages in each one-way frequency table will be based on 300 respondents.
- 2. Number of people asked the particular question.** Because most questionnaires have skip patterns, not all respondents are asked all questions. For example, suppose question 4 on a particular survey asked whether the person owned any dogs and 200 respondents indicated they were dog owners. Since questions 5 and 6 on the same survey were to be asked only of those individuals who owned a dog, questions 5 and 6 should have been asked of only 200 respondents. In most instances, it would be appropriate to use 200 as the base for percentages associated with the one-way frequency tables for questions 5 and 6.
- 3. Number of people answering the question.** Another alternative base for computing percentages in one-way frequency tables is the number of people who actually answered a particular question. Under this approach, if 300 people were asked a particular question but 28 indicated "Don't know" or gave no response, then the base for the percentages would be 272.

Ordinarily, the number of people who were asked a particular question is used as the base for all percentages throughout the tabulations, but there may be special cases in which other bases are judged appropriate. Exhibit 14.9 is a one-way frequency table in which three different bases are used for calculating percentages.

Some questions, by their nature, solicit more than one response from respondents. For example, consumers might be asked to name all brands of vacuum cleaners that come to mind. Most people will be able to name more than one brand. Therefore, when these answers are tabulated, there will be more responses than respondents. If 200 consumers are surveyed and the average consumer names three brands, then there will



The base for each percentage must be determined before one-way frequency tables are run. If a survey question asks whether the person has a dog and 200 respondents indicate that they do, further questions designated for dog owners should have only 200 respondents.

be 200 respondents and 600 answers. The question is, should percentages in frequency tables showing the results for these questions be based on the number of respondents or the number of responses? Exhibit 14.10 shows percentages calculated using both bases. Most commonly, marketing researchers compute percentages for multiple-response

**EXHIBIT 14.9**
**One-Way Frequency Table Using Three Different Bases for Calculating Percentages**
**Q.35 Why would you not consider going to St. Paul for hospitalization?**

	Total* Respondents	Total Asked	Total Answering
Total	300	64	56
	100%	100%	100%
They aren't good/service poor	18	18	18
	6%	28%	32%
St. Paul doesn't have the services/equipment that Minneapolis does	17	17	17
	6%	27%	30%
St. Paul is too small	6	6	6
	2%	9%	11%
Bad publicity	4	4	4
	1%	6%	7%
Other	11	11	11
	4%	17%	20%
Don't know/no response	8	8	8
	3%	13%	

\*A total of 300 respondents were surveyed. Only 64 were asked this question because in the previous question those respondents said they would not consider going to St. Paul for hospitalization. Only 56 respondents gave an answer other than "Don't know."

**EXHIBIT 14.10** Percentages for a Multiple-Response Question Calculated on the Bases of Total Respondents and Total Responses

**Q.34** To which of the following towns and cities would you consider going for hospitalization?

	Total Respondents	Total Responses
Total	300 100%	818 100%
Minneapolis	265 88.3%	265 32.4%
St. Paul	240 80.0%	240 29.3%
Bloomington	112 37.3%	112 13.7%
Rochester	92 30.7%	92 11.2%
Minnetonka	63 21.0%	63 7.7%
Eagan	46 15.3%	46 5.6%

questions on the basis of the number of respondents, reasoning that the client is primarily interested in the proportion of people who gave a particular answer.

### Cross Tabulations

**cross tabulation**  
Examination of the responses to one question relative to the responses to one or more other questions.

**Cross tabulations** are likely to be the next step in analysis. They represent a simple-to-understand, yet powerful, analytical tool. Many marketing research studies go no further than cross tabulations in terms of analysis. The idea is to look at the responses to one question in relation to the responses to one or more other questions. Exhibit 14.11

**EXHIBIT 14.11** Sample Cross tabulation

**Q.30** If you or a member of your family were to require hospitalization in the future, and the procedure could be performed in Minneapolis or St. Paul, where would you choose to go?

	Age				
	Total	18–34	35–54	55–64	65 or Over
Total	300 100%	65 100%	83 100%	51 100%	100 100%
To a hospital in St. Paul	144 48.0%	21 32.3%	40 48.2%	25 49.0%	57 57.0%
To a hospital in Minneapolis	146 48.7%	43 66.2%	40 48.2%	23 45.1%	40 40.0%
Don't know/no response	10 3.3%	1 1.5%	3 3.6%	3 5.9%	3 3.0%

# PRACTICING MARKETING RESEARCH



## Six Practical Tips for Easier Cross Tabulations

Cross tabulation is a valuable method of mining further data and significance and teasing out unsuspected relationships from your basic survey data. Here are six practical tips to improve your cross tabulation gleanings from Custom Insight, a provider of Web-based survey software located in Carson City, Nevada ([www.custominsight.com](http://www.custominsight.com)).

1. **Make Hypotheses.** Probably you already have one or two hunches about what the data might yield in cross tabulation. Articulate your initial hypotheses and use them as a starting point for cross tabulation.
2. **Look for What Is Not There.** After you observe what the data manifestly shows, examine it for what it doesn't show, that is, relationships that you may have assumed to be real or substantive. For example, if your hypothesis postulates that people with higher incomes plan to make more purchases, the data may actually refute that and thus reveal a new set of data—that affluent people are not planning to spend.
3. **Scrutinize for the Obvious.** Some relationships among the data may be obvious (e.g., age and student status). Finding these
4. **Keep Your Mind Open.** Don't be tied to your hypotheses and assumptions. You may see data relationships that you hadn't expected and that are not congruent with your hypotheses. Think about the data from this new angle and formulate new hypotheses to account for them.
5. **Trust the Data.** If your results don't match your initial expectations, maybe they're wrong, not the data. Study the data for new relationships even if they contradict your starting hypotheses.
6. **Watch the "n."** Small totals should raise suspicions, and if you have few respondents in a given category, do not trust the data or look for stronger trends first before drawing final conclusions. For example, your study shows that 38 percent of people under age 15 want a particular product, except that only 8 people comprise that 38 percent. Better is the fact that 88 percent of people under 15 are students. Even though the number of respondents is minimal, the relationship exhibited by the data (88 percent) is much stronger and can be trusted.<sup>7</sup>

shows a simple cross tabulation that examines the relationship between cities consumers are willing to consider for hospitalization and their age. This cross tabulation includes frequencies and percentages, with the percentages based on column totals. This table shows an interesting relationship between age and likelihood of choosing Minneapolis or St. Paul for hospitalization. Consumers in successively older age groups are increasingly likely to choose St. Paul and increasingly less likely to choose Minneapolis.

Following are a number of considerations regarding the setup of cross tabulation tables and the determination of percentages within them:

- The previous discussion regarding the selection of the appropriate base for percentages applies to cross tabulation tables as well.
- Three different percentages may be calculated for each cell in a cross tabulation table: column, row, and total percentages. Column percentages are based on the column total, row percentages are based on the row total, and total percentages are

<b>EXHIBIT 14.12</b>		<b>Cross tabulation Table with Column, Row, and Total Percentages*</b>		
<b>Q.34 To which of the following towns and cities would you consider going for hospitalization?</b>				
	<b>Total</b>	<b>Male</b>	<b>Female</b>	
Total	300	67	233	
	100.0%	100.0%	100.0%	
	100.0%	22.3%	77.7%	
	100.0%	22.3%	77.7%	
St. Paul	265	63	202	
	88.3%	94.0%	86.7%	
	100.0%	23.6%	76.2%	
	88.3%	21.0%	67.3%	
Minneapolis	240	53	187	
	80.0%	79.1%	80.3%	
	100.0%	22.1%	77.9%	
	80.0%	17.7%	62.3%	
Bloomington	112	22	90	
	37.3%	32.8%	38.6%	
	100.0%	19.6%	80.4%	
	37.3%	7.3%	30.0%	

\*Percentages listed are column, row, and total percentages, respectively.

based on the table total. Exhibit 14.12 shows a cross tabulation table in which the frequency and all three of the percentages are shown for each cell in the table.

- A common way of setting up cross tabulation tables is to use columns to represent factors such as demographics and lifestyle characteristics, which are expected to be

<b>EXHIBIT 14.13</b>		<b>A Stub and Banner Table</b>							
<b>North Community College—Anywhere, U.S.A.</b>									
<b>Q.1c. Are you single, married, or formerly married?</b>									
	<b>Total</b>	<b>Zones</b>			<b>Gender</b>		<b>Age</b>		
		<b>1</b>	<b>2</b>	<b>3</b>	<b>M</b>	<b>F</b>	<b>18–34</b>	<b>35–54</b>	<b>55 and Over</b>
Total	300	142	103	55	169	131	48	122	130
	100%	100%	100%	100%	100%	100%	100%	100%	100%
Married	228	105	87	36	131	97	36	97	95
	76%	74%	84%	65%	78%	74%	75%	80%	73%
Single	5	1	2	2	4	1	2	1	2
	2%	1%	2%	4%	2%	1%	4%	1%	2%
Formerly married	24	11	10	3	12	12	3	9	12
	8%	8%	10%	5%	7%	9%	6%	7%	9%
Refused to answer	43	25	4	14	22	21	7	15	21
	14%	18%	4%	25%	13%	16%	15%	12%	16%

predictors of the state of mind, behavior, or intentions data shown as rows of the table. In such tables, percentages usually are calculated on the basis of column totals. This approach permits easy comparisons of the relationship between, say, lifestyle characteristics and expected predictors such as sex or age. For example, in Exhibit 14.11, this approach facilitates examination of how people in different age groups differ in regard to the particular factor under examination.

Cross tabulations provide a powerful and easily understood approach to the summarization and analysis of survey research results. However, it is easy to become swamped by the sheer volume of computer printouts if a careful tabulation plan has not been developed. The cross tabulation plan should be created with the research objectives and hypotheses in mind. Because the results of a particular survey might be cross tabulated in an almost endless number of ways, it is important for the analyst to exercise judgment and select from all the possibilities only those cross tabulations that are truly responsive to the research objectives of the project. Spreadsheet programs such as Excel and nearly all statistics packages (SAS, SPSS, SYSTAT, STATISTICA) can generate cross tabulations. Chapter 15 discusses the chi-square test, which can be used to determine whether the results in a particular cross tabulation table are significantly different from what would be expected based on chance. In other words, confronted with the question of whether the response patterns of men differ significantly from those of women, the analyst can use this statistical procedure to determine whether the differences between the two groups likely occurred because of chance or likely reflect real differences.

A complex cross tabulation, generated using the UNCLE software package, is shown in Exhibit 14.13. UNCLE was designed with the special needs of marketing researchers in mind and is widely used in the marketing research industry. As indicated, this more complex table is sometimes referred to as a *stub and banner table*. The column headings are the banner and the row titles are the stub. In this single table, the relationship between marital status and each of seven other variables is explored. Cross tabulation can be produced in Excel, as described in the Practicing Marketing Research feature on page 455, but it is a cumbersome process.

Race		Family Profile		Vote History		Registered Voter		
White	Black	Other	Child <18 years	Child >18 years	2-3 Times	4 Times or More	Yes	No
268	28	4	101	53	104	196	72	228
100%	100%	100%	100%	100%	100%	100%	100%	100%
207	18	3	82	39	80	148	58	170
77%	64%	75%	81%	74%	77%	76%	81%	75%
5	—	—	—	—	2	3	1	4
2%	—	—	—	—	2%	2%	1%	2%
18	6	—	5	6	10	14	3	21
7%	21%	—	5%	11%	10%	7%	4%	9%
38	4	1	14	8	12	31	10	33
14%	14%	25%	14%	15%	12%	16%	14%	14%



## FROM THE FRONT LINE

### Secrets of Developing a Good Set of Tables



*Suzanne Simpson, Director of Data Processing, DSS Research*

A solid and error-free set of cross tabulation tables is critical to the analysis and report preparation phase of any project. Developing a good set of tables goes well beyond the technical proficiency needed to complete the task. A keen eye for errors in the tables (e.g., tables that don't add up, skip patterns that have inconsistent numbers, incorrect labeling), a strong desire for consistency, and curiosity make the difference between creating clean, professional tables and tables that may not be passable.

Tables are only as good as the data used to create them. Clean data are the result of careful planning and meticulous checking. Familiarity with the data is a natural outcome of the cleaning process. A skilled programmer uses this familiarity to hunt for and find any inconsistencies

that result from data entry or programming (CATI or Internet) problems. Here are some possible inconsistencies to look for: Were all the responses to rating scale questions entered in ascending order (lowest rating is lowest number, highest rating is highest number) with the exception of two or three questions? Did an unusual number of respondents fail to answer only one particular open-ended question? Any discrepancies or errors in data processing are easily identifiable and will be addressed at this stage. Remember, inconsistency is the breeding ground of tabulation error.

Pouring over a clean set of tables is like reading a well-written news article. Everything adds up, everything makes sense. The who, what, when, where, why, and how of your research objective will be answered. If even one of those six questions remains unanswered, it's not yet a finished set of tables. You might want to add another cut to the banner or columns in certain tables (see discussion of banner in this chapter) or perhaps filter some of the tables down to a smaller subset and keep looking. The answers are there waiting to be discovered.

## Graphic Representations of Data

You have probably heard the saying “One picture is worth a thousand words.” Graphic representations of data use pictures rather than tables to present research results. Results—particularly key results—can be presented most powerfully and efficiently through graphs. Some approaches to the display of statistical data are provided in the Practicing Marketing Research feature on page 456.

Marketing researchers have always known that important findings identified by cross tabulation and statistical analysis could be best presented graphically. However, in the early years of marketing research, the preparation of graphs was tedious, difficult, and time-consuming. The advent of personal computers, coupled with graphics software and laser printers, has changed all of this. Spreadsheet programs such as Excel have extensive graphics capabilities, particularly in their Windows versions. In addition, programs designed for creating presentations, such as PowerPoint, permit the user to generate a wide variety of high-quality graphics with ease. With these programs, it is possible to do the following:

- Quickly produce graphs.
- Display those graphs on the computer screen.

# PRACTICING MARKETING RESEARCH



## Doing Frequency and Cross Tabulation Tables in Excel

If you have your data in an Excel spreadsheet or if you can import them into an Excel spreadsheet, then you can use the Pivot Table feature in Excel to produce one-way frequency tables and cross tabulations. The spreadsheet should be prepared so that the columns represent numeric codes for responses to various survey questions and the rows represent responses given by each person surveyed.

To create a one-way frequency table, do the following:

- Select the Data sheet in the Pivot Table template and click on a cell containing data.
- Select the command Pivot Table Report under the Data menu.
- A dialog box will appear. Make sure the Excel list or database is selected. Click the Next button.
- In the Range box, enter the range that contains the database. Click the Next button.
- Another dialog box will appear. It is fairly detailed, with a number of different options. On the right side of the dialog box, you will see a list of all the fields in the database. In this case, they should be survey questions. The different answers to a question should appear as rows in the table. If, for example, you had the label Q1 at the top of the column that includes the responses to Q1 for all respondents, then you would drag the Q1 button into the area labeled ROW and drop it there.
- The button should read Sum of Q1. However, we want a count or frequency for Q1. Double-click on the Sum of Q1 button. In the dialog box that appears, select Count in the list and press the OK button. Count of Q1 should now appear in the DATA area.
- Click the Finish button. The tabulation you requested will appear on a separate sheet.<sup>8</sup>

- Make desired changes and redisplay.
- Print final copies on a laser, inkjet, or dot matrix printer.

All of the graphs shown in this section were produced using a personal computer, a laser printer, and a graphics software package.

## Line Charts

Line charts are perhaps the simplest form of graphs. They are particularly useful for presenting a given measurement taken at several points over time. Exhibit 14.14 shows monthly sales data for Just Add Water, a retailer of women's swimwear. The results reveal similar sales patterns for 2001 and 2002, with peaks in June and generally low sales in January through March and September through December. Just Add Water is evaluating the sales data to identify product lines that it might add to improve sales during those periods.

## Pie Charts

Pie charts are another type of graph that is frequently used. They are appropriate for displaying marketing research results in a wide range of situations. Exhibit 14.15 shows



## PRACTICING MARKETING RESEARCH

### Professional Pointers for the Best Graphic Design of Statistical Data

The graphical presentation of statistical data is a powerful information conveying tool, provided it's done right, and clearly, and effectively. Here are some practical tips from Carl James Schwarz, statistics professor at Simon Fraser University in Burnaby, British Columbia, in Canada, on how to do it:

1. Make sure your graph or chart is complete and self-explanatory, with a title and all axes labeled, so that it presents all the necessary information.
2. Graphing scale is fluid, so any scale will work fine provided the values don't get crowded into a corner or diffused across the page.
3. Typically, the axes of graphs intersect at zero, but you don't have to do it this way provided you clearly mark the starting point.
4. Multiple curves or lines are permissible on a single graph to emphasize comparisons. Try differentiating the different lines in color or with dotted or broken lines. Data values can be distinguished with symbols, such as  $m$  = males and  $f$  = females.

Professor Schwarz also describes seven common graphical errors:

1. Labels and titles are omitted, or axes aren't labeled or units specified.
2. Scales are used incorrectly or their increments change across the graph. Plot your data knowing that most people read a scale left to right.
3. The zero point is misplaced. Readers assume it's at the bottom of the graph, so if it's somewhere else, it can present a misleading impression of the amount of change depicted in the data.
4. The wrong chart type was used. Avoid pie charts. Line charts are preferable (over bar charts) if your horizontal axis represents time.
5. Grid lines were made too dark, left out, or are irrelevant. Good graph paper already has faint grey but serviceable background grid lines.
6. Shading and 3-D effects can often distort a graph rather than enliven it.
7. Dollar amounts were not adjusted for inflation, making comparisons misleading.<sup>9</sup>

#### Exhibit 14.14

#### Line Chart for Sales of Women's Swimwear

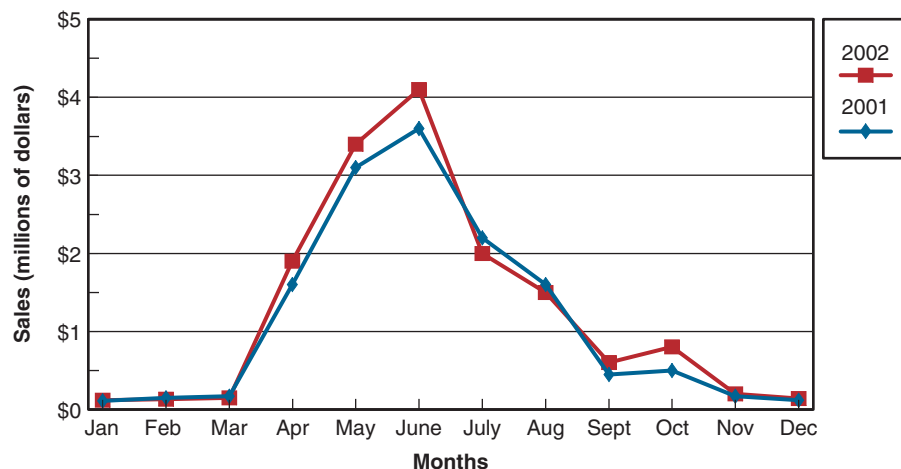
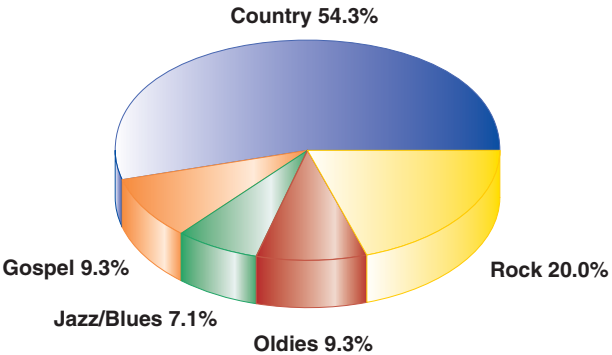


Exhibit 14.15

Three-Dimensional Pie Chart for Types of Music Listened to Most Often



radio music preferences gleaned from a survey of residents of several Gulf Coast metropolitan areas in Louisiana, Mississippi, and Alabama. Note the three-dimensional effect produced by the software.

Bar Charts

Bar charts may be the most flexible of the three types of graphs discussed in this section. Anything that can be shown in a line graph or a pie chart also can be shown in a bar chart. In addition, many things that cannot be shown—or effectively shown—in other types of graphs can be readily illustrated in bar charts. Four types of bar charts are discussed here.

1. *Plain bar chart.* As the name suggests, plain bar charts are the simplest form of bar chart. The same information displayed in the pie chart in Exhibit 14.15 is shown in the bar chart in Exhibit 14.16. Draw your own conclusions regarding whether the pie chart or the bar chart is the more effective way to present this information. Exhibit 14.16 is a traditional two-dimensional chart. Many of the software packages available today can take the same information and present it with a three-dimensional effect, as shown in Exhibit 14.17. Again, decide for yourself which approach is visually more appealing and interesting.
2. *Clustered bar chart.* The clustered bar chart is one of three types of bar charts useful for showing the results of cross tabulations. The radio music preference results are

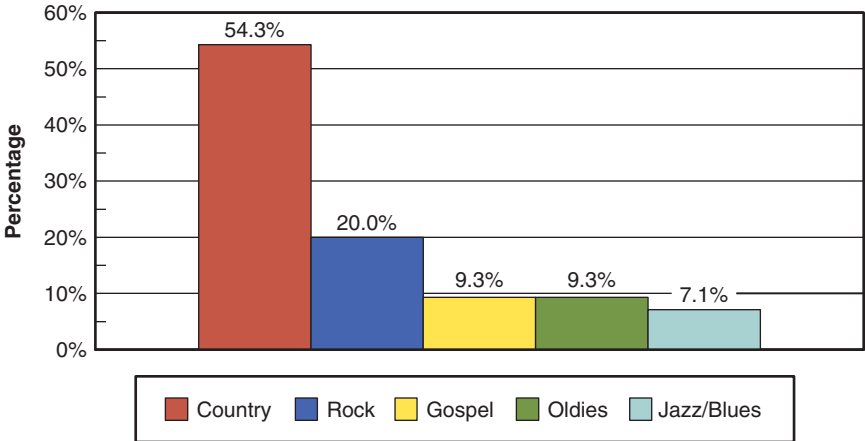
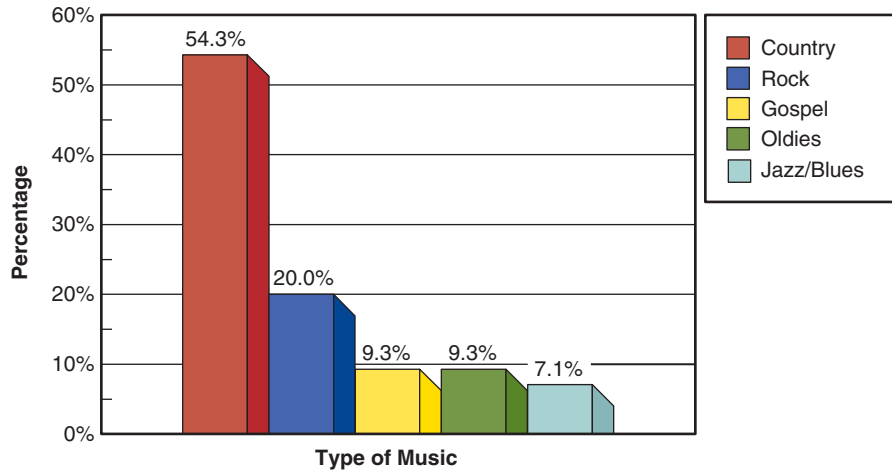


Exhibit 14.16

Simple Two-Dimensional Bar Chart for Types of Music Listened to Most Often

Exhibit 14.17

Simple Three-Dimensional Bar Chart for Types of Music Listened to Most Often



## PRACTICING MARKETING RESEARCH

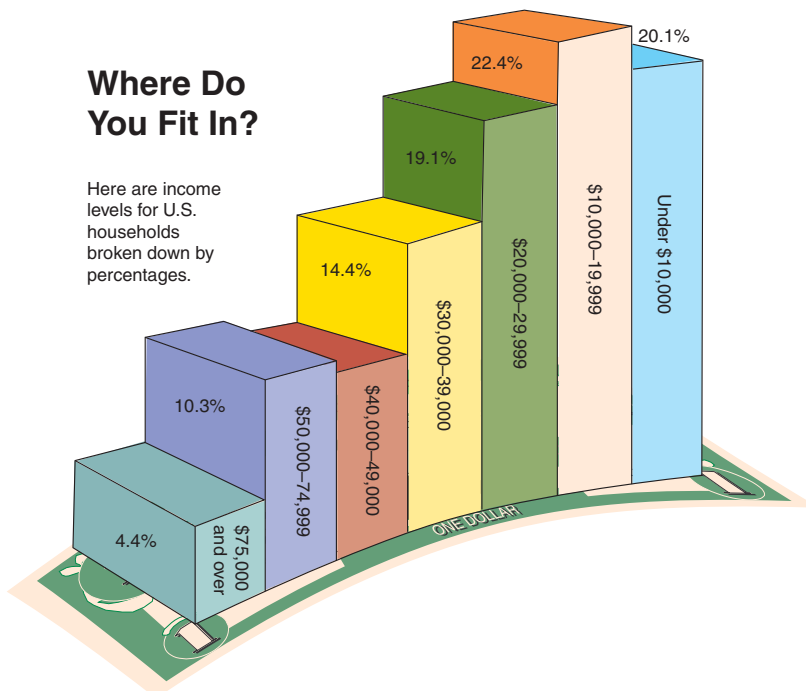
### Learning from the Worst and the Best of Graphic Design for Statistics

**Bad Design:** The three-dimensional effects make reading the bars difficult, and where do you look? The front of the bars, the sides, or the back? The scale, which is nonhorizontal, artificially increases the values for the lower-income

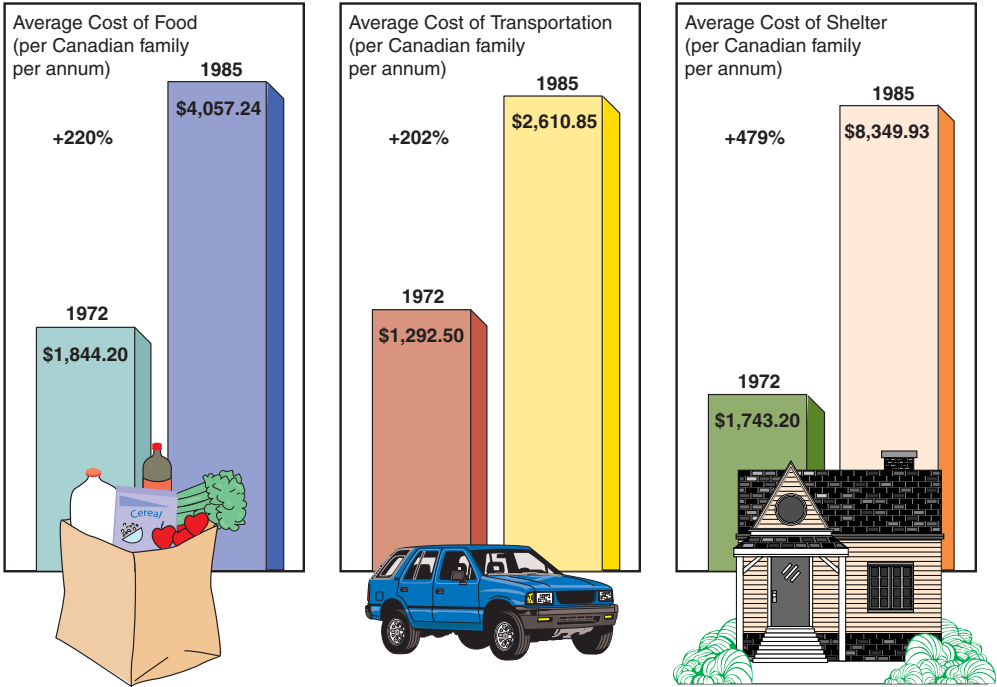
#### Income levels

#### Where Do You Fit In?

Here are income levels for U.S. households broken down by percentages.

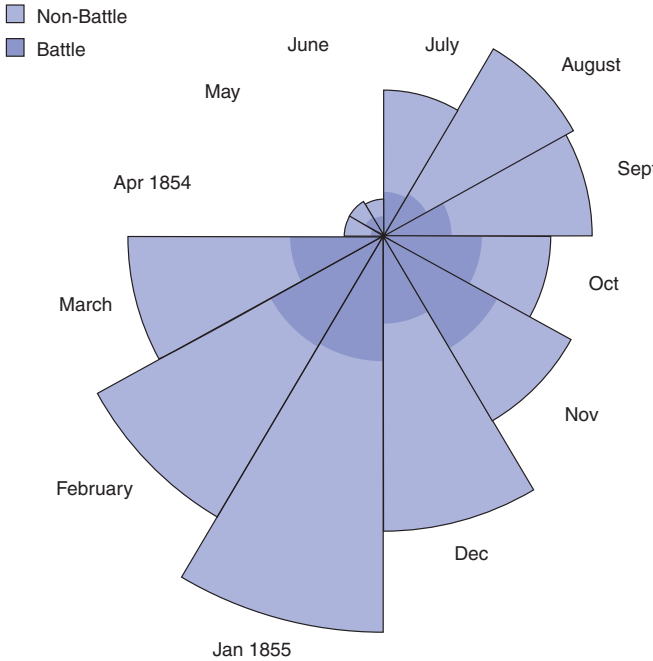


**Cost of Living**



<http://www.math.sfu.ca/~cschwartz/Stat-301/Handouts/node09.html>

**Causes of Mortality in the Army in the East April 1854 to March 1855**



<http://www.math.yorku.ca/SCS/Gallery/images/boxcomb.gif>

bars in comparison to those for upper incomes. At least one bar doesn't have a percentage amount; another is misplaced and hard to find at first glance. Finally, the sizes of the intervals change: first it uses increments of 10,000, then 9,999, then only 9,000, and then 24,999, all of which confuses the representation (analysis by Carl James Schwarz, Simon Fraser University).<sup>10</sup>

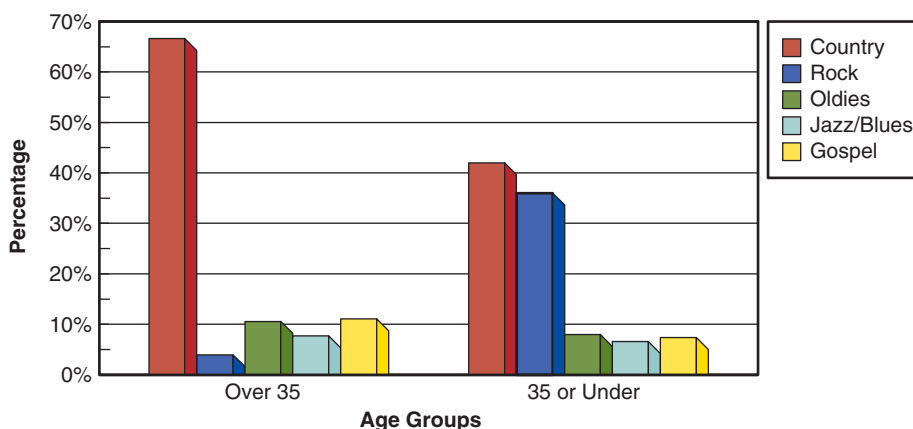
**Bad Design:** The ratio of bar heights in each grouping doesn't reflect the actual ratio, as you can see by comparing the bars for housing with food or transportation. The graphs seem precise, but this is only implied and is unrealistic; it is unlikely that an average can be estimated to the penny. Furthermore, the percentages have been figured incorrectly; for example, when you

double the costs, this represents only a 100 percent increase (analysis by Carl James Schwarz, Simon Fraser University).<sup>11</sup>

**Good Design:** This is a classic in clear, effective, and visually engaging design, even if it is nearly 150 years old. This coxcomb graphical design (also called a polar-area diagram, invented by Florence Nightingale, who was not only a famous nurse but passionate about statistics) is impressive for how it displays frequency by area. It's the same idea as a pie chart but is executed better visually. The coxcomb design maintains constant angles and varies its radius, something the pie chart cannot do (analysis by Michael Friendly, York University).<sup>12</sup>

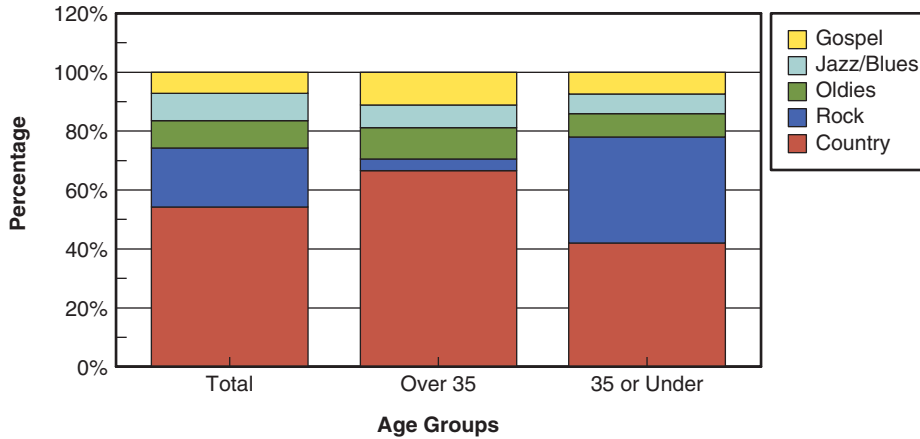
cross tabulated by age in Exhibit 14.18. The graph shows that country music is mentioned most often as the preferred format by those over 35 and those 35 or under. The graph also shows that rock music is a close second for those 35 or under and is least frequently mentioned by those over 35. The results suggest that if the target audience is those in the 35 or under age group, then a mix of country and rock music is appropriate. A focus on country music probably would be the most efficient approach for those over 35.

3. *Stacked bar chart.* Like clustered bar charts, stacked bar charts are helpful in graphically representing cross tabulation results. The same music preference data shown in Exhibit 14.18 are presented as a stacked bar chart in Exhibit 14.19.
4. *Multiple-row, three-dimensional bar chart.* This type of bar chart provides what we believe to be the most visually appealing way of presenting cross tabulation information. The same music preference data displayed in Exhibits 14.18 and 14.19 are presented in a multiple-row, three-dimensional bar chart in Exhibit 14.20.

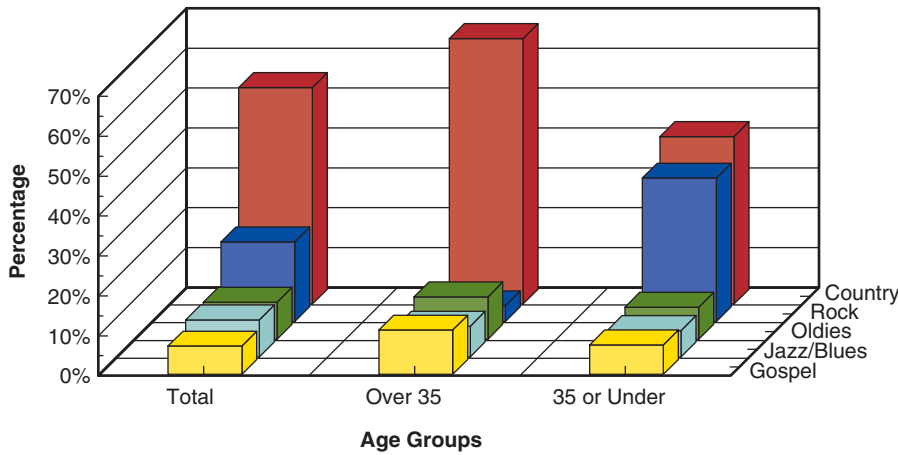


**Exhibit 14.18**

**Clustered Bar Chart  
for Types of Music  
Listened to Most  
Often by Age**



**Exhibit 14.19**  
**Stacked Bar Chart for Types of Music Listened to Most Often by Age**



**Exhibit 14.20**  
**Multiple-Row, Three-Dimensional Bar Chart for Types of Music Listened to Most Often by Age**

## Descriptive Statistics

Descriptive statistics are the most efficient means of summarizing the characteristics of large sets of data. In a statistical analysis, the analyst calculates one number or a few numbers that reveal something about the characteristics of large sets of data.

### Measures of Central Tendency

Before beginning this section, you should review the types of data scales presented in Chapter 9. Recall that there are four basic types of measurement scales: nominal, ordinal, interval, and ratio. Nominal and ordinal scales are sometimes referred to as nonmetric scales, whereas interval and ratio scales are called metric scales. Many of the statistical procedures discussed in this section and in following sections require metric scales, whereas others are designed for nonmetric scales.

The three measures of central tendency are the arithmetic mean, median, and mode. The **mean** is properly computed only from interval or ratio (metric) data. It is computed

**mean**  
 Sum of the values for all observations of a variable divided by the number of observations.

by adding the values for all observations for a particular variable, such as age, and dividing the resulting sum by the number of observations. With survey data, the exact value of the variable may not be known; it may be known only that a particular case falls in a particular category. For example, an age category on a survey might be 18 to 34 years of age. If a person falls into this category, the person's exact age is known to be somewhere between 18 and 34. With grouped data, the midpoint of each category is multiplied by the number of observations in that category, the resulting totals are summed, and the total is then divided by the total number of observations. This process is summarized in the following formula:

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

where  $f_i$  = frequency of the  $i$ th class  
 $X_i$  = midpoint of that class  
 $h$  = number of classes  
 $n$  = total number of observations

#### ▶ median

Value below which 50 percent of the observations fall.

The **median** can be computed for all types of data except nominal data. It is calculated by finding the value below which 50 percent of the observations fall. If all the values for a particular variable were put in an array in either ascending or descending order, the median would be the middle value in that array. The median is often used to summarize variables such as income when the researcher is concerned that the arithmetic mean will be affected by a small number of extreme values and, therefore, will not accurately reflect the predominant central tendency of that variable for that group.

#### ▶ mode

Value that occurs most frequently.

The **mode** can be computed for any type of data (nominal, ordinal, interval, or ratio). It is determined by finding the value that occurs most frequently. In a frequency distribution, the mode is the value that has the highest frequency. One problem with using the mode is that a particular data set may have more than one mode. If three different values occur with the same level of frequency and that frequency is higher than the frequency for any other value, then the data set has three modes. The mean, median, and mode for sample data on beer consumption are shown in Exhibit 14.21.

## Measures of Dispersion

Frequently used measures of dispersion include standard deviation, variance, and range. Whereas measures of central tendency indicate typical values for a particular variable, measures of dispersion indicate how spread out the data are. The dangers associated with relying only on measures of central tendency are suggested by the example shown in Exhibit 14.22. Note that average beer consumption is the same in both markets—3 cans/bottles/glasses. However, the standard deviation is greater in market two, indicating more dispersion in the data. Whereas the mean suggests that the two markets are the same, the added information provided by the standard deviation indicates that they are different.

**EXHIBIT 14.21****Mean, Median, and Mode for Beer Consumption Data**

A total of 10 beer drinkers (drink one or more cans, bottles, or glasses of beer per day on the average) were interviewed in a mall-intercept study. They were asked how many cans, bottles, or glasses of beer they drink in an average day.

Respondent	Number of Cans/ Bottles/Glasses Per Day
1	2
2	2
3	3
4	2
5	5
6	1
7	2
8	2
9	10
10	1

Mode = 2 cans/bottles/glasses  
 Median = 2 cans/bottles/glasses  
 Mean = 3 cans/bottles/glasses

**EXHIBIT 14.22****Measures of Dispersion and Measures of Central Tendency**

Consider the beer consumption data presented in Exhibit 14.21. Assume that interviewing was conducted in two markets. The results for both markets are shown.

Respondent	Number of Cans/ Bottles/Glasses Market One	Number of Cans/ Bottles/Glasses Market Two
1	2	1
2	2	1
3	3	1
4	2	1
5	5	1
6	1	1
7	2	1
8	2	3
9	10	10
10	1	10
Mean	3	3
Standard deviation	2.7	3.7



The formula for computing the standard deviation for a sample of observations is as follows:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

where  $S$  = sample standard deviation  
 $X_i$  = value of the  $i$ th observation  
 $\bar{X}$  = sample mean  
 $n$  = sample size

Occupation is an example of a categorical variable. The only results that can be reported for a variable of this type are the frequency and the relative percentage with which each category was encountered.

The variance is calculated by using the formula for standard deviation with the square root sign removed. That is, the sum of the squared deviations from the mean is divided by the number of observations minus 1. Finally, the range is equal to the maximum value for a particular variable minus the minimum value for that variable.



## Percentages and Statistical Tests

In performing basic data analysis, the research analyst is faced with the decision of whether to use measures of central tendency (mean, median, mode) or percentages (one-way frequency tables, cross tabulations). Responses to questions either are categorical or take the form of continuous variables. Categorical variables such as “Occupation” (coded 1 for professional/man-agerial, 2 for clerical, etc.) limit the analyst to reporting the frequency and relative percentage with which each category was encountered. Variables such as age can be continuous or categorical, depending on how the information was obtained. For example, an interviewer can ask people their actual age or ask them which category (under 35, 35 or older) includes their age. If actual age data are available, mean age can be readily computed. If categories are used, one-way frequency tables and cross tabulations are the most obvious choices for analysis. However, continuous data can be put into categories, and means can be estimated for categorical data by using the formula for computing a mean for grouped data (presented earlier).

Finally, statistical tests are available that can indicate whether two means—for example, average expenditures by men and average expenditures by women at fast-food restaurants—or two percentages differ to a greater extent than would be expected by chance (sampling error) or whether there is a significant relationship between two variables in a cross-tabulation table. These tests are discussed in Chapter 15.

## SUMMARY

Once the questionnaires have been returned from the field, a five-step process takes place. These steps are (1) validation and editing, which are quality control checks, (2) coding, (3) data entry, (4) machine cleaning of data, and (5) tabulation and statistical analysis. The first step in the process, making sure that the data have integrity, is critical. Otherwise, the age-old adage is true: “Garbage in, garbage out.” Validation involves determining with as much certainty as possible that each questionnaire is, in fact, a valid interview. A valid interview in this sense is one that was conducted in an appropriate manner. The objective of validation is

to detect interviewer fraud or failure to follow key instructions. Validation is accomplished by recontacting a certain percentage of the respondents surveyed by each interviewer. Any surveys found to be fraudulent are eliminated from the database. After the validation process is completed, editing begins. Editing involves checking for interviewer and respondent mistakes—making certain that all required questions were answered, that skip patterns were followed properly, and that responses to open-ended questions were accurately recorded.

Upon completion of the editing, the next step is to code the data. Most questions on surveys are closed-ended and precoded, which means that numeric codes already have been assigned to the various responses on the questionnaire. With open-ended questions, the researcher has no idea in advance what the responses will be. Therefore, the coder must establish numeric codes for response categories by listing actual responses to open-ended questions and then consolidating those responses and assigning numeric codes to the consolidated categories. Once a coding sheet has been created, all questionnaires are coded using the coding sheet categories.

The next step is data entry. Today, most data entry is done by means of intelligent entry systems that check the internal logic of the data. The data typically are entered directly from the questionnaires. New developments in scanning technology have made a more automated approach to data entry cost-effective for smaller projects.

Machine cleaning of data is a final, computerized error check of the data, performed through the use of error checking routines and/or marginal reports. Error checking routines indicate whether or not certain conditions have been met. A marginal report is a type of frequency table that helps the user determine whether inappropriate codes were entered and whether skip patterns were properly followed.

The final step in the data analysis process is tabulation of the data. The most basic tabulation involves a one-way frequency table, which indicates the number of respondents who gave each possible answer to each question. Generating one-way frequency tables requires the analyst to determine a basis for percentages. For example, are the percentages to be calculated based on total respondents, number of people asked a particular question, or number answering a particular question? Tabulation of data is often followed by cross tabulation—examination of the responses to one question in relation to the responses to one or more other questions. Cross tabulation is a powerful and easily understood approach to the analysis of survey research results.

Statistical measures provide an even more powerful way to analyze data sets. The most commonly used statistical measures are those of central tendency: the arithmetic mean, median, and mode. The arithmetic mean is computed only from interval or ratio data by adding the values for all observations of a particular variable and dividing the resulting sum by the number of observations. The median can be computed for all types of data except nominal data by finding the value below which 50 percent of the observations fall. The mode can be computed for any type of data by simply finding the value that occurs most frequently. The arithmetic mean is, by far, the most commonly used measure of central tendency.

In addition to central tendency, researchers often want to have an indication of the dispersion of the data. Measures of dispersion include standard deviation, variance, and range.

**validation** Process of ascertaining that interviews actually were conducted as specified.

**editing** Process of ascertaining that questionnaires were filled out properly and completely.

**skip pattern** Sequence in which later questions are asked, based on a respondent's answer to an earlier question or questions.

**coding** Process of grouping and assigning numeric codes to the various responses to a question.

## KEY TERMS & DEFINITIONS



**data entry** Process of converting information to an electronic format.

**intelligent data entry** Form of data entry in which the information being entered into the data entry device is checked for internal logic.

**scanning technology** Form of data entry in which responses on questionnaires are read in automatically by the data entry device.

**machine cleaning of data** Final computerized error check of data.

**error checking routines** Computer programs that accept instructions from the user to check for logical errors in the data.

**marginal report** Computer-generated table of the frequencies of the responses to each

question, used to monitor entry of valid codes and correct use of skip patterns.

**one-way frequency table** Table showing the number of respondents choosing each answer to a survey question.

**cross tabulation** Examination of the responses to one question relative to the responses to one or more other questions.

**mean** Sum of the values for all observations of a variable divided by the number of observations.

**median** Value below which 50 percent of the observations fall.

**mode** Value that occurs most frequently.

## QUESTIONS FOR REVIEW & CRITICAL THINKING

1. What is the difference between measurement validity and interview validation?
2. Assume that Sally Smith, an interviewer, completed 50 questionnaires. Ten of the questionnaires were validated by calling the respondents and asking them one opinion question and two demographic questions over again. One respondent claimed that his age category was 30–40, when the age category marked on the questionnaire was 20–30. On another questionnaire, in response to the question “What is the most important problem facing our city government?” the interviewer had written, “The city council is too eager to raise taxes.” When the interview was validated, the respondent said, “The city tax rate is too high.” As a validator, would you assume that these were honest mistakes and accept the entire lot of 50 interviews as valid? If not, what would you do?
3. What is meant by the editing process? Should editors be allowed to fill in what they think a respondent meant in response to open-ended questions if the information seems incomplete? Why or why not?
4. Give an example of a skip pattern on a questionnaire. Why is it important to always follow the skip patterns correctly?
5. It has been said that, to some degree, coding of open-ended questions is an art. Would you agree or disagree? Why? Suppose that, after coding a large number of questionnaires, the researcher notices that many responses have ended up in the “Other” category. What might this imply? What could be done to correct this problem?
6. Describe an intelligent data entry system. Why are data typically entered directly from the questionnaire into the data entry device?
7. What is the purpose of machine cleaning data? Give some examples of how data can be machine cleaned. Do you think that machine cleaning is an expensive and unnecessary step in the data tabulation process? Why or why not?

8. It has been said that a cross tabulation of two variables offers the researcher more insightful information than does a one-way frequency table. Why might this be true? Give an example.
9. Illustrate the various alternatives for using percentages in one-way frequency tables. Explain the logic of choosing one alternative method over another.
10. Explain the differences among the mean, median, and mode. Give an example in which the researcher might be interested in each of these measures of central tendency.
11. Calculate the mean, median, mode, and standard deviation for the following data set:

Respondent	Times Visited Whitehall Mall in Past 6 Months	Times Visited Northpark Mall in Past 6 Months	Times Visited Sampson Mall in Past 6 Months
A	4	7	2
B	5	11	16
C	13	21	3
D	6	0	1
E	9	18	14
F	3	6	8
G	2	0	1
H	21	3	7
I	4	11	9
J	14	13	5
K	7	7	12
L	8	3	25
M	8	3	9

12. Enter the following data into an Excel spreadsheet. Include the column headings (Q1, Q2, and Q3), as well as the numeric values. The definitions of the numeric values are provided at the bottom of the table. Use the Pivot Table feature in Excel (found under the Data option) to cross tabulate the likelihood of purchase (row) by gender (column) and income level (column). What conclusions can you draw about the relationship between gender and likelihood of purchase and that between income and likelihood of purchase?

Respondent	Likelihood of Purchase	Gender	Income
A	5	2	3
B	4	2	3
C	4	2	2
D	3	1	2
E	1	1	2
F	5	2	3
G	5	2	3
H	4	1	3
I	1	1	2
J	1	1	2
K	2	1	1
L	5	2	3

Respondent	Likelihood of Purchase	Gender	Income
M	5	2	3
N	4	1	3
O	3	1	2
P	3	1	2
Q	4	2	3
R	5	2	3
S	2	1	1
T	2	1	1

Likelihood of purchase: very likely = 5, likely = 4, undecided = 3, unlikely = 2, very unlikely = 1

Gender: male = 1, female = 2

Income: under \$30,000 = 1, \$30,000 to \$75,000 = 2, over \$75,000 = 3

13. Using data from a newspaper or magazine article, create the following types of graphs:
- Line graph
  - Pie chart
  - Bar chart

## WORKING THE NET

What is statistics? To find out, go to the Web site of the American Statistical Association at <http://www.amstat.org/>. Here you will find many resources by going to the menu items on the left side of the page, including career information, publications, and various links.

## REAL-LIFE RESEARCH • 14.1

### Taco Bueno

Taco Bueno has recently opened its 15th store in Utah. Currently, the chain offers tacos, enchiladas, and burritos. Management is considering offering a supertaco that would be approximately two and a half times as large as a regular taco and would contain 5 ounces of ground beef. The basic taco simply has spiced ground beef, lettuce, and cheese. Management feels that the supertaco ought to have more toppings. Therefore, a marketing research study was undertaken to determine what those toppings should be. A key question on the survey was, “What, if anything, do you normally add to a taco that you have prepared at home besides meat?” The question is open-ended, and the coding categories that have been established for the question are shown in the table.

Responses	Code
Avocado	1
Cheese (Monterey Jack/cheddar)	2
Guacamole	3
Lettuce	4
Mexican hot sauce	5

Responses	Code
Olive (black/green)	6
Onion (red/white)	7
Peppers (red/green)	8
Pimienta	9
Sour cream	0
Other	X

## Questions

- How would you code the following responses?
  - I usually add a green, avocado-tasting hot sauce.
  - I cut up a mixture of lettuce and spinach.
  - I'm a vegetarian; I don't use meat at all. My taco is filled only with guacamole.
  - Every now and then, I use a little lettuce, but normally I like cilantro.
- Is there anything wrong with having a great number of responses in the "Other" category? What problems does this present for the researcher?

## PrimeCare

PrimeCare is a group of 12 emergency medical treatment clinics in the Columbus and Toledo, Ohio, markets. The management group for PrimeCare is considering a communications campaign that will rely mainly on radio ads to boost its awareness and quality image in the market. The ad agency of Dodd and Beck has been chosen to develop the campaign. Currently, the plan is to focus on PrimeCare's experienced, front-line health professionals. Two themes for the ad campaign are now being considered. One theme ("We are ready!") focuses primarily on the special training given to PrimeCare's staff, and the second theme ("All the experts—all the time") focuses primarily on PrimeCare's commitment to having the best, trained professionals available at each PrimeCare facility 24/7. Dodd and Beck's research team has conducted a survey of consumers to gauge the appeal of the two campaigns. Overall results and results broken down by gender and by location are shown in the table.

	Total	Gender		Location	
		Male	Female	Columbus	Toledo
Total	400	198	202	256	144
	100%	100.0%	100.0%	100.0%	100.0%
Prefer "Ready" campaign	150	93	57	124	26
	37.5%	47.0%	28.2%	48.4%	18.1%
Prefer "All the Time" campaign	250	105	145	132	118
	62.5%	53.0%	71.8%	51.6%	81.9%

## REAL-LIFE RESEARCH • 14.2

## Questions

1. Which theme appears to have more appeal overall? Justify your answer.
2. How does the appeal of the campaigns differ between men and women? What is your basis for that conclusion?
3. Are the campaign themes equally attractive to residents of Columbus and Toledo? Why do you say that?



## SPSS EXERCISES FOR CHAPTER 14



### Exercise #1: Machine Cleaning Data

1. Go to the Wiley Web site at [www.wiley.com/college/mcdaniel](http://www.wiley.com/college/mcdaniel) and download the *Segmenting the College Student Market for Movie Attendance* database to SPSS Windows. This database will have several errors for you to correct. In the SPSS Data Editor, go to the *variable view* option and notice the **computer coding** for each variable.
2. Also from the Wiley Web site, download a copy of the *Segmenting the College Student Market for Movie Attendance* questionnaire. Notice the computer coding for each of the variables; which is the same as that in the *variable view* option on the SPSS Data Editor. This information will be important in finding errors in the database.
3. In the SPSS Data Editor, invoke the *analyze/descriptive statistics/frequencies* sequence to obtain frequencies for all of the variables in the database.
4. From the SPSS Viewer *output screen*, determine which variables have input errors. Summarize the errors using the template below as a guide.

Questionnaire Number	Variable Containing error	Incorrect Value	Correct Value
----------------------	---------------------------	-----------------	---------------

Going back to the *data view* screen of the *SPSS Data Editor*.

5. Another possible source of errors is in question 8. Notice that in this question the sum of the answers should be 100 percent. Create a summated variable for question 8 (Q8a + Q8b + Q8c + Q8d) to check for errors by invoking the *transform/compute* sequence. Now, compute a frequency distribution for Q8sum. The values that are not “100” indicate an input error. (Such an error could be the result of the respondent not totaling percentages to 100, but for this machine cleaning exercise, the assumption is that it is an input error.) Summarize the errors using the template above.
6. Once you have completed summarizing the variables containing errors, go back to the *data view* screen of the *SPSS Data Editor*. Position the cursor on each of the

variables containing errors. Use the *ctrl-f* function to find the questionnaire numbers where the errors occurred. At this point, you will need the corrected database, or the database with no errors. Your professor has access to this database with no errors. After getting the corrected database, finish filling in the table in part (4) above with the correct values. Then make the changes in your database, so that you have a database with no errors. Be sure to resave your database after correcting it for errors.

7. After machine cleaning your data, rerun the *analyze/descriptive statistics/frequencies* sequence to obtain frequencies for your corrected database.
8. You will use the results of this exercise to answer the questions in Exercises #2 and #4.

## Exercise #2: Analysis of Data with Frequency Distributions

If you did not complete Exercise #1, you will need the corrected database from your professor. After getting the corrected database, use the *analyze/descriptive statistics/frequencies* sequence to obtain frequency distributions for all of the variables in your database except the questionnaire number (Q No).

If you completed Exercise #1, you will have a corrected database, which consists of frequency distributions for each of the variables in the database.

Answer the following questions.

1. What percentage of all respondents attended at least **1** movie in the past year?  
\_\_\_\_\_ %
2. What percentage of all respondents *never buy food items* at a movie? \_\_\_\_\_ %
3. Produce a table indicating the percentage of all respondents that consider each of the movie theater items in question 5 of the questionnaire *very important*. List the top 5 movie items in descending order (start with the movie items that have the highest percentage of *very important* responses).

### For Example:

Movie Item	Percentage of Respondents
Movie item with the highest percentage	75.0%
Movie item with the 2nd highest percentage, etc.	39.2%

4. What percentage of respondents consider the “newspaper” a *very important* source of information about movies playing at movie theaters? \_\_\_\_\_ %
5. What percentage of respondents consider the “Internet” a *very unimportant* source of information about movies playing at movie theaters? \_\_\_\_\_ %
6. By observing the distribution of responses for Q8a, Q8b, Q8c, and Q8d, which is the most popular *purchase option* for movie theater tickets? \_\_\_\_\_
7. Produce a table listing in descending order the percentage of respondents that consider each of the movie theater information sources (Q7) *very important*.



For example:

Movie Theater Information Sources	Percentage of Respondents Indicating <i>Very Important</i>
Internet	55%
Newspaper	31%

### Exercise #3: Analysis of Data with Descriptive Statistics

If you did not complete Exercises #1 or #2, you will need the corrected database from your professor. The objective of this exercise is to analyze data using measures of central tendency and measures of dispersion. To analyze means and standard deviations, use the *analyze>descriptive statistics>descriptives* sequences. To analyze medians and modes, use the *analyze>descriptive statistics>frequencies* sequence, and select *statistics*. You will see the box with all three measures of central tendency (mean, median and mode).

On the questionnaire, question 5 utilizes a 4-point Itemized Rating scale (illustrated below). This scale is balanced and can be assumed to yield interval scale/metric data. Given the preceding, invoke SPSS to calculate the mean and standard deviation for all of the variables in question 5 (Q5a–Q5i).

Very unimportant	Somewhat unimportant	Somewhat important	Very important
1	2	3	4

Answer the following questions.

- Using only the **mean** for each of the variables, which of the movie theater items was considered “most important”? \_\_\_\_\_
- Using only the **standard deviation** for each of the variables, for which question was there the greatest amount of agreement? \_\_\_\_\_  
*Hint:* Least amount of dispersion regarding the response to the movie item
- Questions 4 and 6 utilize multiple-choice questions that yield nonmetric data, but that are ordinal scale. The appropriate measures of central tendency for nonmetric data are the median and the mode.
  - What is the *median* response for question 4, concerning the amount a person spends on food/drink items at a movie? \_\_\_\_\_

Never buy food items at movies (0)	Up to \$7.49 (1)	\$7.50 to \$14.99 (2)	\$15.00 or more (3)
------------------------------------	------------------	-----------------------	---------------------

- Concerning question 6, the distance a person would drive to see a movie on a “big screen,” what is the *mode* of that distribution of responses?

Zero (0)	1 to 9 miles (1)	11 to 24 miles (2)	25 to 49 miles (3)	50+ miles (4)
-------------	---------------------	-----------------------	-----------------------	------------------

4. In this question the objective will be to compare the results of median and mean responses for Q3.
- a. Mean response: \_\_\_\_\_
  - b. Median response: \_\_\_\_\_
  - c. Standard deviation: \_\_\_\_\_
  - d. Minimum response: \_\_\_\_\_
  - e. Maximum response: \_\_\_\_\_
5. When the responses to a question contain extreme values, the mean response can lie in the upper or lower quartile of the response distribution. In such a case, the median value would be a better indicator of an average response than the mean value. Given the information you obtained from answering #4 above, is the mean or median a better representative of the “average” response to Q3?

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

### Exercise #4: Analysis of Demographic Characteristics Using Charts

If you completed Exercise #1 and/or Exercise #2 you will have the information to complete this exercise.

If you did not complete either Exercise #1 or #2, you will need to get a corrected soft drink database from your professor. After getting the database, use the *analyze/descriptive statistics/frequencies* sequence to obtain frequency distributions for the demographic questions (questions 11–14).

Complete the following.

1. Display the demographic data for each of the four demographic variables in tables.
2. For each demographic variable, illustrate the table results using some type of graphic representation of the data (pie charts, line charts, or bar charts).

*Note:* Some students who are proficient in Excel may want to paste their databases into an Excel spreadsheet for the geographical depiction of the demographic variables.

