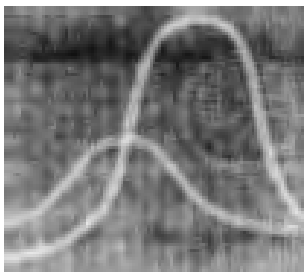


Statistical Methods for Health Care Research

FIFTH EDITION

Barbara Hazard Munro, PhD, FAAN
Dean and Professor
William F. Connell School of Nursing
Boston College
Chestnut Hill, Massachusetts

LIPPINCOTT WILLIAMS & WILKINS
A Wolters Kluwer Company
Philadelphia • Baltimore • New York • London
Buenos Aires • Hong Kong • Sydney • Tokyo

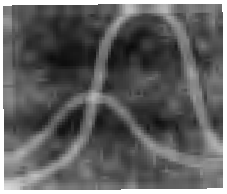
Senior Manufacturing Manager: William Alberti
Compositor: TechBooks
Printer: R.R. Donnelley—Crawfordsville

5th Edition

9 8 7 6 5

Library of Congress Cataloging-in-Publication Data
Munro, Barbara Hazard.
Statistical methods for health care research / Barbara Hazard Munro.—5th ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 13: 978-0-7817-4840-7
ISBN 10: 0-7817-4840-2 (alk. paper)
1. Nursing—Research—Statistical methods. 2. Medical care—Research—Statistical methods. I. Title.
[DNLM: 1. Health Services Research—methods. 2. Statistics. WA 950 M968s 2005]
RT81.5.M86 2005
610'.7'27—dc22
2004007098

LWW.com



Contributors

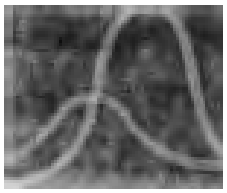
Karen J. Aroian, PhD, RN, CS, FAAN
Professor of Nursing Research
College of Nursing
Wayne State University
Detroit, Michigan

Jane Karpe Dixon, PhD
Professor, Doctoral Program
School of Nursing
Yale University
New Haven, Connecticut

Mary E. Duffy, PhD, FAAN
Professor & Director, Center for Nursing Research
School of Nursing
Boston College
Chestnut Hill, Massachusetts

Barbara S. Jacobsen, MS
Professor (Retired)
The University of Pennsylvania School of Nursing
Philadelphia, Pennsylvania

Anne E. Norris, PhD, RN, CS
Associate Professor
School of Nursing
Boston College
Chestnut Hill, Massachusetts



Reviewers

Denise Côté-Arsenault, PhD, RN
Associate Professor of Nursing
Syracuse University
Syracuse, New York

Karen K. Badros, EdD, CRNP
Professor of Nursing
Salisbury University
Salisbury, Maryland

Vera Brancato, EdD, MSN, RN, BC
Associate Professor of Nursing
Kutztown University
Kutztown, Pennsylvania

Anne Folta Fish, PhD, RN
Associate Professor
University of Missouri-St. Louis
St. Louis, Missouri

Stephen D. Krau, PhD, RN, BSN, BA, MSN, MA
Professor and Coordinator of Continuing Education
Middle Tennessee State University
Murfreesboro, Tennessee

Sarah Newton, PhD, RN
Associate Professor
Oakland University School of Nursing
Rochester, Michigan

Linn Stranak, PhD
Professor and Department Chair
of Physical Education, Wellness and Sport
Union University
Jackson, Tennessee

Donald E. Stull, PhD
Associate Professor
University of Maryland School of Nursing
Baltimore, Maryland

Laura Talbot
Johns Hopkins University
Baltimore, Maryland

Mary A. (Sandy) Wyper, PhD, RN
Associate Professor
Ursuline College
Pepper Pike, Ohio

Nashat Zuraikat, PhD, RN
Graduate School Coordinator
Indiana University of Pennsylvania
Indiana, Pennsylvania

Preface

The purpose of the first edition of Statistical Methods for Health Care Research was to acquaint the reader with the statistical techniques most commonly reported in the research literature of the health professions. We attempted to make the book user-friendly by keeping mathematical symbolism to a minimum and by using computer printouts and examples from the literature to demonstrate specific techniques. In the second edition, we further reduced mathematical equations, moved from mainframe to personal computer examples, and added new techniques such as logistic regression. In the third edition, we added a dataset for use with the exercises at the end of the chapters. In the fourth edition, we incorporated suggestions from students and reviewers to clarify complex statistical procedures.

Once again, in this fifth edition, we have updated the examples from the literature and included additional content primarily in the area of preparing data for statistical analyses. We believe that it is essential that one spend time preparing the data prior to running statistical analyses. We, therefore, have added a section on the principles for preparing data for analyses and included more detail on carrying out data transformations. We also have expanded the sections on handling outliers and dealing with missing data.

The concepts of sensitivity and specificity are now part of data analyses. We have added an introductory section on these concepts, and, in the chapter on logistic regression, we demonstrate how to interpret the printout in terms of sensitivity and specificity.

The statistical software has been updated using SPSS version 12.0. The exercises at the end of each chapter are based on the database provided on a CD-ROM in the back of the book. Each year, our students add additional cases to the database, and we encourage you to have your students add cases for their use, as well. We continue to underplay the role of mathematical calculations, assuming that readers will be using a personal computer for statistical analyses.

As a support for instructors, we have produced PowerPoint presentations for each chapter. Go to the LWW connection web site to access the full set of PowerPoint slides, which include guidelines for using SPSS. www.connection.LWW.com

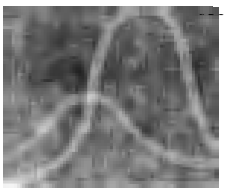
TEXT ORGANIZATION

We have organized the text into two sections:

Section I includes three chapters that present content essential to Understanding the Data. Content includes organizing and displaying data, univariate descriptive statistics, inferential statistics, hypothesis testing, and dealing with missing data and outliers.

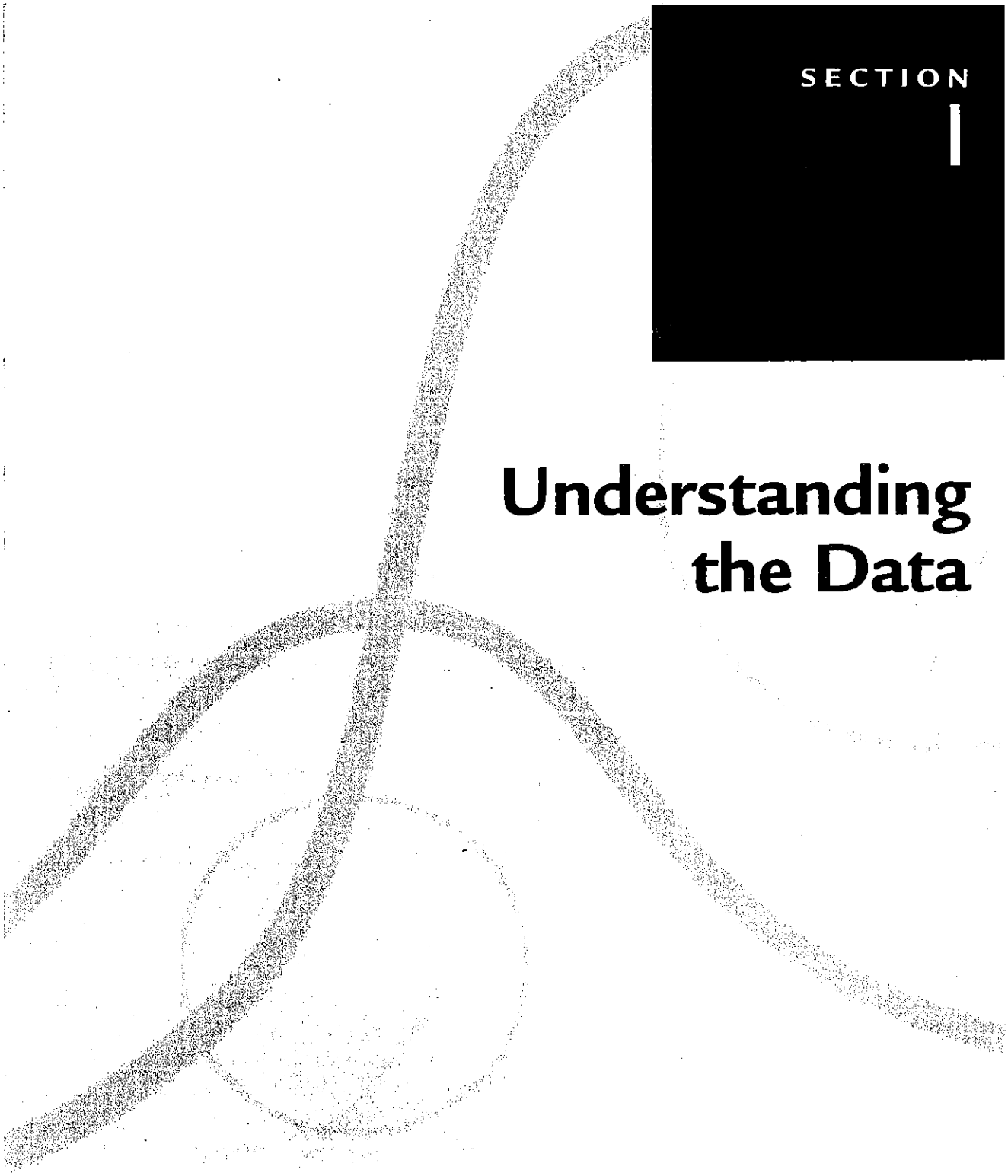
Section II presents 14 chapters that address Specific Statistical Techniques including nonparametric techniques, *t* tests, one-way and multifactorial analysis of variance, analysis of covariance, repeated measures analysis of variance, correlation, regression, canonical correlation, logistic regression, factor analysis, confirmatory factor analysis, path analysis, and structural equation modeling.

Once again, we would like to thank the users and reviewers of the first four editions who made very helpful suggestions for this fifth edition. The students at Boston College, The University of Pennsylvania, and Yale University who have taken courses taught by authors of this text have most definitely played a role in the continuing development of this text, and we thank them too.



Contents

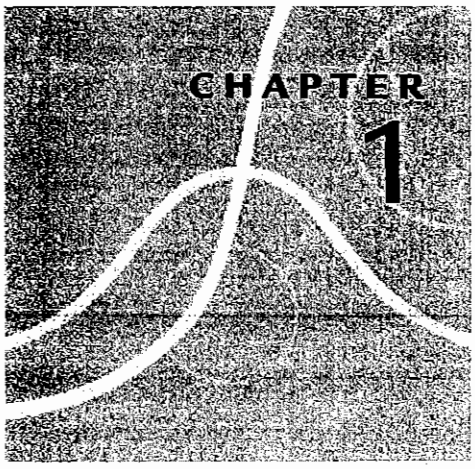
APPENDICES



SECTION

I

Understanding the Data



Organizing and Displaying Data

Mary E. Duffy
Barbara S. Jacobsen

Objectives for Chapter 1

After reading this chapter, you should be able to do the following:

1. Discuss the nature, purpose, and types of statistics.
2. Discuss variables, levels of measurement, and their relationship to statistical analysis.
3. Discuss principles of data handling.
4. Interpret a frequency distribution created by a computer program.
5. Organize data into a table.
6. Interpret data presented in a chart.

Research is the systematic study of one or more problems, usually posed as research questions germane to a specific discipline. Quantitative research uses specific methods to advance the science base of the discipline by studying phenomena relevant to the goals of that discipline. Quantitative research methods include experiments, surveys, correlational studies of various types, and some commonly encountered procedures such as meta-analysis and psychometric evaluations (Knapp, 1998).

Usually when researchers collect data to answer specific quantitative research questions, they want to draw conclusions about a broader base of people, events, or objects than those actually included in the particular study. For example, a researcher may want to draw conclusions about how effective a telephone-delivered coaching intervention delivered by Registered Nurses (RNs) is in relieving postoperative distress in patients having knee surgery in a same-day, ambulatory surgical setting. Yet the researcher selects only a specific number of these patients to study, not the total group of knee surgery patients. The larger group of patients the researcher wants to draw conclusions about is called the *population*; the term *parameter* is

used when describing the characteristics of the population. The group of patients the researcher actually studies is called the *sample*; the term *statistic* is used to describe the characteristics of this group. In most studies, the population parameters are not known and must be estimated from the sample statistics (Norusis, 2002).

THE NATURE OF STATISTICS

Statistics is a branch of applied mathematics that deals with collecting, organizing, and interpreting data using well-defined procedures. Researchers use a variety of techniques to gather these data, which become the observations used in statistical analyses. Thus, the raw materials of research are data, gathered from a sample that has been selected from a population. Applying statistics to these data permits the researcher to draw conclusions and to understand more about the sample from whence the data were obtained.

The purpose of statistics is threefold: to describe and summarize information, thereby reducing it to smaller, more meaningful sets of data; to make predictions or to generalize about occurrences based on observations; and to identify associations, relationships, or differences between the sets of observations. When our goal is to summarize data, we take a large mass of unorganized bits of information and reduce them to smaller sets that describe the original data without sacrificing critical elements. If our goal is to make predictions or to generalize about occurrences of data, we use statistics as an inferential measuring tool. This permits us to state the degree of confidence we have in the accuracy of the measurements we make in a specific research context. When we want to identify associations, relationships, or differences between variables of interest, we are using knowledge about one set of data to infer or predict characteristics about another set of data. Each statistical technique discussed in this book serves one or more of these purposes.

There are two main types of statistics: descriptive and inferential. *Descriptive statistics* are used to describe or characterize data by summarizing them into more understandable terms without losing or distorting much of the information. Summary tables, charts, frequencies, percentages, and measures of central tendency are the most common statistics used to describe basic sample characteristics. In contrast, *inferential statistics* consist of a set of statistical techniques that provide predictions about population characteristics based on information in a sample from that population. The primary focus of most research is the parameters of the population under study; the sample and statistics describing it are important only insofar as they provide information about the population parameters. Thus, an important aspect of statistical inference involves reporting the likely accuracy, or degree of confidence, of the sample statistic that predicts the value of the population parameter (Agresti & Finlay, 1997).

VARIABLES AND THEIR MEASUREMENT

Data are the raw materials of research. The most common way a researcher acquires data is to design a study that will answer a specific research question. The researcher then attempts to answer the question by collecting information (data)

about the characteristics of interest in the study, usually people, events, or objects. Once collected, the data must be organized, examined, and interpreted using well-defined procedures. Almost all quantitative studies involve data that are entered into a computer-based statistical spreadsheet for subsequent data analysis. The logistics and time required to collect data, enter it into a statistical spreadsheet, and prepare it for data analysis are often greatly underestimated and poorly understood. Davidson (1996) recommends taking control of the structure and flow of your data from the beginning. Hopefully, this will help eliminate faulty data leading to faulty conclusions.

In research, the specific characteristics of interest are commonly called variables. *A variable* is a characteristic being measured that varies among the persons, events, or objects being studied. Measurement, in the broadest sense, is the assignment of numerals to objects or events according to a set of rules (Stevens, 1946). For example, the length of a piece of paper can be measured by following a set of rules for placing a graduated straightedge (eg, a ruler) and then reading the numeral that corresponds to the concept of the paper's length. This definition can be broadened to include the assignment of numerals to abstract, intangible concepts such as resilience, self-esteem, and health status. After determining a method of measurement for the concept, it is then called a variable (ie, a measured characteristic that takes on different values). Stevens (1946) noted four types of measurement scales for variables: nominal, ordinal, interval, and ratio. When analyzing data, the first task is to be aware of the type of measurement scale for each of the variables, because this knowledge helps in deciding how to organize and display data.

Nominal Scales

This type of scale, the lowest form of measurement, allows the researcher to assign numbers that classify characteristics of people, objects, or events into categories. Sometimes nominal variables are called *categorical* or *qualitative*. These numeric values are usually assigned to the categories as labels for computer storage, but the choice of numerals for those labels is absolutely arbitrary. Some examples follow:

Variables	*Values*
Group Membership	1 = Experimental
	2 = Placebo
	3 = Routine
Gender	0 = Female
	1 = Male
Adherence to Scheduled Appointment	0 = Did not keep appointment
	1 = Kept appointment

Ordinal Scales

In this case, the characteristics are placed in categories *and* the categories are ordered in some meaningful way (ie, the assignment of numerals is not arbitrary). Ordinal measures can be ranked from high to low. The distance between the categories, however, is unknown. Summated rating scales, as exemplified by the popular Likert scale, are examples of ordinal scales. For example, an RN who works in long-term care could rank patients on their ability to carry out activities of daily living (ADLs): 3 = fully able to perform all ADLs, 2 = partially able to perform ADLs, or 1 = not able to perform any ADLs independently. For the record, however, it is irrelevant how many ADLs fall into each category. Some other examples follow:

Variables	***Values***
Socioeconomic status	1 = Low
	2 = Middle
	3 = High
Health Status	1 = Very Poor
	2 = Poor
	3 = Fair
	4 = Good
	5 = Excellent
Pain Intensity	0 = No pain
	1 = Minor/Little Pain
	2 = Moderate Pain
	3 = Severe Pain

Interval Scales

For this type of scale, the distances between these ordered category values are equal because there is some accepted physical unit of measurement. Because the units are in equal intervals, it is possible to add and subtract across an interval scale. You can say that the difference between 5 and 10 is the same amount (5) as the difference between 75 and 80. An interval scale provides information about the rank ordering of categories and the magnitude of the difference between different values on the scale. Interval variables may be *continuous* (ie, in theory, they may take on any numerical value within the variable's range), or they may be discrete (ie, takes on only a finite number of values between two points). A good example of interval-level measurement is the Fahrenheit scale of temperature.

Ratio Scales

The fourth and most precise level of measurement consists of meaningfully ordered characteristics with equal intervals between them and the presence of a zero point that is not arbitrary but is determined by nature. Blood pressure, pulse rate, and weight are

examples of ratio variables. With ratio scales, all mathematical operations (addition, subtraction, multiplication, division) are possible. Thus, one can say that a 200-pound man is twice as heavy as a 100-pound man. The distinction between interval and ratio variables is interesting, but for the purposes of this text, these two types of variables are handled the same way in analyzing data when the assumptions underlying the statistical test are met.

Measurement Scale Considerations

Researchers need to be very clear about the measurement levels of their study variables, particularly when it comes to classifying variables as either ordinal, interval, or ratio (Burns & Grove, 2001). When measuring variables derived from psychosocial scales, psychological inventories, or tests of knowledge, there may be differences of opinion as to the variable's level of measurement. Many of these scales have arbitrary zero points as determined by the test developer, and they have no accepted unit of measurement comparable to a standard ruler measurement of inches and feet. Technically, these variables are ordinal in nature, but in practice, researchers often think of them as interval- or ratio-level scales. This has been a controversial issue in the research literature for years. Gardner (1975) reviewed the early literature on this conflict, and Knapp (1990) has commented on more recent literature. In his original article on measurement (1946) and in a later article (1968), Stevens noted that treating ordinal scales as interval or ratio scales may violate a technical canon, but in many instances the outcome had demonstrable use. More recently, Knapp (1990, 1993) and Wang, Yu, Wang, and Huang (1999) pointed out that such considerations as measurement perspective, the number of categories that make up an ordinal scale, the concept of *meaningfulness*, and keeping in mind the relevancy of measurement scales to permissible statistics may be important in deciding whether to treat a variable as ordinal or interval. We recommend these articles for further reading on this topic.

It is usually best to gather data at the highest level of measurement for research variables because this permits the researcher to perform more mathematical operations and gain greater precision in measurement. However, interval or ratio variables can be converted to ordinal or nominal variables. For example, diastolic blood pressure, as measured by a sphygmomanometer, is a ratio variable. However, for research purposes, if blood pressure is recorded as either controlled or uncontrolled, then it is a nominal variable. In this case, there is a physiologic basis for such a dichotomous division. But when no such reason exists, converting interval or ratio variables to lower-level nominal or ordinal variables can be unwise because it results in a loss of information. Cohen (1983) detailed the amount of degradation of measurement as a consequence of dichotomization and urged researchers to use all of the original measurement information.

PRINCIPLES OF DATA HANDLING

Traditionally, very little has been written about the principles of getting research data ready for statistical analysis. In recent years, however, the principles of data handling have begun to be written about. Davidson (1996) delineates major

principles of statistical data handling to fill the gap between getting data into the computer and running statistical tests. These principles are based on his view that data handling has certain universal concepts that apply no matter what the data-gathering context or the computer software used. We encourage you to read Davidson's book for a full listing of his principles. We have listed below those principles that relate directly to data collection, input, manipulation, and debugging.

Atomicity Principle: You cannot analyze below the data level that you observe. For example, you gather information about study participants' ages by asking them to circle the number that best reflects their chronological age. For example,

1. 21–25 years
2. 25.1–29.9 years
3. 30–39.9 years
4. 40 or more years

This age variable is measured at the nominal (categorical) level, the lowest form of measurement. Had you asked for participants' ages using a higher form of measurement (ie, What is your age in years?_____), you would be able to manipulate the age variable to produce measures of central tendency, including respondents' average age, standard deviation, and variance. With the lower-level nominal-level age data, the frequency and related percent of persons falling within the stated categories are the best information about age available to you.

Appropriate Data Principle: You cannot analyze what you do not measure. If you want to know a respondent's age but do not gather such information, then you can't use age as a variable in subsequent analyses. Adhering to this principle requires that you anticipate what variables might be needed to explain the results of your data analyses.

Social Consequences Principle: Data about people are about people. Data can have social consequences. Suppose you are gathering information on how effective (ie, efficacy) a specific nonpharmocologic pain management intervention is in relieving pain in patients with chronic headaches. You find that the intervention does not significantly differ from the standard method of giving nonsteroidal anti-inflammatory drugs (NSAIDS) every 4 hours as needed. Thus, it would not be appropriate, and possibly unethical, to counsel them to use the nondrug intervention rather than take a dose of NSAIDS that works.

Data Control Principle: Take control of the structure and flow of your data. Even if you are not going to be the person who enters data into a statistical or other computer program, you should take responsibility for developing and monitoring the procedure for the layout of each respondent's data record (ie, a codebook for how data will be entered into the program; data entry and data manipulations such as recoding variable levels and computing new variables).

Data Efficiency Principle: Be efficient in getting your data into a computer, but not at the cost of losing crucial information. For example, do not hand-total respondents' scores on a 10-item self-esteem scale and then enter only the total score as the measure of their self-esteem in each electronic respondent's data record. By so doing, you are not able to determine the internal consistency reliability of the self-esteem scale because you chose not to enter the items that formed the self-esteem scale score into respondents' data records. Thus, you have sacrificed scientific rigor in favor of efficient data entry.

Change Awareness Principle: Data entry is an iterative process. Keep a list of the changes you have to make (computations), the values you will have to change (recoding), and the problems you will have to solve (debugging), but try to use the computer to do as much computing and debugging as possible.

Data Manipulation Principle: Let the computer do as much work as possible. Instruct it to do tasks such as recoding, variable computation, dataset catenation (linking), dataset subsetting, data merging, and similar tasks that would, frankly, waste your time. Let the computer manipulate your data for you.

Original Data Principle: Always save a computer file of the original, unaltered data. In this way, if you make a mistake in manipulating data through improper recoding of variables or computing new variables, you will have the original data file to use to rectify any mistakes.

Kludge Principle: Sometimes the best way to manipulate data is not elegant and seems to waste computer resources. A kludge is sometimes justifiable; the end CAN justify the means. (In information technology, a kludge [*pronounced clue-j*] is an awkward or clumsy patching together of a series of computer commands to make the data do what you want.)

Default Principle: Know your software's default settings. Know whether these settings meet your needs. In particular, be aware of the default handling of missing values in your software. Not being aware of such settings can produce study results that are inaccurate. For example, the SPSS computer program has a default option that prints only the results of data analyses in the output unless the user specifies that a log of what commands were used to compute the analyses is set prior to running the analysis. The same thing applies to recoding variables and/or computing scores from several item variables. Not setting this option can result in not knowing what, if any, mistakes were made in undertaking data analyses.

Complex Data Structure Principle: If your software can accommodate complex data structures (eg, hierarchical relational databases), then you might benefit from using that software feature. Alternatively, you might prefer a kludge (eg, copying the same information to each record). The choice is yours as to how best to achieve your data entry purpose.

Software's Data Relations Principle: Know whether your software can perform the following four relations and, if so, what commands are necessary for it to do so: subsetting (Can subgroups be formed from the

larger dataset?), catenation (Can two subgroups of data be joined to form one larger dataset?), merging (Can two separate datasets of cases and/or variables be joined together to form one larger dataset?), and relational database construction (Can two separate datasets be joined together in a hierarchical fashion?).

Software's Sorting Principle: Know how to perform a sort in your software and whether your software requires a sort before a by-group analysis or before merging. For example, prior to merging two datasets containing the same cases but different variables, the variable on which you will match cases (normally, the subject identification code) may need to be sorted in an ascending (smallest to highest numbers) or descending (largest to smallest numbers) order in both data files. Thus, the larger dataset variables will be matched to the correct participant data record.

Impossibility/Implausibility Principle: Use the computer to check for impossible and implausible data. This should be done routinely by computing frequencies and measures of central tendency (eg, descriptive statistics) on all study variables and examining them for mistakes and/or bugs. If found, correct them immediately and then save the dataset.

Burstein's Data Sensibility Principle: Run your data all the way through to the final computer analysis and ask yourself whether the results make sense. Be prepared to decide that they do not, and hence, be prepared to treat the analysis not as final, but as another debugging step. You need to know your data as completely as possible so as not to be surprised by unexpected findings.

Extant Error Principle: Data bugs exist. Even if you have corrected one or more mistakes in your dataset, it is still possible that you missed something. Thus, always maintain an attitude of healthy skepticism when examining your data analysis results. And don't be surprised if you find another bug that needs fixing.

Manual Check Principle: Nothing can replace another pair of eyes to check over a dataset. Either check your data entry, input, and manipulation yourself, or get somebody else to do it. Determine the criticality of your dataset before expending human resources to check it manually. Highly critical datasets require manual checking regardless, possibly a priori, certainly a posteriori. Ideally, all datasets require manual checking. You should debug data by computer (Impossibility/Implausibility Principle) before you check it manually so that manual checking is easier.

Error Typology Principle: Debugging includes detection and correction of errors. To ease correction, try to classify each error as you uncover it. The two most common types of error are entry errors and logic errors. An entry error, a mistake in typing one or more responses correctly, is quite common and, once detected, is simple to fix. Locate the respondent's identification number, then retrieve the original data record and correct the mistake. In contrast, a logic error may be less detectable and more serious. For example, suppose the scoring instructions for computing a total score for a

health status measure (10 items measured on a 5-point Likert scale) directs you to reverse-score (ie, make 1 = 5, 2 = 4, 3 = 3, 4 = 2, 5 = 1) one or more items prior to adding them together. You neglect to reverse-score these items and just add the items together, forming a health-status score. The resulting sum is incorrect because some of the items are not correctly recoded prior to being tallied.

In summary, Davidson's principles summarize the key dilemmas faced by researchers and the decisions they may have to make as they work with data destined for statistical analyses. The thing to keep in mind is to avoid the worst-case scenario: that of finding yourself with data that are inappropriate for the intended statistical analyses that will achieve study aims. Thus, it is extremely important from the earliest possible moment to foresee the form of statistical analysis that is intended to achieve study aims.

UNIVARIATE ANALYSES

As our society has grown more dependent on statistics and other numeric information, the need to present data in an appropriate way has become extremely important. As the first step, researchers should examine each variable separately, whether those variables are demographic, prognostic, group membership, or outcomes. Univariate analyses are helpful in cleaning and checking the quality of data that have been entered into a statistical computer program. The data values for each variable in the dataset must be examined visually or via computer. If the data indicate that a pregnant woman is 86 years old, then an error most likely occurred in data entry for that variable in that individual case. The researcher can then locate the individual case identification number in the dataset, check the original test information, correct the data entry error, and save the new information in the data file.

Univariate analyses are also helpful in examining the variability of data, describing the sample, and checking statistical assumptions before performing more complex analyses. In some cases, data analysis may end here if the research questions can be answered solely by univariate analyses.

PRESENTING DATA

A set of data can be presented in a table or in a chart. Tables offer two main advantages: They condense data into a form that can make them easier to understand (Morgan, Reichert, & Harrison, 2002); and they show many details in summary fashion. But tables have one major disadvantage: Because the reader sees only numbers, the table may not be readily understood without comparing it with other tables. In contrast, charts speak directly to the reader; despite their lack of exact details, charts are very effective in giving the reader a picture of differences and patterns in a set of data (Wallgren et al., 1996). They are often a very effective way to describe, explore, and summarize a set of numbers (Morgan et al., 2002; Tufte, 1983).

Tables

When data are organized into values or categories and then described with titles and captions, the result is called a ***statistical table.*** A researcher begins to construct a table by tabulating data into a frequency distribution—that is, by counting how often each value or category occurs in a variable or set of variables.

For nominal and ordinal variables, the categories should be listed (in some natural order if possible) and then the frequencies indicated for each category. Table 1-1 is an example of such a table, as produced by a computer, for the nominal variable of marital status. It is helpful to state the percentage in each category. Then the reader can quickly see that the majority of subjects in this sample were widowed (53.5%). The Percent column displays the percentage in each category, calculated on the total number of cases, including those with missing data on this variable. The next column, Valid Percent, provides the percentage of cases in each category based on the number of cases with no missing data. The Cum. Percent column refers to cumulative percentages, again with missing values excluded. By summing the valid percents (14.1%, 53.5%, 14.5%, and 12.4%), the cumulative percentage of 94.6% was formed, indicating that all but 5.4% of persons in this sample were either married or formerly married (ie, widowed, divorced, separated combined).

For interval or ratio variables, an ordered array of values (Table 1-2) is usually the first step in constructing a table. This frequency distribution table might be termed a ***working table.*** If the difference between the maximum and the minimum

TABLE 1-1 *Example of Frequency Distribution Produced by SPSS 12.0 for Windows: Marital Status in a Sample of 246 Older Black Women**

Program

FREQUENCIES VARIABLES = MARITAL

Output

MARITAL Marital Status

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum. Percent*
Married	1	34	13.8	14.1	14.1
Widowed	2	129	52.4	53.5	67.6
Divorced	3	35	14.2	14.5	82.2
Separated	4	30	12.2	12.4	94.6
Never married	5	13	5.3	5.4	100.0
Total				100.0	
System missing		5	2.0		
Total		246	100.0		

*Data from Wood, R. Y. (1997). The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.

value exceeds 15, the researcher may want to group the data into classes or categories before forming the final table (this also may be true for some ordinal variables). In Table 1-2, the ages of older women go from 60 to 105, with a range of 45 (ie, 105 − 60 = 45). Therefore, grouping the values in a meaningful way will make the data more comprehensible.

As the next step, the computer printout for Table 1-3 shows a frequency distribution for the same data, with the values grouped into 3 classes, each containing those women whose age fit within the category of young-old, old-old, and oldest-old, a common method of grouping older persons. The young-old group contained 173 older women between the ages of 60 and 74.9 years; the old-old group had 50 older women between the ages of 75 and 84.9 years; and the oldest-old group had

TABLE 1-2 *Example of Frequency Distribution (Condensed) Produced by SPSS 12.0 for Windows: Age of Older Black Women in Sample**

Program

FREQUENCIES VARIABLES = AGE

Output

AGE Older Women's Age

Value Label	*Frequency*	*Value Label*	*Frequency*
60	14	78	7
61	13	79	5
62	15	80	2
63	13	81	6
64	11	82	6
65	19	83	1
66	11	84	4
67	13	85	3
68	7	86	3
69	9	87	4
70	5	88	6
71	22	89	1
72	4	90	1
73	8	92	1
74	9	94	1
75	5	98	1
76	4	100	1
77	10	105	1

*Data from Wood, R. Y. (1997). The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.

TABLE 1-3 *Example of Frequency Distribution (Condensed) Produced by SPSS 12.0 for Windows: Age Groups of Older Black Women in Sample**

Program

RECODE AGE (60 through 74.9 = 1) (75 through 84.9 = 2) (85 through 105 = 3) INTO RECAGE.
EXECUTE.

Output

RECAGE Older Women's Recoded Age Group

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum. Percent*
Young-Old Women (60–74.9 years)	1	173	70.3	70.3	70.3
Old-Old Women (75–84.9 years)	2	50	20.3	20.3	90.7
Oldest-Old Women (85–105 years)	3	23	9.3	9.3	100.0

*Data from Wood, R. Y. (1997). The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.

23 women 85 or more years. Again, it is most helpful to know the percentage falling into each group. The groupings fall in the expected direction with the youngest young-old group being far more numerous than the oldest-old group. By looking at the "Cum. Percent" column, the reader can quickly see that almost 91% of the sample were less than 85 years old.

By using the Recode command in a computer program, the researcher can easily form this new nominal level group (RECAGE) variable from the original interval level (AGE) variable. When creating such a variable, it is wise to create the new variable by using the *Recode into a Different Variable* command rather than to permanently change the original variable by recoding into the same variable. It is best to preserve the variable in its original form.

Computer programs can also group variable values automatically; however, some programs have defaults for the interval width and the number of classes produced, resulting in an inconveniently constructed table. Most statistical programs let the researcher control the choice of interval and the number of classes. Using a multiple of five for the interval width is helpful because it is easier to think about numbers that are divisible by five.

Authorities differ on their recommendations for the number of classes. Glass and Hopkins (1996) suggest there should be at least ten times as many observations as classes until there are between 20 and 30 intervals. Freedman et al. (1991) suggest 10 to 15 classes; Ott and Mendenhall (1990) suggest 5 to 20; and Freund (1988) suggests 6 to 15 classes. Thus, it is up to the researcher to determine the number of intervals in a frequency distribution of a variable. Usually, the clustering that best

depicts the important features of the distribution of scores for the intended audience should be the major consideration. Too few or too many classes will obscure important features of a frequency distribution. Some detail is lost by grouping the values, but information is gained about clustering and the shape of the distribution.

The final presentation of the data from Table 1-3 depends on the format requirements of each journal or of the dissertation or thesis. If a table is included, it should be mentioned in the text of the research report. The discussion of a table should reinforce the major points for which the table was developed (Burns & Grove, 2001). Researchers should comment on the important patterns in the table as well as the major exceptions (Chatfield, 1988), but should not rehash every fact in the table.

Suggestions for the Construction of Tables for Research Reports

The specific content of a table will vary depending on the statistical analysis you are summarizing and/or the hypothesis you are testing. It is wise to use a table only to highlight major facts. Most of the tables examined by researchers while analyzing their data do not need to be published in a journal. If a finding can be described well in words, then a table is unnecessary. Too many tables can overwhelm the rest of a research report (Burns & Grove, 2001).

The table should be as self-explanatory as possible. The patterns and exceptions in a table should be obvious at a glance once the reader has been told what they are (Ehrenberg, 1977). With this goal in mind, the title should state the variable, when and where the data were collected (if pertinent), and the size of the sample. Headings within the table should be brief and clear. Find out the required format for tables in the research report. If the report is being submitted to a particular journal, examine tables in recent past issues. Follow the advice about table format for publication in a manual of style, such as the *Publication Manual of the American Psychological Association* (APA, 2001). Rudestam and Newton (1992) also suggest that tables should be numbered as whole numbers, such as Table 1, Table 2, and the like. They recommend not using a chapter number-table number form like Table 2.1 or Table 2.2. However, in a book chapter or a dissertation, table titles need to conform to the publisher's or university's requirements. If the data being presented in the table are not original, notes, including the data source, should be included.

Morgan and colleagues (2002) offer several principles that should guide table construction:

1. Don't try to do too much in a table. Model tables after published exemplars of similar research to find the right balance for how much a table should contain.
2. Use white space effectively so as to make the layout of the table pleasing to the eye and aid in comprehension and clarity.
3. Make sure tables and text refer to each other; but not everything displayed in a table needs to be mentioned in the text.
4. Use some aspect of the data to order and group rows and columns. This could be size (largest to smallest), chronology (first to last), or to show similarity or invite comparison.

5. If appropriate, frame the table with summary statistics in rows and columns to provide a standard of comparison. Remember when making a table that values are compared down columns more easily than across rows.
6. It is useful to round numbers in a table to one or two decimal places because they are more easily understood when the number of digits is reduced.
7. When creating tables for publication in a manuscript, they should be double-spaced unless contraindicated by the journal.

Charts

Although there are many different kinds of charts, most are based on several basic types that are built with lines, areas, and text. These include bar charts, histograms, pie charts, scatter plots, line charts, flow charts, and box plots. Charts can quickly reveal facts about data that might be gleaned from a table only after careful study. They are often the most effective way to describe, explore, and summarize a set of numbers (Tufte, 1983). Charts, the visual representations of frequency distributions, provide a global, bird's-eye view of the data and help the reader gain insight.

Choosing which type of chart to use in a given situation depends on what we wish to convey. When drawing a chart, Wallgren et al. (1996) suggest three things should be considered: data structure, variable type, and measurement characteristics. The researcher should ask these questions:

- Do the data represent one point in time, indicating *cross-sectional data*, or do they represent several points in time, called *time series data*?
- What type of variable do we wish to illustrate?
  - Is the variable *qualitative*, consisting of words, or *quantitative*, consisting of numbers?
  - If quantitative, is the variable *discrete*, which can take on only certain values, or *continuous*, which can take all the numbers in a range?
- What level of measurement is the variable of interest?

Answering these questions will help the researcher choose the type of chart that best illustrates a variable's characteristics.

BAR CHART

A bar chart, the simplest form of chart, is used for nominal or ordinal data. When constructing such charts, the category labels usually are listed horizontally in some systematic order, and then vertical bars are drawn to represent the frequency or percentage in each category. A space separates each bar to emphasize the nominal or ordinal nature of the variable. The spacing and the width of the bars are at the researcher's discretion, but once chosen, all the spacing and widths should be equal. Figure 1-1 is an example of a bar chart for ordinal data. If the category labels are lengthy, it may be more convenient to list the categories vertically and draw the bars horizontally, as in Figure 1-2.

Bar charts also make it easier to compare univariate distributions. Two or more univariate distributions can be compared by means of a cluster bar chart (Fig. 1-3). Current computer graphics, statistics, and spreadsheet programs offer

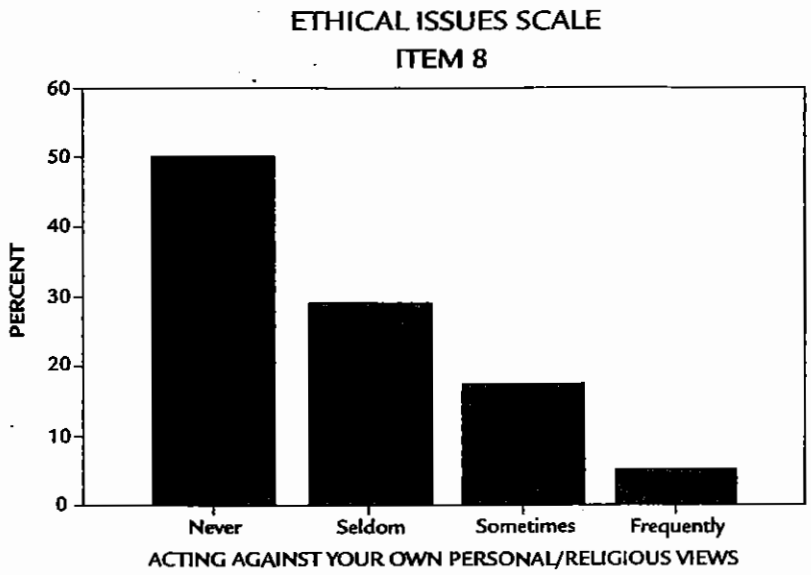


FIGURE 1-1. Degree of frequency that nurses experience the ethical issue of acting against their personal or religious beliefs. (Data from Fry, S., & Duffy, M. [2000]. *Ethics and Human Rights in Nursing Practice: A Study of New England Registered Nurses.* Chestnut Hill, MA: Nursing Ethics Network & The Center for Nursing Research, Boston College.)

many tempting patterns for filling in the bars. The legend explaining Figure 1-3 is outside the chart to avoid clutter (Cleveland, 1985). Wallgren et al. (1996) recommend filling bars with either shading of various depths or simple dot or line patterns, avoiding complex patterns, slanting lines in different directions, or a combination of horizontal and vertical lines in the same chart.

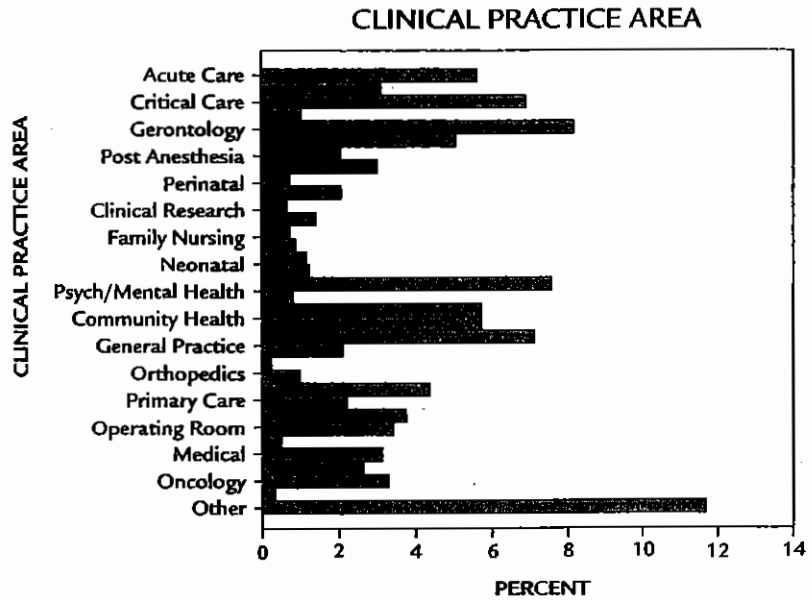


FIGURE 1-2. Major clinical practice area of registered nurses working in the six New England states in 1997 (N = 2,090). (Data from Fry, S., & Duffy, M. [2000]. *Ethics and Human Rights in Nursing Practice: A Study of New England Registered Nurses.* Chestnut Hill, MA: Nursing Ethics Network & The Center for Nursing Research, Boston College.)

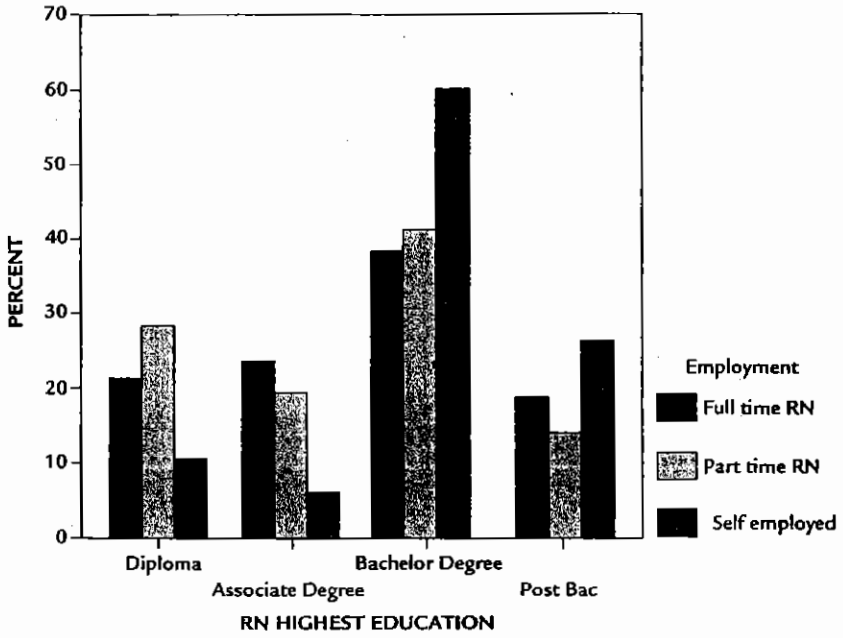


FIGURE 1-3. Employment status by RN highest level of education. (Data from Fry, S., & Duffy, M. [2000]. *Ethics and Human Rights in Nursing Practice: A Study of New England Registered Nurses.* Chestnut Hill, MA: Nursing Ethics Network & The Center for Nursing Research, Boston College.)

PIE CHART

The pie chart, an alternative to the bar chart, is simply a circle that has been partitioned into percentage distributions of qualitative variables. Simple to construct, the pie chart has a total area of 100%, with 1% equivalent to 3.6° of the circle. Figure 1-4 is an example of a pie chart displaying RNs' need for ethics and human rights education.

When constructing a pie chart, Wallgren et al. (1996) recommend the following:

- Use the pie chart to provide overviews: readers find it difficult to get precise measurements from a circle.
- Place the different sectors in the same order as would be found in the bar chart, beginning either in an ascending or a descending order. Retain the order between the variables.
- Use the percentages corresponding to each category rather than the absolute frequency of each category.
- Read the pie chart by beginning at the 12 o'clock position and proceeding clockwise.
- Use no more than six sectors in a given pie chart; clarity is lost with more than six sectors.

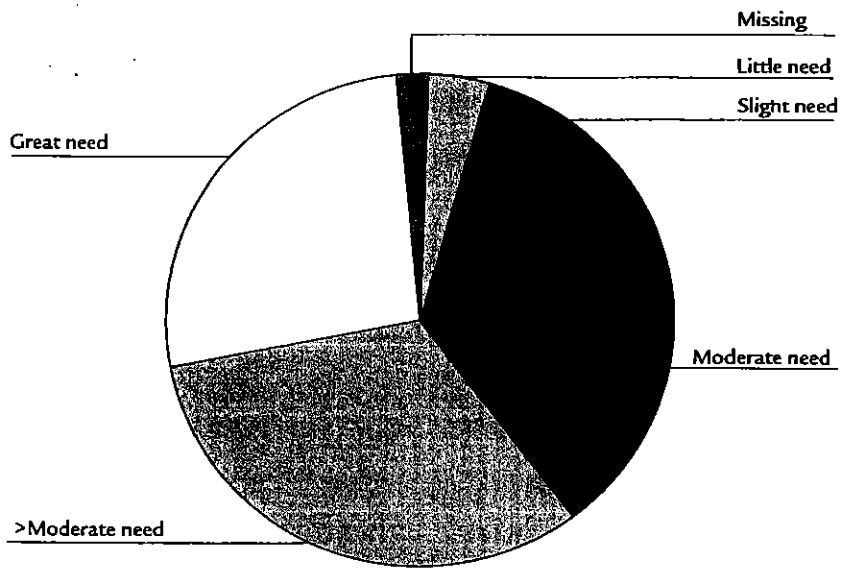


FIGURE 1-4. New England RNs' reported need for ethics and human rights education. (Data from Fry, S., & Duffy, M. [2000]. *Ethics and Human Rights in Nursing Practice: A Study of New England Registered Nurses.* Chestnut Hill, MA: Nursing Ethics Network & The Center for Nursing Research, Boston College.)

- Use a low-key shading pattern that does not detract from the meaning of the pie chart.
- If using more than one pie chart, give the number on which the percentages are based for each circle.
- Make sure the sum of the pie chart sectors equals 100%.

HISTOGRAM

Histograms, appropriate for interval, ratio, and sometimes ordinal variables, are similar to bar charts, except the bars are placed side by side. The bar length represents the number of cases (frequency) falling within each interval. Histograms are often used to represent percentages instead of, or in addition to, frequencies because percentages are more meaningful than simple number counts. Therefore, each histogram has a total area of 100%.

The first decision is to select the number of bars. With too few bars, the data will be clumped together; with too many, the data will be overly detailed. Figure 1-5 shows how the choice for the number of bars affects the appearance of a histogram. The top chart presents a jagged appearance; the bottom chart clumps the data into only four bars and makes the data seem skewed. The middle chart, with 10 bars, presents a smoother appearance.

Computer programs are handy for a preliminary chart of a variable, but the researcher should be aware of built-in defaults and should think about the adjustments

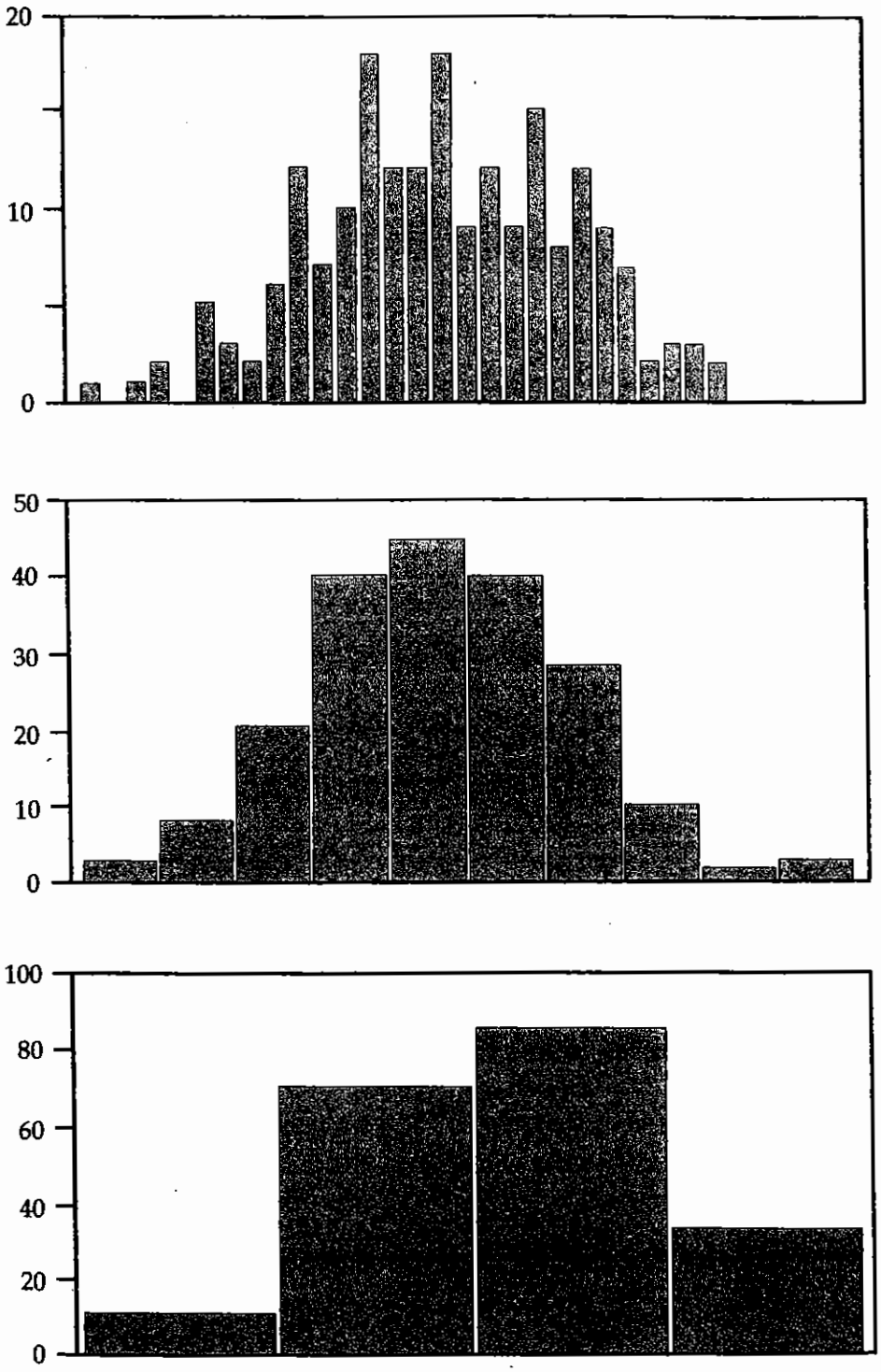


FIGURE 1-5. Illustration of the importance of the number of bars when designing a

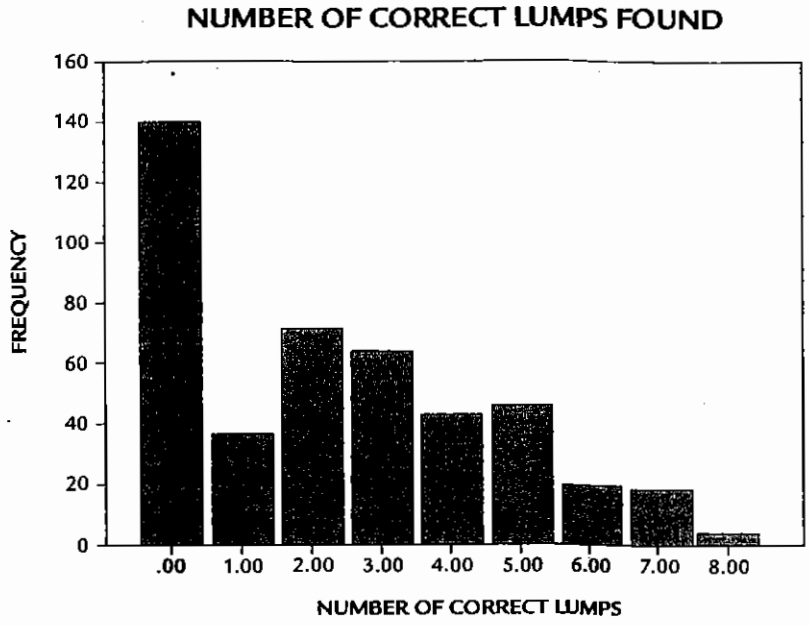


FIGURE 1-6. Number of correct breast lumps identified by older Black women ($N = 246$). (Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)

that are necessary. The advice given in the previous section for constructing frequency distributions for interval or ratio variables is helpful here. For example, if the difference between the maximum and minimum values exceeds 15, the researcher should consider grouping the data. The interval and the starting point should be divisible by five. Most histograms have 5 to 20 bars.

For interval or ratio variables that are discrete, the numerals representing the values should be centered below each bar to emphasize the discrete nature of the variable. Figure 1-6 illustrates a histogram for the discrete variable of number of correct lumps identified by older Black women. For continuous variables, the numerals representing the values should be placed at the sides of the bars to emphasize the continuous nature of the distribution.[1] Figure 1-7 depicts a histogram for the continuous variable of number of cigarettes smoked per day for a national sample of community-dwelling adults. Tick marks are placed outside the data region to avoid clutter (Cleveland, 1985).

Once the number of bars has been determined, the next decision concerns the height of the vertical axis. If the chart is horizontal, Tufte (1983) recommends a height of approximately half the width. Other authorities, such as Schmid (1983), recommend a height approximately two thirds to three fourths the width. The reason for these recommendations is the different effect that can be produced by altering the height of a chart. Figure 1-8 shows the different impressions that can be created for

[1]The grouping interval of "25 to 29" has a lower limit of 25 and an upper limit of 29. These are called the written limits. The real or mathematical limits are understood to extend half a unit above and below the written class limits. For convenience, researchers almost always use the written class limits in tables and charts.

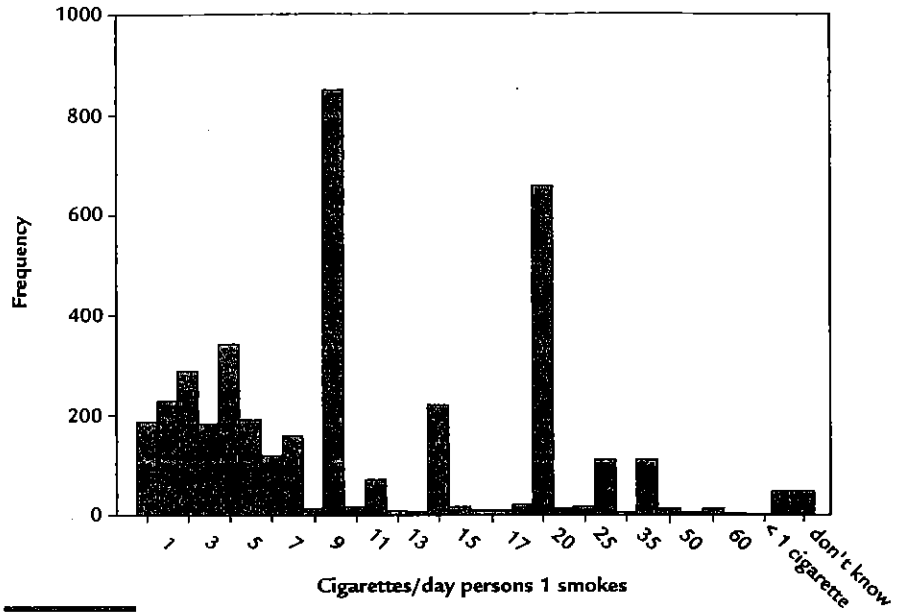


FIGURE 1-7. Number of cigarettes smoked per day in a national sample of 16,197 community-dwelling adults. (Data from US Department of Health and Human Services [DHHS]. [1996]. *Third National Health and Nutrition Examination Survey, 1988–1994*, NHANES III Laboratory Data File [CD-ROM]. US Department of Health and Human Services. National Center for Health Statistics. Public Use Data File Documentation Number 76200. Hyattsville, MD: Centers for Disease Control.)

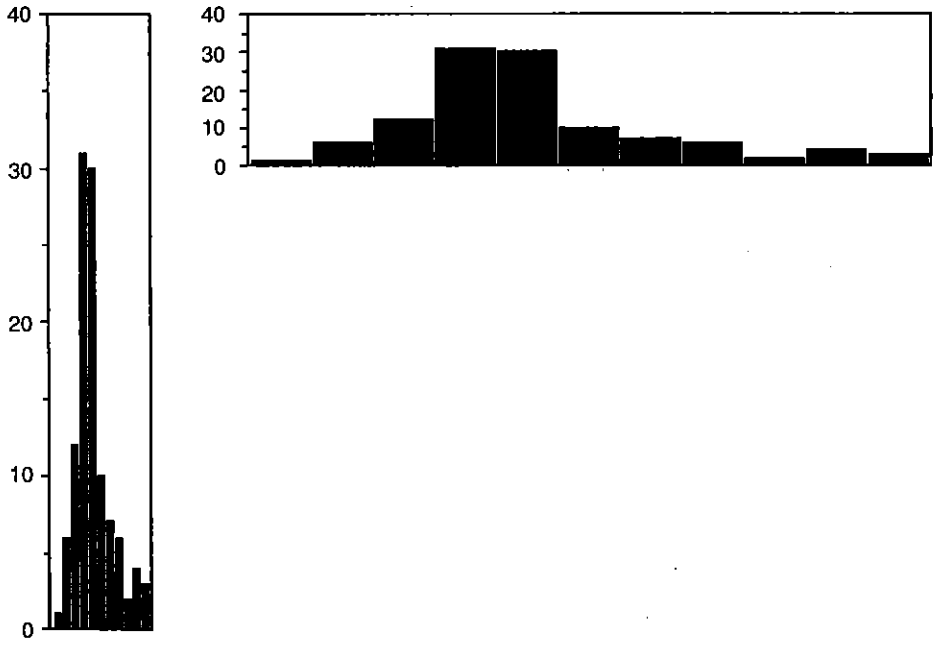


FIGURE 1-8. Illustration of the importance of the graph's height when designing a

the same data by a tall, narrow chart and by a flat, wide chart. The tall, narrow chart seems to emphasize the clustering of the data in the middle, whereas the flat, wide chart appears to emphasize the scatter of the data to the right.

POLYGON

The polygon, a chart for interval or ratio variables, is equivalent to the histogram but appears smoother. For any set of data, the histogram and the polygon will have equivalent total areas of 100%. The polygon is constructed by joining the midpoints of the top of each bar of the histogram and then closing the polygon at both ends by extending lines to imaginary midpoints at the left and right of the histogram. Figure 1-9 illustrates a polygon superimposed on a histogram. In the process of construction, triangles of area are removed from the histogram, but congruent triangles are added to the polygon. Two such congruent triangles are shaded in Figure 1-9 to show why the areas of the two types of chart are equivalent.

Polygons are especially appropriate for comparing two univariate distributions by superimposing them (Fig. 1-10). The percentages were used on the vertical scale because the sizes of the two samples differed.

WHAT TO LOOK FOR IN A HISTOGRAM OR POLYGON

A chart can help us see quickly the shape of a distribution. Frequency distributions have many possible shapes. Often they have a bell-shaped appearance, as in the computer printout in Figure 1-11. In this case, older women rated themselves on the

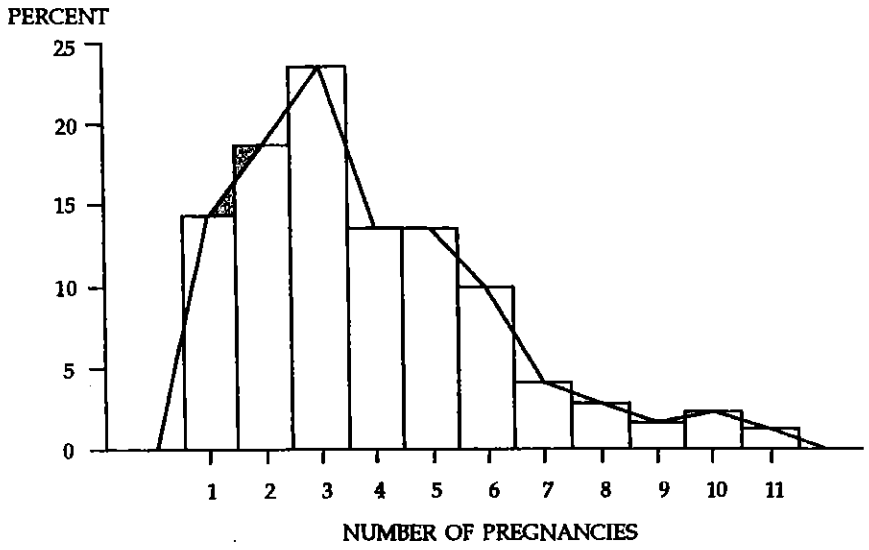


FIGURE 1-9. Polygon superimposed on histogram. The two shaded triangles are congruent. (Data collected with a grant funded by the National Institute of Nursing Research, NR-02867. P.I., Brooten, D. University of Pennsylvania School of Nursing, *Nurse Home Care for High Risk Pregnant Women: Outcome and Cost.*)

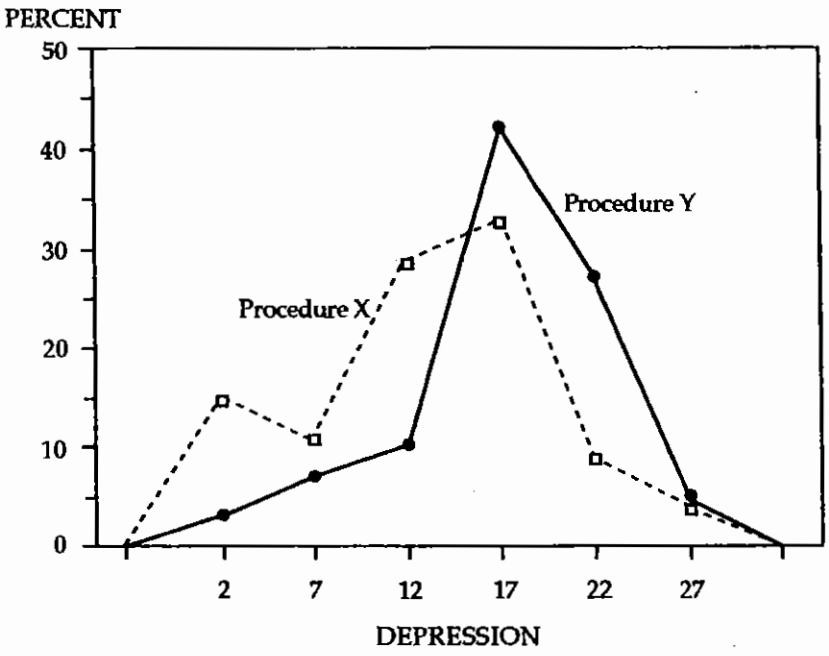


FIGURE 1-10. Comparison of depression scores for patients having surgical procedure X ($N = 104$) and patients having surgical procedure Y ($N = 61$).

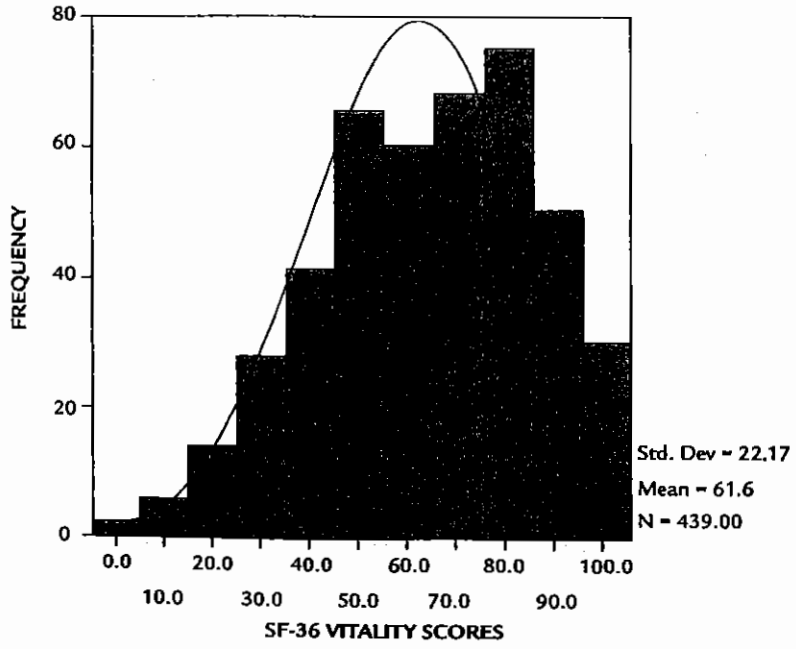


FIGURE 1-11. Example of a histogram produced by SPSS 12.0 for Windows: SF-36 Transformed Vitality Scores from a sample of 439 older women. (Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)

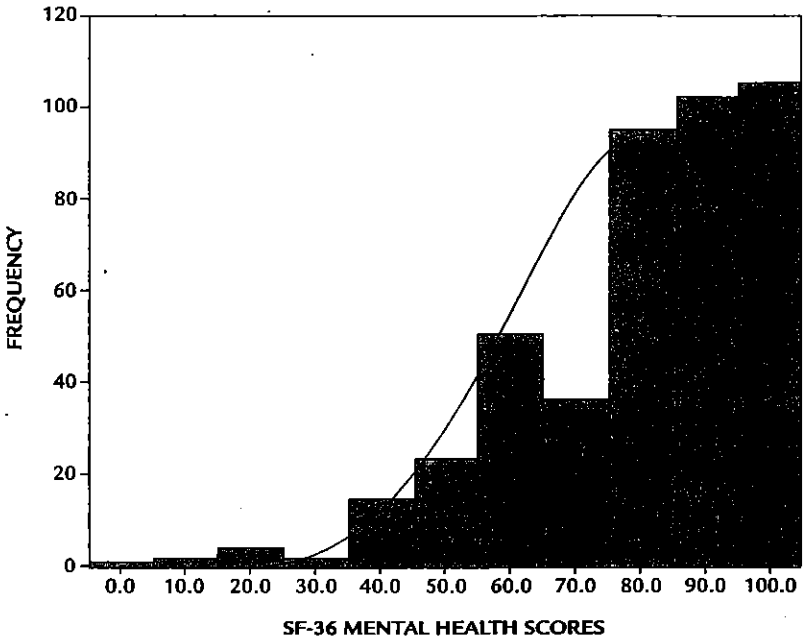


FIGURE 1-12. Relative frequency distribution of SF-36 transformed mental health scale scores from a sample of 439 older women. (Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)

Vitality subscale of the Short Form-36 Health Survey (SF-36), consisting of four items, with each item rated for evidence of vitality, energy, or fatigue on a 6-point scale, transformed so that scores ranged from 1 to 100 for comparison purposes. Technically, such a scale is ordinal, because there is no accepted physical unit of vitality, and the zero point is arbitrary. An ordinal scale with such a large range, however, is usually treated as interval in the research literature. In Figure 1-11, the frequency count is given at the left. The programmer chose an interval width of 10 and a starting point of 0. Thus, the first class is 0 to 10. In addition, the programmer instructed the computer program to plot the bell-shaped (normal) curve atop the histogram with a Line. The reader can then visually compare the distribution of transformed SF-36 vitality scores with the theoretical bell-shaped, or normal, curve.

Distributions also may be skewed, as in Figure 1-12. Occasionally, data may clump at several places, as in Figure 1-13.

Charts also can be helpful in spotting where the data cluster, how the data are scattered around the clustering points, whether there are far-out observations that may be outliers, and whether there are gaps in the data. These are the kinds of features that researchers need to know, and they become immediately evident with simple graphic representation (Cohen, 1990). With the assistance of a computer, researchers have no excuse for failing to know their data (Jacobsen, 1981). As with tables, all charts included in a research report should be cited in the text, and the important features of the distribution should be discussed.

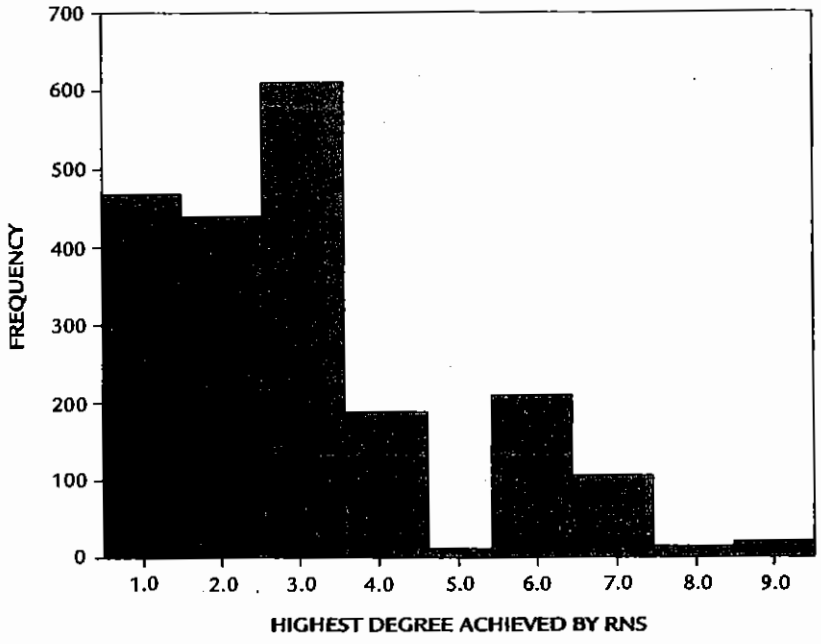


FIGURE 1-13. Relative frequency distribution of the highest degree achieved by registered nurses. (Data from Fry, S., & Duffy, M. [2000]. *Ethics and Human Rights in Nursing Practice: A Study of New England Registered Nurses.* Chestnut Hill, MA: Nursing Ethics Network & The Center for Nursing Research, Boston College.)

GENERAL SUGGESTIONS FOR CONSTRUCTING CHARTS

The purpose of a chart is to promote understanding without distorting the facts; therefore, the chart should make the desired point *honestly*. Because gross misuses of charts are not generally found in respected research journals, some researchers believe that they need not pay attention to the construction of their charts because editors and reviewers will tell them how to fix the charts. This is not true: Read some of the references cited on developing charts, and follow their advice to avoid having your research report rejected.

Wallgren et al. (1996) advise researchers to ask themselves the following questions after completing a chart:

- **Is the chart easy to read?** Simplicity is the hallmark of a good chart. What you want to display in a chart should be quickly and clearly evident. Keep in mind the target audience for whom you are constructing the chart. Keep grid lines and tick marks to a minimum. Avoid odd lettering and ornate patterns (Schmid, 1983).
- **Is the chart in the right place?** Locate the chart close to the place in the text where the topic is discussed. Make sure the chart is well positioned on the page.
- **Does the chart benefit from being in color (if color is used)?** Color should have a purpose; it should not be used solely for decorative reasons.
- **Have you tried the chart out on anybody?** Try the chart out on someone whom you consider to correspond to the target group before you make the

final diagram. Ask that person questions about the chart to gain information on how the person perceives the chart.

Wallgren et al. offer the caveat that "a poor chart is worse than no chart at all" (1996, p. 89).

SUMMARY

The first steps toward understanding data are univariate analyses. The researcher should study each variable separately by means of tables and charts. The type of table or chart varies according to the type of measurement scale. For nominal variables, the table should be a simple listing of categories with corresponding frequencies and percentages, and the bar chart is appropriate for graphic display. For interval or ratio variables, it may be necessary first to group the data into appropriate numeric intervals before constructing a frequency table, histogram, or polygon. For ordinal variables, the researcher must decide whether the values should be treated as nominal data or as continuous (interval or ratio) data. Once this decision has been made, the researcher can then apply the rules for either nominal or interval and ratio variables. The best tables and charts are self-explanatory and present data in a clear and straightforward manner.

Application Exercises and Results

General Introduction

Appendix G contains a survey instrument that was developed at Boston College's William F. Connell School of Nursing for doctoral students to gather data for use in a statistics class. Each student is responsible for getting 10 people to fill out the questionnaire. Students are asked to have variety in terms of the respondents' gender, age, and so forth. We also ask them to try to minimize missing data by checking questionnaires for completeness.

The students then enter the data from their 10 subjects into a data file, called a *dataset*. They examine a printout of file information, run frequencies, and examine the output carefully to be sure they have entered their data correctly. The students make corrections as necessary, and then their data files are merged and we provide them with a large dataset they can use for all their homework assignments.

The CD at the back of this book contains data collected by these students on 701 respondents. If the same survey is used for several years, a fairly substantial dataset can be developed. The reader may use our survey form, collect data, and add it to the dataset we have provided.

When the students collect and enter data and then clean the datasets, they achieve a much better understanding of data and how to manage it. Although our students use their dataset for all homework exercises, other large datasets that we get from researchers in the school are used for the midterm and final exams. Each student is provided with data from a randomly drawn subset of one of these large datasets. Thus, each student has a slightly different sample of respondents but the same variables so they can answer the same questions on the exam. Their answers will differ because they have different cases in their sample dataset.

The major dataset for the exercises throughout the book is named MUNRO04.SAV and was created in SPSS for Windows, version 12.0. For the purpose of this book, we have posed

specific research questions for students to answer. In our courses, however, we often just ask the students to state a research question or hypothesis that can be answered using their dataset and the statistical technique being studied that week. They then run the analysis and write up the results. We have found that students need more guidance in how to write the results in a manner that would be acceptable for a research journal than in how to run the analysis. We have not provided step-by-step guidance in the use of statistical software for several reasons: so many statistical packages are available, students are more computer literate today, and statistics software is very user-friendly. When we first began teaching doctoral students how to use SPSS for Windows several years ago, it took a full-day workshop to accomplish this task. Now, it takes students about 1 hour to use later versions of SPSS for Windows.

Exercises

1. Access the dataset called MUNRO04.SAV, which contains data collected using the survey form contained in Appendix G. Either bring it into SPSS or convert it into a file for SAS or whatever software you are using. Print the dictionary, which contains a list of the variables, formats, and labels. In most versions of SPSS for Windows, this is done by clicking on File Info, then on Display Data File Information, then on Working File. Once the file is in the output screen, it can be printed from the File menu.
2. Compare the file information with the survey form in Appendix G. Note that the variable names have been selected to reflect each variable, making it easy to recognize them when working with the file. Variable labels and value labels have been added to enhance the output. Look for any discrepancies between the survey form and the file information.
3. Produce charts/graphs. Many options are available for producing charts in statistical software programs. They may be produced within specific techniques and in separate graphics sections. We will confine ourselves to requesting graphics that are available with the specific techniques. The following can be requested as part of the output from frequencies in most software programs. Within the frequencies program, request a bar graph for GENDER, a histogram for SATCURWT, and a histogram with a polygon (normal curve) for SATCURWT.

Results

1. Exercise Figure 1-1 contains a portion of the dictionary.
2. If you look carefully, you should note the following:
   a. Compared to what is printed in the survey form in Appendix G, the value labels for the following items from the Inventory of Personal Attitudes (IPPA) have been reversed: 1, 2, 4, 6, 8, 13, 15, 20, 22, 24, 27, and 29. This had to be done to prepare these items for scoring the scale. For example, look at item 1. On the questionnaire, we see that a very high level of energy is scored 1 and a very low level is scored 7. Because the inventory measures positive attitudes, the originators (Kass et al., 1991) reversed this item before adding it to the scale. We have already done the reverse-scoring for you. We recoded all of these items so that 1 = 7, 2 = 6, 3 = 5, 4 = 4, 5 = 3, 6 = 2, and 7 = 1. Thus, with item 1, if someone checked a 1, it would now be scored a 7. We also reversed the value labels to reflect the new scoring.

      If you add your own data to our dataset, recode these items and be sure the value labels are correct **before** adding them to our dataset. If you add your data to the dataset first, then reverse-score the IPPA items, you will also change the previously reverse-scored items back to their original, unreversed-score values.

   b. Three "extra" variables are listed in the dictionary. These new variables follow the 30 IPPA items.

```
                    List of variables on the working file
Name                                                                Position

CODE        subject's identification number                             1
            Print Format: F3
            Write Format: F3

GENDER      gender                                                      2
            Print Format: F1
            Write Format: F1

            Value    Label
                0    male
                1    female

AGE         subject's age                                               3
            Print Format: F3
            Write Format: F3

MARITAL     marital status                                              4
            Print Format: F1
            Write Format: F1

            Value    Label
                1    Never Married
                2    Married
                3    Living with Significant Other
                4    Separated
                5    Widowed
                6    Divorced

DEPRESS     depressed state of mind                                     9
            Print Format: F1
            Write Format: F1

            Value    Label
                1    Rarely
                2    Sometimes
                3    Often
                4    Routinely

IPA1        energy level                                                19
            Print Format: F8
            Write Format: F8

            Value    Label
                1    very low
                7    very high

IPA2        reaction to pressure                                        20

            Print Format: F1
            Write Format: F1

            Value    Label
                1    I get tense
                7    I remain calm
```

EXERCISE FIGURE 1-1. A portion of the SPSS dictionary.

```
FREQUENCIES
  VARIABLES=marital
  /BARCHART     FREQ
  /ORDER=     ANALYSIS
```

Frequencies

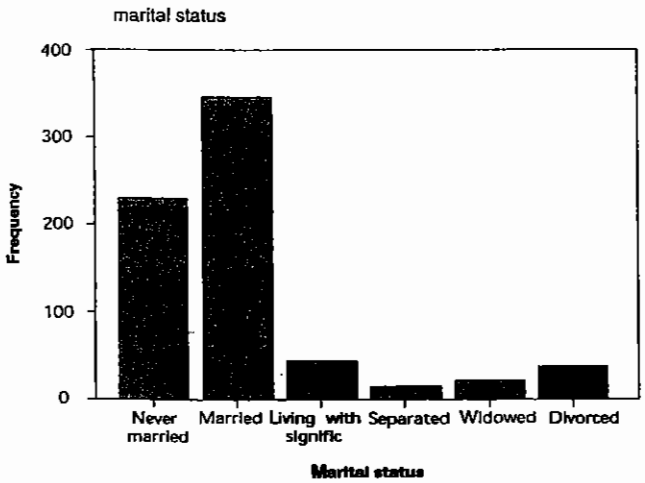
Statistics

MARITAL marital status

N	Valid	688
	Missing	13

MARITAL marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 never married	230	32.8	33.4	33.4
	2 married	346	49.4	50.3	83.7
	3 living with significant other	44	6.3	6.4	90.1
	4 separated	13	1.9	1.9	92.0
	5 widowed	20	2.9	2.9	94.9
	6 divorced	35	5.0	5.1	100.0
	Total	688	98.3	100.0	
Missing	System	13	1.7		
Total		701	100.0		



EXERCISE FIGURE 1-2. Frequencies for marital status (MARITAL) and bar graph.

```
FREQUENCIES
  VARIABLES=qolcur
  /HISTOGRAM
  /ORDER=    ANALYSIS
```

Frequencies

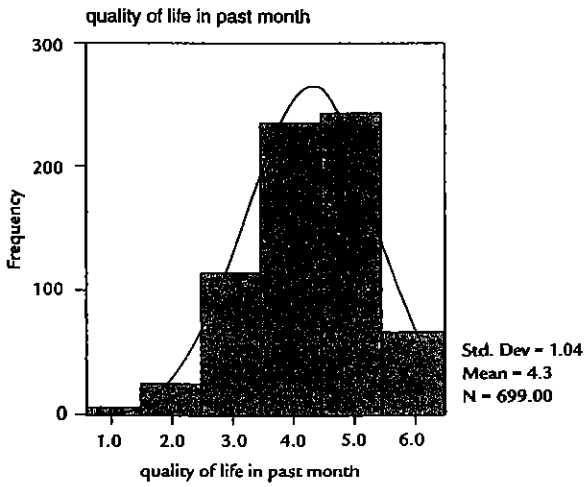
Statistics

QOLCUR quality of life in past month

N	Valid	699
	Missing	2

QOLCUR quality of life in past month

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 very dissatisfied, unhappy most of time	8	1.1	1.1	1.1
	2 generally dissatisfied, unhappy	27	3.9	3.9	5.0
	3 sometimes fairly satisfied, sometimes fairly unhappy	113	16.1	16.2	21.2
	4 generally satisfied, pleased	238	34.0	34.0	55.2
	5 very happy most of time	245	35.0	35.1	90.3
	6 extremely happy, could not be more pleased	68	9.7	9.7	100.0
	Total	699	99.7	100.0	
Missing	System	2	.3		
Total		701	100.0		



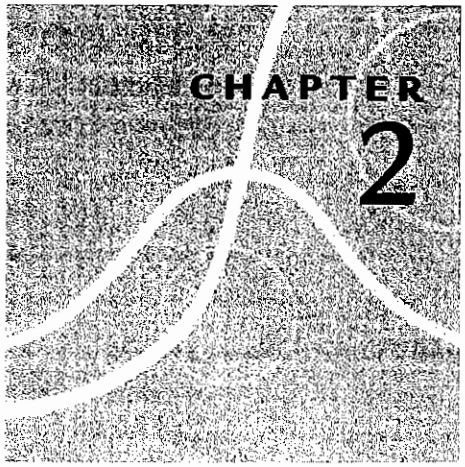
EXERCISE FIGURE 1-3. Frequencies for quality of life in the past month (QOLCUR) and two histograms.

CONFID is the sum of the following IPPA items: 2, 5, 7, 10, 14, 15, 16, 17, 18, 22, 24, 26, and 29. It is defined as self-confidence during stressful situations. Because it includes 13 items and each item is rated on a scale from 1 to 7, the potential range of scores for CONFID is 13 to 91.

LIFE is the sum of the following IPPA items: 1, 3, 4, 6, 8, 9, 11, 12, 13, 19, 20, 21, 23, 25, 27, 28, and 30. It is defined as life purpose and satisfaction and includes 17 items, with a potential range of scores of 17 to 119.

IPPATOT is the sum of all 30 items and is the total score on the IPPA. The potential range of scores is 30 to 210.

3. Exercise Figure 1-2 contains the frequencies for marital status (MARITAL) and its associated bar graph. Exercise Figure 1-3 contains the frequencies for quality of life in the past month (QOLCUR), and the histogram with the normal curve superimposed.



Univariate Descriptive Statistics

Mary E. Duffy and
Barbara S. Jacobsen

Objectives for Chapter 2

After reading this chapter, you should be able to do the following:

1. Define measures of central tendency and dispersion.
2. Select the appropriate measures to use for a particular dataset.
3. Discuss methods to identify and manage outliers.
4. Discuss methods to handle missing data.

Although charts may bring facts to life vividly, the information they present for our inspection is often inexact. Frequency distribution tables provide many details, but often a researcher will want to condense a distribution further. After the data have been organized, quantitative measures are frequently calculated to capture the essence of the four basic characteristics of a distribution: central tendency, variability, skewness, and kurtosis. These statistics may be used not only in a descriptive summary, but also in statistical inference.

Symbols and formulas for descriptive statistics vary depending on whether one is describing a sample or a population. As mentioned in Chapter 1, a population includes all members of a defined group; a sample is a subset of a population. Characteristics of populations are called *parameters*; characteristics of samples are called *statistics*. To distinguish between them, different sets of symbols are used. Usually, lowercase Greek letters are used to denote parameters, and Roman letters are used to denote statistics.

MEASURES OF CENTRAL TENDENCY

The typical value of a variable is summarized using measures of central tendency. These statistics, commonly called *averages*, describe where the values of a variable's distribution cluster. The most commonly reported measures of central tendency are the *mean*, the *median*, and the *mode*.

Mean

The *mean*, the best known and most widely used average, describes the center of a frequency distribution. The mean of a sample[1] is represented symbolically by $\overline{X}$, which is read "X bar." Many journals simply use "M" to represent the mean.

To compute the mean, add up all the values in the distribution and divide by the number of values. Expressed as a formula, the sample mean is defined as:

$$M = \Sigma X/N$$

The uppercase Greek letter sigma (Σ) means "the sum of." If the letter X represents a single quantitative value in a distribution, then ΣX means "sum up all the values."

For example, the following list of values for length of stay (in hours) in the hospital in the past year for a sample of older women has 10 entries: 8, 10, 10, 18, 24, 29, 36, 48, 60, 72. The mean is:

$$8 + 10 + 10 + 18 + 24 + 29 + 36 + 48 + 60 + 72 = 315/10$$
$$= 31.5 \text{ hours}$$

In this example, the mean is located near the middle of the 10 values. It is clear from the formula and the example that each value in the distribution contributes to the mean. Because the mean is influenced by all of the data points, it is not appropriate as a descriptive statistic for a variable when not all the data points are known. For instance, not everyone with cancer will have a recurrence of that disease; therefore, some of the values of the variable "time to recurrence" may be absent or "censored."

Any extreme values in the distribution also influence the mean. For example, in the previous distribution relating to length of stay in hours in the hospital for the group of older women, suppose the value of 72 hours was instead 224 hours. The new mean would be

$$8 + 10 + 10 + 18 + 24 + 29 + 36 + 48 + 60 + 224 = 467/10$$
$$= 46.7 \text{ hours}$$

This mean would not be located in the middle of the 10 values; only three women would have a length of hospital stay greater than the mean. Thus, the mean works best as an average for symmetrical frequency distributions that have a single peak, more commonly called a normal distribution.

[1]The mean of a population (N) is represented by the lowercase Greek letter mu (μ). The formula is the same as that for the sample mean.

TABLE 2-1 *Demonstration of Several Important Properties of the Mean*

X	$X - M$	$(X - M)^2$
4	$4 - 6 = -2$	$(-2)^2 = 4$
4	$4 - 6 = -2$	$(-2)^2 = 4$
10	$10 - 6 = +4$	$(+4)^2 = 16$
5	$5 - 6 = -1$	$(-1)^2 = 1$
7	$7 - 6 = +1$	$(+1)^2 = 1$
$\Sigma X = 30$	$\Sigma(X - M) = 0$	$\Sigma(X - M)^2 = 26$
$N = 5$		sum of squares
$M = 6$		

The mean has several other interesting properties. First, for any distribution, the sum of the deviations of the values from the mean always equals zero. This helps to explain why the mean is the center of a distribution. Table 2-1 demonstrates this property. The mean (6) is subtracted from each value to form *deviations* (X − M). These deviations from the mean sum to zero. If any value other than the mean is subtracted from each value, the sum of the deviations will not be zero.

A second property of the mean relates to the sum of the squared deviations—that is, $\Sigma(X - M)^2$. In Table 2-1, each of the deviations from the mean has been squared, and the sum of these squared deviations equals 26. This sum, called the *sum of squares* in statistics, is at a minimum; that is, it is smaller than the sum of squares around any other value. If any value other than the mean (6) is subtracted from each value and squared, the total will exceed 26. This characteristic of the mean underlies the idea of *least squares*, which is important in later chapters.

Third, because the mean has a formula, it is algebraic and can be manipulated in equations. For example, if two or more means are available from samples of different sizes, a mean of the total group can be calculated. By transposing terms in the formula for the mean, the following shows that the sum of the values is equal to the mean multiplied by the size of the sample.

$$\Sigma X = Mn$$

Therefore, a formula for a combined mean for two samples (which can be easily extended to include more than two samples), weighted according to sample size, logically follows:

$$M_{total} = M_1 n_1 + M_2 n_2 / n_1 + n_2$$

Finally, when repeatedly drawing random samples from the same population, means will vary less among themselves and less from the true population mean than other measures of central tendency. Thus, the mean is the most reliable average when making inferences from a sample to a population.

The mean is intended for interval or ratio variables when values can be added, but many times it is also sensible for ordinal variables. Computer programs, however, will compute means for nominal level variables, reporting such uninterpretable results for a sample as "the mean gender = 0.75."

Median

The median, the middle value of a set of ordered numbers, is the point or value below which 50% of the distribution falls. Thus, 50% of the sample will be below the median regardless of the shape of the distribution. The median is sometimes called the 50th percentile and symbolized as P_{50}. It may also be conceived as the bisector of the total area of the histogram or polygon. There is no algebraic formula for the median, just a procedure:

1. Arrange the values in order.
2. If the total number of values is odd, count up (or down) to the middle value. If there are several identical values clustered at the middle, the median is that value.
3. If the total number of values is even, compute the mean of the middle values.

In the previous example relating to length of stay of older women in the hospital, the 10 values, arranged in order, were 8, 10, 10, 18, 24, 29, 36, 48, 60, 72. Counting to the center of these 10 entries (an even number), the two middle values are 24 and 29. Thus, the median is $(24 + 29)/2 = 26.5$. Note that the mean for these data was 31.5, slightly higher than the median.

From the procedure, it is clear that every value does not enter into the computation of the median; only the number of values and the values near the midpoint of the distribution enter the computation. If the value of 72 is changed to 224 in the previous example, the new distribution is 8, 10, 10, 18, 24, 29, 36, 48, 60, 224. The median of this distribution is still located midway between 24 and 29 and is still 26.5 hours. Thus, the median is not sensitive to extreme scores. It may be used with symmetrical or asymmetrical distributions, but is especially useful when the data are skewed. However, this property of not summarizing all values in a distribution is also the median's chief shortcoming: it means that the median cannot be algebraically derived. The median merely represents the point in a distribution below which 50% of the scores fall.

The median is appropriate for interval or ratio data and for ordinal data but not for nominal data. It can be used for open-end or censored data, such as "time to recurrence," if more than half of the sample has contributed a value to the distribution.

Mode

The mode, the most frequent value or category in a distribution, is not calculated but is simply spotted by inspecting the values in a distribution. In the previous example of length of hospital stay in hours, the 10 entries were 8, 10, 10, 18, 24, 29, 36, 48, 60, and 72. The mode for this distribution is 10 because that score occurs most frequently.

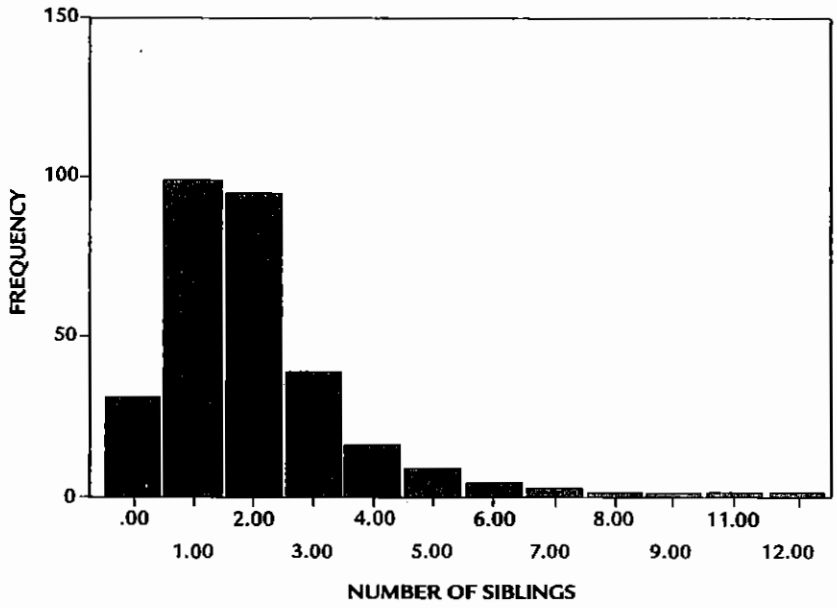


FIGURE 2-1. Relative frequency distribution of number of siblings a child has. (Data collected with a grant funded by the National Institute of Nursing Research, R01 NR04838-01A2. P.I., Vessey, J. (2000). *Development of the CATS: Child-Adolescent Teasing Scale.* The William F. Connell School of Nursing, Boston College.)

If all scores in a distribution are different, the mode does not exist. If several values occur with equal frequency, then there are several modes. If the values of a distribution cluster in several places but with unequal frequency, then there are primary and secondary modes. For example, when discussing the bar chart showing the frequency distribution of number of siblings a child has in Fig. 2-1, it is helpful to note that the primary mode was 1 (the midpoint of the second bar) with a secondary mode of 2 (the midpoint of the third bar). Alternatively, the primary mode for Fig. 2-1 could be reported as 1 sibling and the secondary mode 2 siblings.

For strictly nominal-level variables, the mode is the only appropriate measure of central tendency. It is reported as the modal category. For instance, in Fig. 2-2, the modal category for data collection site is "Albuquerque, NM."

The mode can also be used with interval, ratio, or ordinal variables as a rough estimate of central tendency. Obtaining the mode for numeric data consists of noting which value occurs most frequently. The modal value is the most frequently occurring actual value in the distribution, not the value that has the largest frequency of scores.

Comparison of Measures of Central Tendency

The mean is the most common measure of central tendency. It has a formula and is the most trustworthy estimate of a population average. Generally, researchers prefer

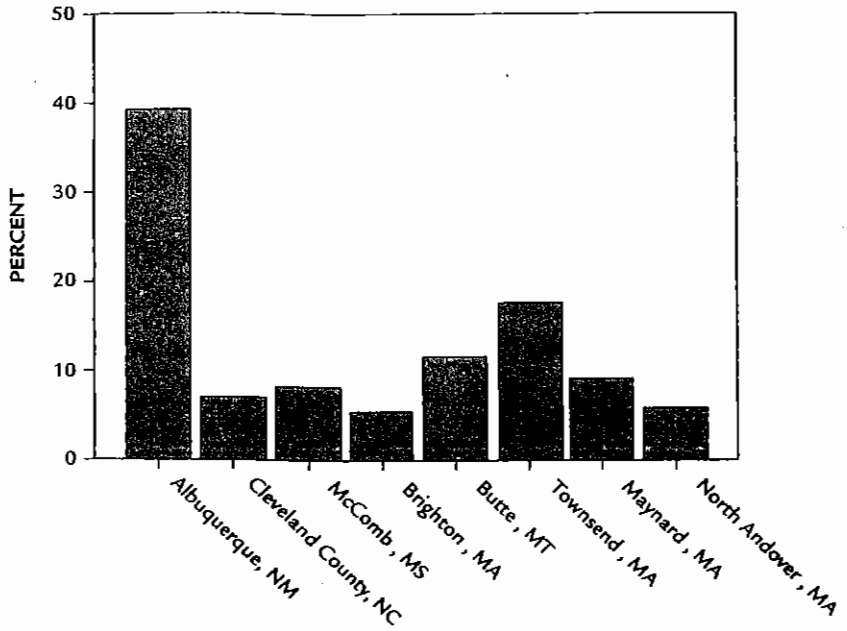


FIGURE 2-2. Data collection site, $N = 764$. (Data collected with a grant funded by the National Institute of Nursing Research, R01 NR04838. P.I., Vessey, J. (2000). *Development of the CATS: Child-Adolescent Teasing Scale.* The William F. Connell School of Nursing, Boston College.)

to use the mean, unless there is a good reason for not doing so. The most compelling reason for not using the mean is a distribution that is badly skewed. The effect of extreme values on the mean diminishes as the size of the sample increases; therefore, another good reason for not using the mean is a small sample with a few extreme values. The mean is best when used with distributions that are reasonably symmetrical and that have one mode.

The median is easy to understand as the 50th percentile of a distribution or the bisector of the area of a histogram. It has no formula but is calculated by a counting procedure, and is usually produced by statistical computer programs. The median may be used with distributions of any shape but is especially useful with very nonsymmetrical distributions because it is not sensitive to skewness.

The main use of the mode is for calling attention to a distribution in which the values cluster at one or more places. It can also be used for making rough estimates. In addition, the mode is the only measure of central tendency available for nominal data.

When a distribution has only one mode and is symmetrical, the mean, median, and mode will have, or very nearly have, the same value. In a skewed, or nonsymmetrical, distribution like that in Fig. 2-3, the mode is the value under the high point of the polygon, the mean is pulled to the right by the extreme values in the tail of the distribution, and the median usually falls in between. Thus, if the mean is greater than the median, then the distribution is *positively skewed,* with the mean being

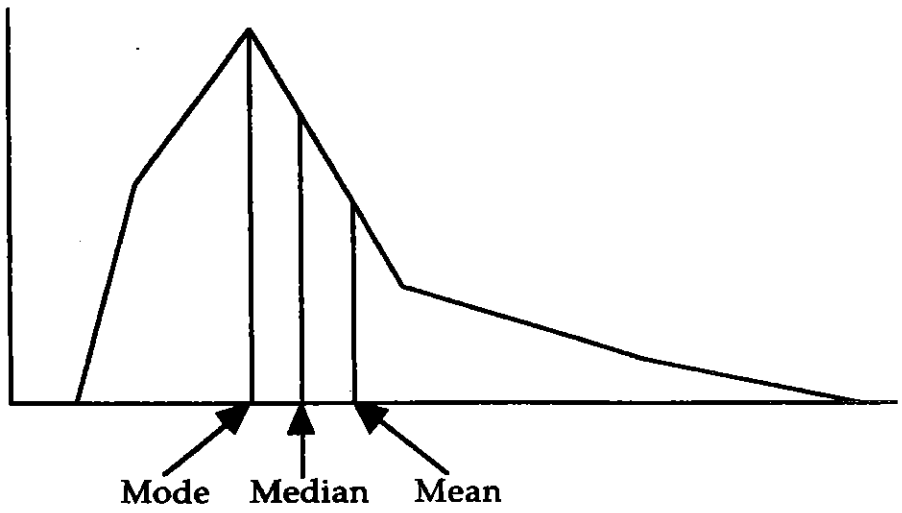


FIGURE 2-3. Sketch of frequency polygon for a distribution skewed to the right, indicating the relative positions of mean, median, and mode.

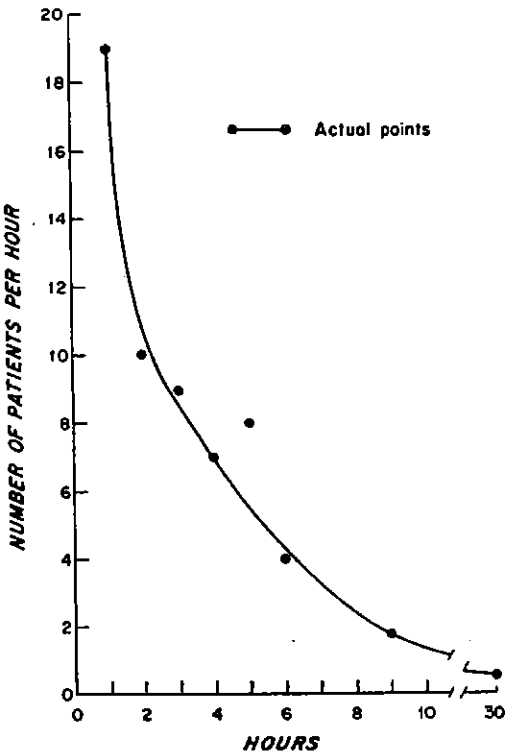


FIGURE 2-4. Graph illustrating a distribution (rate at which patients seek medical care for coronary symptoms as a function of time from onset of symptoms) in which mean, median, and mode are quite different. (From Hackett, T. P., & Cassem, N. H. [1969]. Factors contributing to delay in responding to the signs and symptoms of acute myocardial infarction. *American Journal of Cardiology, 24,* 653. With permission from Excerpta Medica Inc.) Mean, 10.6 hours; median, 4 hours; mode, 1 hour.

dragged to the right by a few high scores. If the mean is less than the median, then the distribution is *negatively skewed*, with the mean being pulled to the left by a small number of low scores.

Weisberg (1992) points out that it is not always necessary to select only a single measure of central tendency because these statistics provide different information. Sometimes it is useful to examine multiple aspects of a distribution. An example from a research journal is presented in Fig. 2-4. In this case, the mode for delay in seeking treatment was 1 hour, the median was 4 hours, and the mean was 10.6 hours. If the objective of reporting an average is to present a fair view of the data, consider which average (or averages) should be used here.

MEASURES OF VARIABILITY OR SCATTER

Reporting only an average without an accompanying measure of variability, or dispersion, is a good way to misrepresent a set of data. A common story in statistics classes tells of the woman who had her head in an oven and her feet in a bucket of ice water. When asked how she felt, the reply was, "On the average, I feel fine." Researchers tend to focus on measures of central tendency and neglect how the data are scattered, but variability is at least equally important (Tulman & Jacobsen, 1989). Two datasets can have the same average but very different variabilities (Fig. 2-5). If scores in a distribution are similar, they are *homogeneous* (having low variability); if scores are not similar, they are *heterogeneous* (having high variability).

The three measures of variability discussed in this text are the standard deviation (SD), range, and interpercentile measures. Unlike averages, which are points

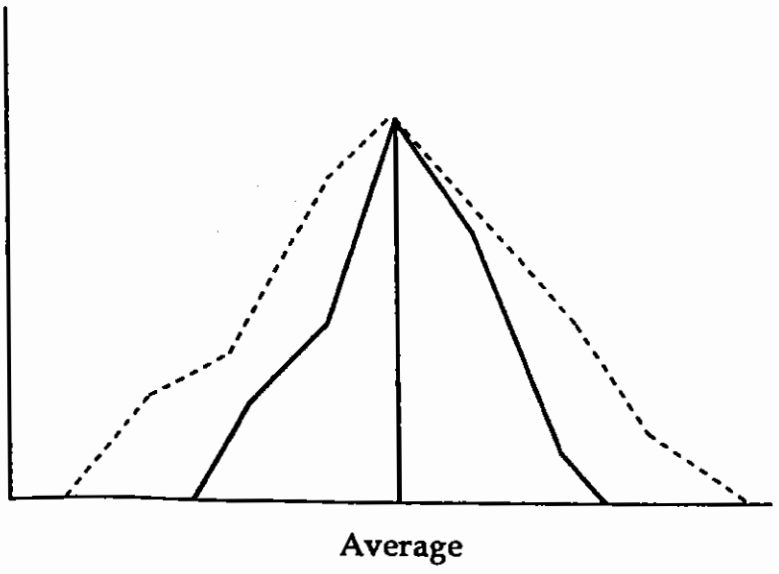


FIGURE 2-5. Two frequency distributions with equal averages but different variabilities.

TABLE 2-2 *Demonstration of the Calculation of the Sample Standard Deviation for Length of Stay (in Hours) in the Hospital for a Sample of Older Women*

X	$X - M$	$(X - M)^2$
8	$8 - 31.5 = -23.5$	$(-23.5)^2 = 552.25$
10	$10 - 31.5 = -21.5$	$(-21.5)^2 = 462.25$
10	$10 - 31.5 = -21.5$	$(-21.5)^2 = 462.25$
18	$18 - 31.5 = -13.5$	$(-13.5)^2 = 182.25$
24	$24 - 31.5 = -7.5$	$(-7.5)^2 = 56.25$
29	$29 - 31.5 = -2.5$	$(-2.5)^2 = 6.25$
36	$36 - 31.5 = +4.5$	$(+4.5)^2 = 20.25$
48	$48 - 31.5 = +16.5$	$(+16.5)^2 = 272.25$
60	$60 - 31.5 = +28.5$	$(+28.5)^2 = 812.25$
72	$72 - 31.5 = +40.5$	$(+40.5)^2 = 1640.25$
$\Sigma X = 315$	$\Sigma(X - M) = 0$	$\Sigma(X - M)^2 = 4466.50 =$ Sum of squares
$M = 31.5$		

Variance = 4466.50/9 = 496.28 square hours

SD = Square root of 496.3 = 22.28 hours

representing a central value, measures of variability should be interpreted as distances on a scale of values.

Standard Deviation

This is the most widely used measure of variability. The sample[2] SD is defined as:

$$SD = \text{square Root of } \Sigma(X - M)^2/n - 1$$

The reason for dividing by the quantity $(n - 1)$ involves a theoretical consideration called *degrees of freedom*. This concept is discussed later in this text. Briefly, it can be shown that using $(n - 1)$ produces, for a random sample, an unbiased estimate of a population variance. This consideration assumes more importance with small samples.

Table 2-2 illustrates the calculation of the SD for the list of 10 values for the variable "length of stay in the hospital" for a sample of older women. The first step is to calculate the mean and then subtract it from each value, making sure that the sum

[2]The SD of a population is represented symbolically by the lowercase Greek letter sigma (σ). The formula differs from the sample SD in that the denominator is simply N, not $n - 1$.

of the deviations is zero. Next, each deviation is squared. The sum of the squared deviations (or sum of squares) is then divided by (n − 1). This quantity is called the *variance*. Although it is a measure of variability, the variance is not used as a descriptive statistic because it is not in the same unit as the data. For example, the variance of the data in Table 2-2 is 496.28 square hours. Most people would have difficulty interpreting a "square hour." Therefore, the square root is taken to return the statistic to its original scale of measurement. The resulting statistic of 22.28 hours is the SD. Again, as with the mean, it is clear that every value in the distribution enters into the calculation of the SD. It is also clear from the formula that the SD is a measure of variability around the mean. The formula in Table 2-2 provides the basic understanding of the SD.

The SD, like the mean, is sensitive to extreme values. For example, in Table 2-2, if the value of 72 is changed to 224, the new SD is 35.15 hours, a large inflation from the original SD of 22.28 hours. Therefore, the SD serves best for distributions that are symmetrical and have a single peak. In general, if it is appropriate to calculate the mean, then it is appropriate to calculate the SD.

The SD has a straightforward interpretation if the distribution is bell-shaped or normal (the normal curve is discussed in detail in the next chapter). If the distribution is perfectly bell-shaped, 68% of the values are within 1 SD of the mean, 95% of the values are within 2 SD of the mean, and more than 99% of the data will be within 3 SD of the mean. For example, Table 2-3 displays the basic statistics for the approximately bell-shaped distribution of a set of denial scores from a sample of 152 heart attack patients. The mean for the denial scale is 46.5, and the SD is 16.4. To determine the range of patients falling ±1 SD from the mean, you subtract the SD from the mean to determine the lower limit (46.5 − 16.4 = 30.1) and add the SD to the mean to determine the upper limit (46.5 + 16.4 = 62.9). After rounding

TABLE 2-3 *Descriptive Statistics Produced by SPSS for a set of Denial Scores From a Sample of 152 Heart Attack Patients*

Program

FREQUENCIES VARIABLES = DENIAL/FORMAT = NOTABLE/STATISTICS = ALL.

Output

Mean	46.533	Std err	1.328	Median	45.500
Mode	45.000	SD	16.374	Variance	286.092
Kurtosis	−.249	S E Kurt	.391	Skewness	.195
S E Skew	.197	Range	76.000	Minimum	11.000
Maximum	87.000	Sum	7073.000		
Valid cases	152	Missing cases	0		

Jacobsen, B. S., & Lowery, B. J. (1992). Further analysis of the psychometric properties of the Levine Denial of Illness Scale. *Psychosomatic Medicine, 54*, 372–381.

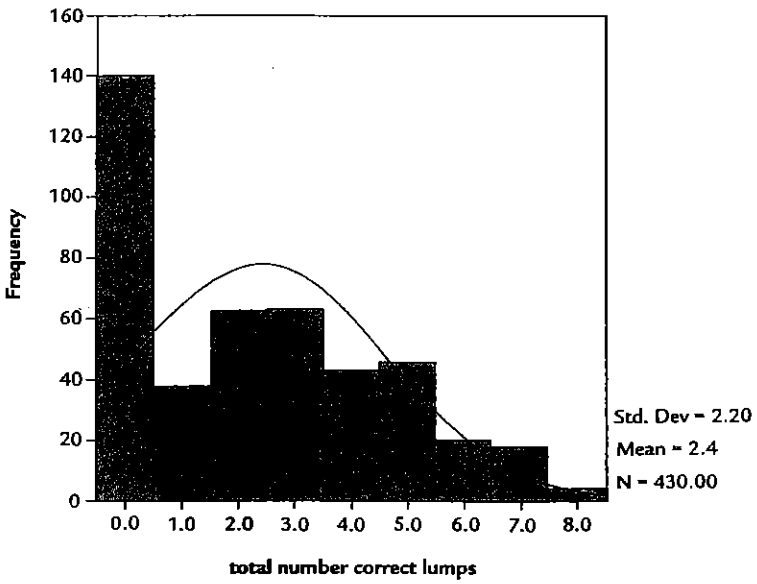


FIGURE 2-6. Number of correct breast lumps identified by a sample of 246 older Black women. (Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)

to whole numbers, 46 of the 152 heart attack patients, or 68% of the sample, had denial scores ranging from 30 to 63 and falling within 1 SD of the mean on the denial scale.

Even if the distribution is not perfectly symmetrical, however, this percentage holds fairly well. Chebyshev's theorem maintains that even in oddly shaped distributions, at least 75% of the data will fall within 2 SD of the mean (Freund, 1988). Figure 2-6 displays a positively skewed distribution, with a mean of 2.4 and an SD of 2.2. By actual count, about 87% of the values lie within the interval of mean ± 1 SD. Because this distribution is decidedly not bell-shaped, the percentage in this interval is different from the expected 68%. Subtracting 2 SD from the mean of 2.4 leads to the absurd conclusion that some older women had a -2.0 correct identification of breast lumps score!

Because the SD, like the mean, is algebraic, formulas have been developed for combining SD from several distributions with different sample sizes to compare measures of variability across different samples from different studies. The *coefficient of variation* (CV) is a useful statistic for comparing SD between several investigations examining the same variable (Daniel, 1987). This statistic is defined as:

$$CV = 100\ (SD/\bar{X}) \text{ or } 100\ (SD/m)$$

Because the CV expresses the SD as a percentage of the mean value, it lets the researcher compare the variability of different variables (Norusis, 2002). For example, Spielberger (1983) reported the following statistics on the State-Trait Anxiety Inventory for a sample of depressed patients: mean = 54.43 and SD = 13.02. For general medical or surgical patients without depression, the statistics were: mean = 42.68 and SD = 13.76. The CV for the depressed group was 24%; the

CV for the nondepressed group was 32%. Thus, the nondepressed group was more variable relative to their mean than the depressed group.

Range

The range, the simplest measure of variability, is the difference between the maximum value of the distribution and the minimum value. In Table 2-2, the range is $72 - 8 = 64$. If the range is reported in a research journal, it would ordinarily be given as a maximum and a minimum, without the subtracted value.

The range can be unstable because it is based on only two values in the distribution and because it tends to increase with sample size. It is sensitive to extreme values. For example, in Table 2-2, if the single value of 72 is changed to 224, the range would then be $224 - 8 = 216$, a tremendous increase.

The main use of the range is for making a quick estimate of variability, but it can be informative in certain situations. For example, a health researcher who is considering subgroup analyses may be interested in knowing the most extreme values in a particular variable's distribution. A researcher who intends to report the SD may also choose to report the range for the additional information it provides about the two endpoints of a distribution.

Interpercentile Measures

A *percentile* is a score value above which and below which a certain percentage of values in a distribution fall (Norusis, 2002). Percentiles are symbolized by the letter P, with a subscript indicating the percentage below the score value. Hence, P_{60} refers to the 60th percentile and stands for the score below which 60% of values fall. The statement "$P_{40} = 55$" means that 40% of the values in the distribution fall below the score 55.

Percentiles allow us to describe a score in relation to other scores in a distribution. The 25th percentile is called the *first quartile*; the 50th percentile, the *second quartile* or more commonly the *median*; and the 75th percentile, the *third quartile*. A score is not said to fall within a quartile, because the quartile is only one point. Therefore, the third quartile is not from 50 to 75; it is just the 75th percentile.

There are several interpercentile measures of variability, the most common being the *interquartile range* (IQR). The IQR is defined as the range of the values extending from the 25th percentile to the 75th percentile. To locate the first quartile, first locate the median of the distribution. The first quartile is the middle value of all the data points below the median; the third quartile is the middle value of all the data points above the median. In the previous example, the set of ordered values was 8, 10, 10, 18, 24, 29, 36, 48, 60, 72. The 50th percentile was noted to be 26.5; there are five values below 26.5. The median of those five values is 10, and the median of the five values above the 50th percentile is 48. Thus, the IQR is 48 to 10.

Other frequently used interpercentile ranges are (P_{10} to P_{90}) and (P_3 to P_{97}). The latter interpercentile range identifies the middle 94% of a distribution, a percentage similar to that identified in a bell-shaped distribution by the mean ±2 SD. Table 2-4

TABLE 2-4 *Selected Percentiles Produced by SPSS 12.0 for the Data in FIGURE 2-6; Number of Correct Breast Lumps Identified by Older Black Women (N = 246).(Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)*

Frequencies

ACLUMPS	***Statistics*** ***Total Number of Correct Lumps***	
N	*Valid*	*430*
	Missing	*9*
Percentiles	10	.0000
	20	.0000
	30	.0000
	40	1.0000
	50	2.0000
	60	3.0000
	70	3.0000
	80	4.0000
	90	5.0000

contains a printout of selected computer percentiles for the variable, number of correct breast lumps identified by a sample of older Black women.

These interpercentile ranges, like the median, are not sensitive to extreme values. If a distribution is badly skewed and the researcher judges that the median (P_{50}) is the appropriate average, then the IQR (or other interpercentile measure) is also appropriate. One of the most common uses of interpercentile measures is for growth charts.

Comparison of Measures of Variability

The SD is the most widely reported measure of variability. It has a formula and is the most reliable estimate of population variability. Generally, researchers prefer to use the SD, unless there is a good reason for not doing so. Like the mean, the most compelling reason for not using the SD is a distribution that has extreme values. The SD is best with distributions that are reasonably symmetrical and have only one mode.

The main uses of the range are to call attention to the two extreme values of a distribution and for quick, rough estimates of variability. The range has a serious shortcoming as a measure of variability because it is greatly influenced by sample size. Because the range is determined by only the smallest and largest values in a distribution, other things being equal, the larger the sample, the larger the range (Glass & Hopkins, 1996).

Interpercentile measures are easy to understand. In a histogram, they mark off a certain percentage of area around the median. For example, the IQR, extending from P_{25} to P_{75}, delineates the middle 50% of a distribution. These measures have no formulas but are calculated by a counting procedure. They can be used with distributions of any shape but are especially useful with very skewed distributions.

To choose the appropriate measures of variability, the researcher must know how a set of scores on a variable is distributed. All of the above measures of variability are intended for use with interval or ratio variables, and often they are sensible for ordinal values. There are no measures of variability for nominal data in common use (Weisberg, 1992).

MEASURES OF SKEWNESS OR SYMMETRY

In addition to central tendency and variability, *symmetry* is an important characteristic of a distribution. A *normal distribution* is *symmetrical* and bell-shaped, having only one mode. When a variable's distribution is asymmetrical, it is skewed. A skewed variable is one whose mean is not in the center of the distribution. If there is positive skewness, there is a pileup of cases to the left and the right tail of the distribution is too long. Negative skewness results in a pileup of cases to the right and a too-long left tail (Tabachnick & Fidell, 2001).

Two sets of data can have the same mean and SD but different skewness (see Fig. 2-5). Two measures of symmetry are considered here: Pearson's measure and Fisher's measure. Although rarely mentioned in research reports, these statistics are very useful in determining the degree of symmetry of a variable's distribution. Researchers routinely compute them using statistics produced when running frequency distributions and descriptive statistics on study variables.

Pearson's Skewness Coefficient

This measure of skewness is nonalgebraic but is easily calculated and is useful for quick estimates of symmetry. It is defined as:

$$\text{Skewness} = (\text{mean} - \text{median})/\text{SD}$$

For a perfectly symmetrical distribution, the mean will equal the median, and the skewness coefficient will be 0. If the distribution is positively skewed, as in Fig. 2-1, the mean will be more than the median, and the coefficient will be positive. If the coefficient is negative, the distribution is negatively skewed, and the mean will be less than the median. In general, skewness values will fall between −1 and +1 SD units. Values falling outside this range indicate a substantially skewed distribution (Hair et al., 1998). Hildebrand (1986) states that skewness values above 0.2 or below −0.2 indicate severe skewness.

For the denial score data of Table 2-3, the skewness coefficient is (46.53 − 45.50)/16.37. The resulting value of 0.06 is close to zero. Using Hildebrand's guideline, the value of 0.06 indicates minor, not severe, skewness. The reader should verify this result visually by means of the chart of the denial score data in Fig. 2-7.

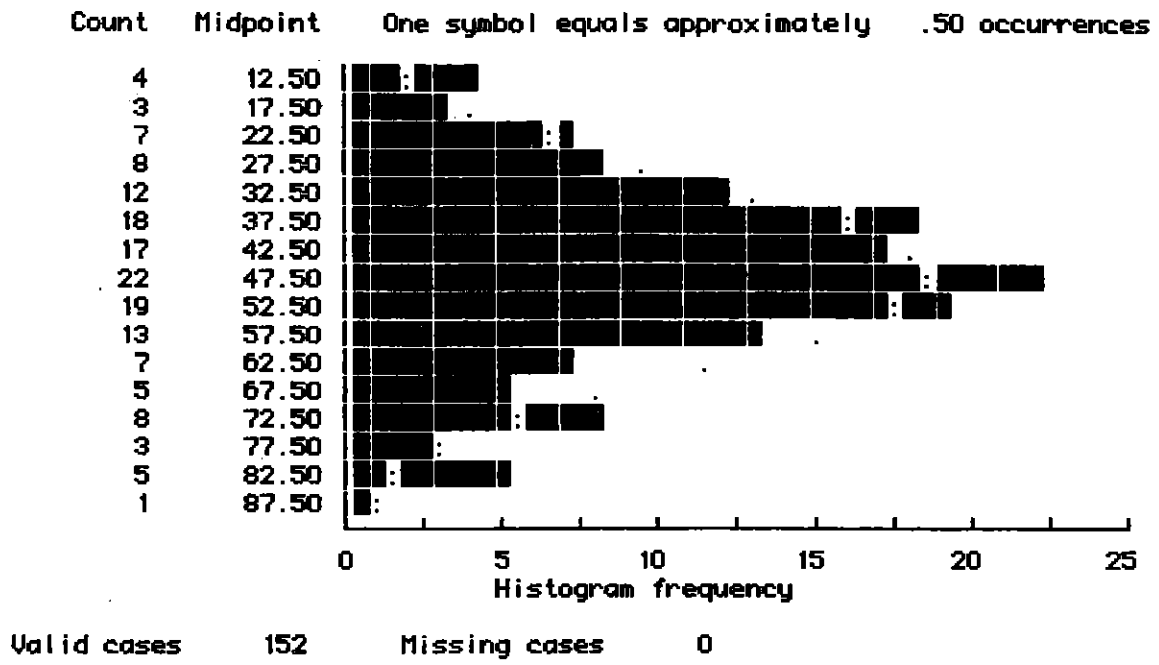


FIGURE 2-7. Example of a histogram produced by SPSS: Denial scores from a sample of 152 heart attack patients. (Jacobsen, B., & Lowery, B. [1992]. Further analysis of the psychometric characteristics of the Levine Denial of Illness Scale. *Psychosomatic Medicine, 54,* 372–381.)

It is a different story for the negatively skewed data in Fig. 2-8. The SF-36 transformed mental health score mean for that distribution is 79.7, the median is 84.0, and the SD is 18.5. Therefore, Pearson's coefficient is (79.7 − 84.0)/18.5, producing a value of −0.23, indicating severe negative skewness.

Fisher's Measure of Skewness

The formula for Fisher's skewness statistic, found in Hildebrand (1986), is based on deviations from the mean to the third power. A symmetrical curve will result in a value of 0. If the skewness value is positive, then the curve is skewed to the right, and vice versa for a distribution skewed to the left. For the denial score data in Table 2-3, Fisher's skewness measure is 0.195. The measure of skewness can be interpreted in terms of the normal curve. (This concept is explained further in the next chapter.) A z-score is calculated by dividing the measure of skewness by the standard error for skewness (0.195/0.197 = 0.99). Values above +1.96 or below −1.96 are significant at the 0.05 level because 95% of the scores in a normal distribution fall between +1.96 and −1.96 SD from the mean. Our value of 0.99 indicates that this distribution is not significantly skewed. Because this statistic is based on deviations to the third power, it is very sensitive to extreme values.

Statistics

TMH

N	Valid	439
	Missing	0
Mean		79.7267
Median		84.0000
Std. Deviation		18.50085

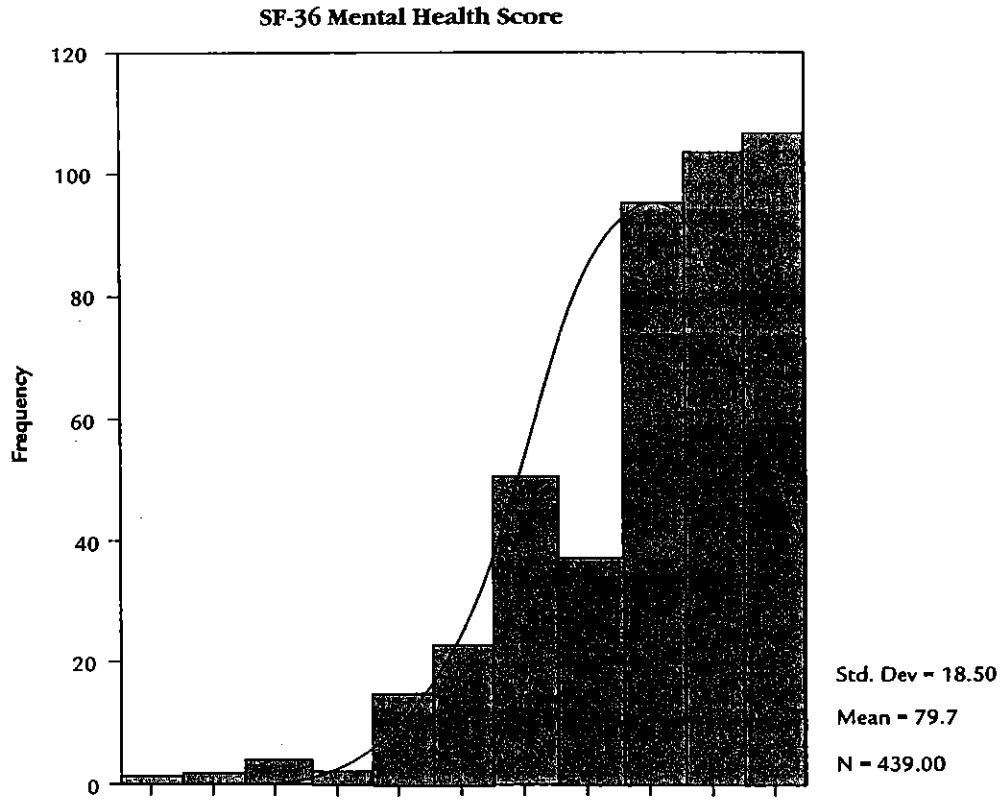


FIGURE 2-8. Example of a histogram produced by SPSS 12.0 for Windows: SF-36 transformed mental health scores from a sample of 439 older women. (Data from Wood, R. Y. [1997]. The development and testing of video breast health kits for older women. National Cancer Institute Small Business Innovation Research (SBIR) Phase II R43 CA 63935-02.)

Types of Data Transformations

Markedly skewed data indicate that the mean is not a good measure of central tendency of scores in the distribution. It is often possible to transform the skewed data so that the new scores display normality and equality of variances. Because variables differ in the extent to which they deviate from normal, Tabachnick and Fidell (2001) recommend the following:

- For moderate skewness, use a square root transformation.
- For substantial skewness, use a log transformation.
- For severe skewness, use an inverse transformation.

Although these authors do not define "moderate," "substantial," and "severe," a practical approach is to start with a square root transformation and see if that results in a more normally distributed variable. If not, then proceed to use a log transformation on the original variable and so on, always checking to see if the transformation reduces the skewness problem. If it does, then use the transformed variable in subsequent statistical analyses.

The direction of the skewness is also considered. For example, when data have a positive skewness, one can proceed directly to undertake either a square root or logarithmic transformation, which often produces data that are more nearly normal. In some cases, the same transformation also achieves equality of variances (Maxwell & Delaney, 1990; Tabachnick & Fidell, 2001). With negative skewness, however, an additional step is required, that of "reflecting" the variable to make the negative skewness a positive skewness. This means that the variable is reverse-scored. For example, with moderate or severe negative skewness, the following procedure needs to be done:

1. "Reflect" the variable by finding the largest score in the distribution, and add one to it to form a constant that is larger than any other score in the distribution.
2. Form a new variable by subtracting each person's score from the constant. Thus, the negative skewness is converted to a positive skewness before transformation. At this stage, the resulting variable, because it was derived from a "reflected" variable, means just the opposite of what it meant before reflection. Thus, if high scores on a self-esteem total score mean high self-esteem, they now mean low self-esteem after reflection.
3. Then apply the appropriate transformation to the newly formed variable.
4. Check the skewness for the transformed variable; if close to zero, then use the transformed variable in subsequent analyses.

If you then use the transformed variable in subsequent analyses, remember that its meaning has been reversed so that a high score now means just the opposite of what it meant before reflecting the variable and transforming it. In order to change the transformed variable back to its original meaning, it is often useful to perform another reflection on the transformed variable. This can be accomplished by finding the largest score in the transformed variable's distribution, add one to it to form a constant that is larger than any other score in the distribution, and form a new variable by subtracting each person's score from the constant. The resulting variable now is interpreted exactly as it was interpreted prior to the first reflection (#s 1 and 2 in the earlier list). If high numbers meant more of that characteristic (ie, self-esteem) before the first reflection, then high numbers again mean more of that characteristic (ie, high self-esteem) after the second reflection.

As a rule, it is best to transform significantly skewed variables to normality unless the transformed scores make interpretation impractical (Tabachnick & Fidell,

2001). Once transformed, always check that the transformed variable is normally or nearly normally distributed. If one type of transformation does not work, try another until you achieve a transformation that produces variables with skewness close to zero and/or the fewest outliers. Finally, if transforming the variable does not work, the best thing might be to create a categorical variable.

There are potential disadvantages to transforming data. Chief among them is that transformed variables may be harder to interpret. Whether or not to transform depends on the scale that measures the variable. If the scale is widely known and used, transformations often hinder interpretation. If the scale is not well known, transformations often do not particularly increase the difficulty of interpretation (Tabachnick & Fidell, 2001). Most computer programs permit various types of transformations through the use of the Compute command.

Hair et al. (1998) recommend keeping several guidelines in mind when carrying out data transformations:

1. For a transformation to have a noticeable effect, the ratio of the variable's mean to its SD should be less than 4.0.
2. When the transformation can be done on either of two variables, transform the variable with the smallest ratio from guideline #1.
3. Transformations should be applied to the independent variable except when heteroscedasticity, or the failure of the assumption of homoscedasticity, is present. (Homoscedasticity is the assumption that the dependent variable displays equal levels of variance across the range of predictor variables.)
4. Heteroscedasticity can be corrected only by transforming the dependent variable in a dependence relationship. If a heteroscedastic relationship is also nonlinear, the dependent variables and possibly the independent variables must also be transformed.
5. Transformations may change how you interpret the variable's score. Thus, you should carefully explore the possible interpretations of the transformed variables.

MEASURES OF KURTOSIS OR PEAKEDNESS

Fisher's Measure of Kurtosis

This statistic, indicating whether a distribution has the right bell shape for a normal curve, measures whether the bell shape is too flat or too peaked. Fisher's measure, based on deviations from the mean to the fourth power, can also be found in Hildebrand (1986). However, the calculation is tedious and is ordinarily done by a computer program. A curve with the correct bell shape will result in a value of zero. If the kurtosis value is a large positive number, the distribution is too peaked to be normal (*leptokurtic*). If the kurtosis value is negative, the curve is too flat to be normal (*platykurtic*). For the denial score data in Table 2-3, the kurtosis statistic is given as −0.249, a value close to zero, indicating that the shape of the bell for this distribution can be called normal. Dividing this value by the standard error for kurtosis

(−0.249/0.391 = −0.64), our distribution is not significantly kurtosed; that is, the value is not beyond ±1.96 SD. Because this statistic is based on deviations to the fourth power, it is very sensitive to extreme values. If a distribution is markedly skewed, there is no particular need to examine kurtosis because the distribution is not normal.

ROUNDING DESCRIPTIVE STATISTICS FOR TABLES

When reporting descriptive statistics in a table, too many digits are confusing. Even though a computer program has provided the statistic to the fourth decimal place, not all the digits need to be reported. If diastolic blood pressure is measured to the nearest whole number, why report descriptive statistics for blood pressure to the nearest 10,000th?

In rounding to the nearest 10th (or 100th), if the last digit to be dropped is less than 5, round to the lower number; if it is higher than 5, round to the higher number. If the last digit to be dropped is exactly 5, no change is made in the preceding digit if it is even, but if it is odd, it is increased by 1. Thus, 4.25 to the nearest 10th is 4.2, but 4.35 becomes 4.4.

CHARTS USING DESCRIPTIVE STATISTICS

Line Charts

In health care research, the line chart is frequently used to display longitudinal trends. Time points in equal intervals are placed on the horizontal axis and the scale for the statistic on the vertical axis. Dots representing the statistic (eg, means, medians, or percentages) at each time point are then connected. The line chart presents a smoother appearance than drawing bars over each time point. Frequently, vertical error bars are added to each time point to indicate the accuracy of the statistic as an estimate of a population parameter. These error bars represent standard errors, which are discussed in detail in Chapter 3. Examples of several types of line charts from research journals are given in Figs. 2-9 through 2-11. When several groups are being compared in the same line chart, Tufte (1983) recommends that labels be integrated into the chart rather than having a separate legend, so the eye is not required to go back and forth.

The issue of whether to place a zero on the vertical axis of a time series line chart is determined by the purpose of the chart and the target group for whom it was designed. Different authors of books on charting have differing views on this subject. Cleveland (1985) and Tufte (1983) both maintain that the vertical axis should start immediately below the lowest value in the dataset. It is best to choose the scale for the vertical axis so that the data fill up the chart. You may assume that the reader of a scientific journal will look at tick mark labels and breaks in the axes or line plot and understand them (Cleveland, 1985).

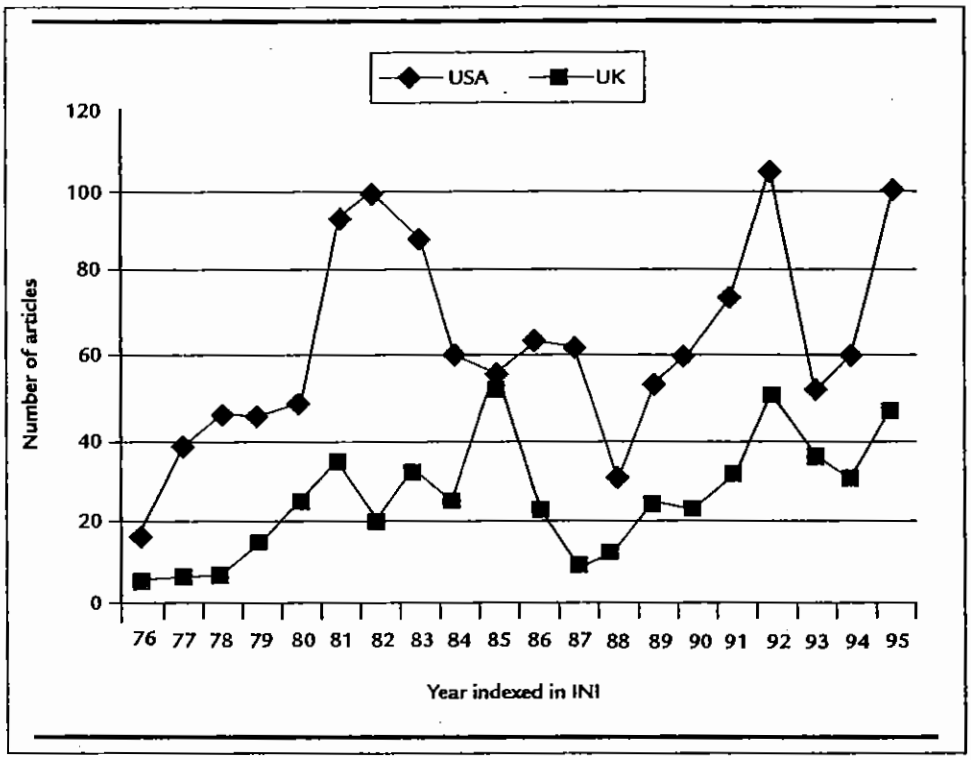


FIGURE 2-9. Comparison of numbers of articles published on patient advocacy in the USA and the UK (1876–1995). (From Mallik, M., & Rafferty, A. [2000]. Diffusion of the concept of patient advocacy. *Journal of Nursing Scholarship, 32*(4), 402.)

Box Plots

A *box plot,* also called a *box-and-whiskers plot,* is a graphic display that uses descriptive statistics based on percentiles (Tukey, 1977). It simultaneously displays the median, the IQR, and the smallest and largest values for a group (Norusis, 2002). Although more compact than a histogram, it does not provide as much detail.

The first step in constructing the box plot is to draw the box. Its length corresponds to the IQR; that is, the box begins with the 25th percentile and ends with the 75th percentile (Fig. 2-12). A line (or other symbol) within the box indicates the location of the median or 50th percentile. Thus, the box provides information about the central tendency and the variability of the middle 50% of the distribution.

The next step is to locate the wild values of the distribution, if any. Calculate the IQR ($P_{75} - P_{25}$), and then multiply this value by 3. Individual scores that are more than three times the IQR from the upper and lower edges of the box are extreme outlying values and are denoted on the plot by a symbol such as E. Next, multiply the IQR by 1.5. Individual scores between 1.5 times the IQR and 3 times the IQR

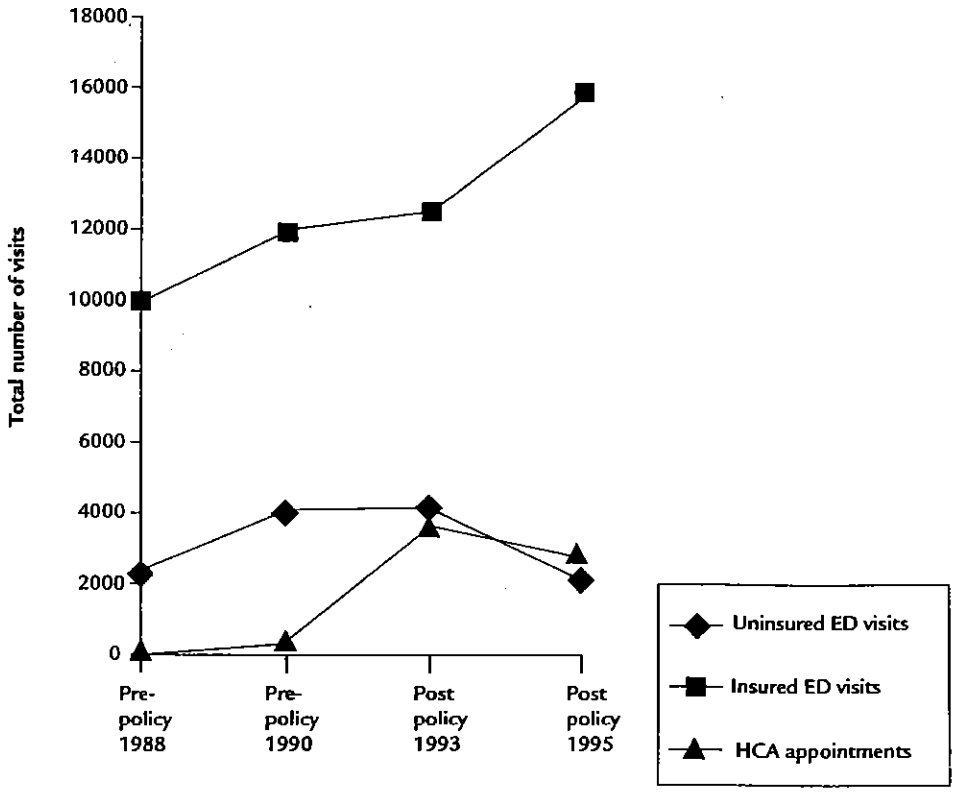


FIGURE 2-10. Total emergency department (ED) visits and HCA (health care access) appointments by year, before and after state funding. (From Smith-Campbell, B. [2000]. Access to health care: Effects of public funding on the uninsured. *Journal of Nursing Scholarship, 32*(3), 298.)

away from the edges of the box are minor outlying values. They are denoted on the box plot with a different symbol, such as O. Finally, draw the whiskers of the box. These lines should extend to the smallest and largest values that are not minor or extreme outlying values. Thus, the whiskers and designation of the outlying values provide more detail about how the lower 25% and upper 25% of the distribution are scattered.

The box plot is particularly well suited for comparisons among several groups. Examples of box plots are given in Fig. 2-13, comparing psychological adjustment to illness in a sample of breast cancer patients according to stage of cancer. As the stage of cancer became higher, adjustment worsened (ie, the average score increased). Also, it is clear that the variability of the adjustment scores was greater as the stage became higher. The subject identification numbers can be placed on the plot for convenient reference.

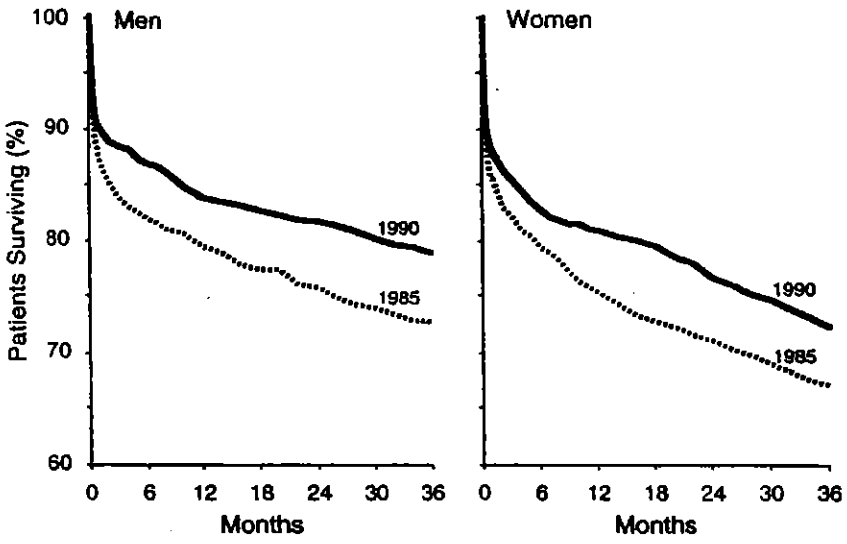


FIGURE 2-11. Trends in survival in the 3 years after hospitalization for definite acute myocardial infarction in 1985 and 1990 among residents of the Twin Cities area who were 30 to 74 years of age. (From McGovern, P. G., Pankow, J. S., Shahar, E., Doliszny, K. M., Folsom, A. R., Blackburn, H., & Leupker, R. V. [1996]. Recent trends in acute coronary heart disease. *New England Journal of Medicine, 34*(14), 887. © 1996, Massachusetts Medical Society. All rights reserved.)

OUTLIERS

Outliers are values that are extreme relative to the bulk of scores in the distribution. They appear to be inconsistent with the rest of the data. Outliers must be appraised by the types of information they provide. In some cases, outliers, despite being different from most of the sample, may be beneficial: They may indicate characteristics of the population that would not be known in the normal course of analysis. In other cases, outliers may be problematic because they do not represent the population, run counter to the objectives of the analysis, and can seriously distort statistical tests (Hair et al., 1998). Thus, it is important to detect outliers to ascertain their type of influence.

The source of an outlier may be any of the following:

1. An error in the recording of the data
2. A failure of data collection, such as not following sample criteria (eg, inadvertently admitting a disoriented patient into a study), a subject not following instructions on a questionnaire, or equipment failure
3. An actual extreme value from an unusual subject

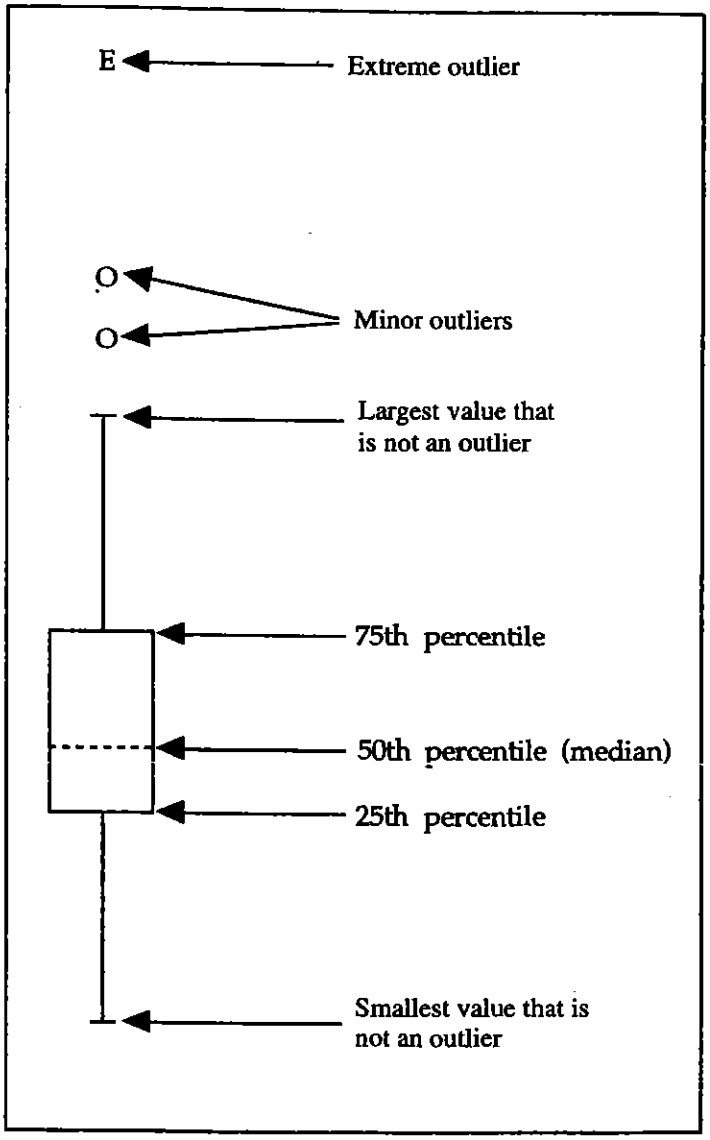


FIGURE 2-12. Schematic diagram of the construction of a box plot.

Outliers must first be identified by an objective method. A traditional way of labeling outliers has been to locate any values that are more than 3 SD from the mean. The problem with this method is that outliers inflate the SD, making it less likely that a value will be 3 SD away from the mean. Tukey's (1977) recommendation was described earlier in connection with the box plot. Values that are more than 3 IQRs from the upper or lower edges of the box are extreme outliers. Values between 1.5 and 3 IQRs from the upper and lower edges of the box are minor outliers. The reason for having an objective method is to prevent undue (perhaps

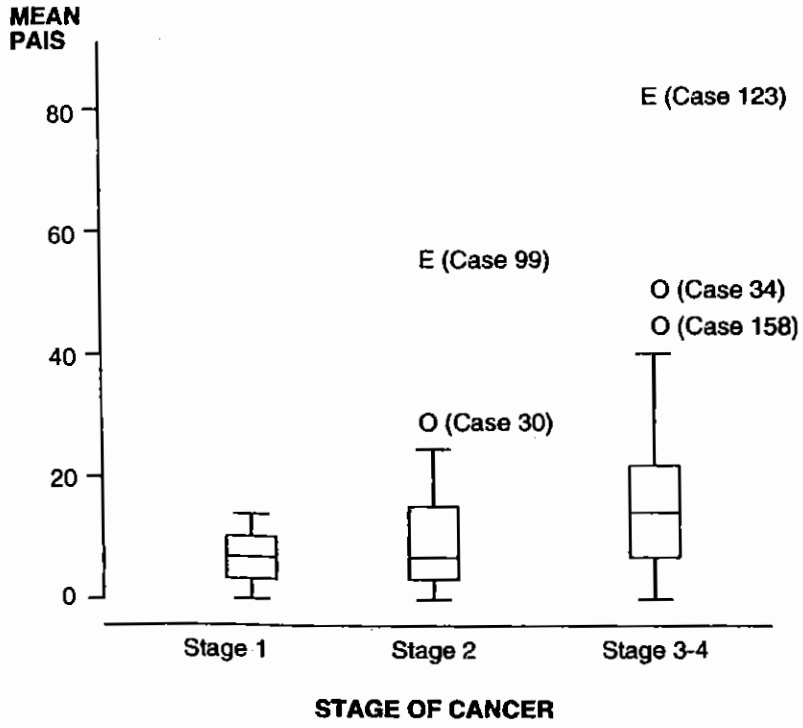


FIGURE 2-13. Box plots of psychological adjustment to illness (PAIS) in a sample of breast cancer patients by stage of cancer (hypothetical data). Higher scores indicate poorer adjustment.

unethical) data manipulation, such as pruning very high or very low values that are not really outliers.

Outliers in a dataset can be identified from univariate, bivariate, and multivariate perspectives. In data analysis, it is best to use as many of these perspectives as possible. Although in-depth discussion of the various approaches is beyond the scope of this chapter, Hair et al. (1998) and Tabachnick and Fidell (2001) provide excellent discussions on detecting and handling outliers in various circumstances.

Once outliers have been identified, the next step is to try to explain them. If they represent errors in coding or a failure in the data collection, then those observations are either discarded or corrected. If the outliers represent actual values or their occurrence in the distribution cannot be explained, the researcher must decide how to deal with them.

Handling Outliers

A frequent suggestion for handling outliers is to analyze the data two ways: with the outliers in the distribution and with the outliers removed. If the results are similar, as they are likely to be if the sample size is large, then the outliers may be ignored. If the results are not similar, then a statistical analysis that is resistant to outliers can be used (eg, median and IQR).

If the researcher wants to use a mean with outliers, then the ***trimmed mean*** is an option. This statistic is calculated with a certain percentage of the extreme values removed from both ends of the distribution. For example, if the sample size is 100, then the 5% trimmed mean is the mean of the middle 90% of the observations. Formulas for using the trimmed mean in statistical inference are given by Koopmans (1987).

Another alternative is a ***winsorized mean***. In the simplest case, the highest and lowest extremes are replaced, respectively, by the next-to-highest value and by the next-to-lowest value. If the sample size is 100, the resulting 100 data points are then processed as if they were the original data. Winer (1971) outlines the techniques for handling statistics computed from winsorized samples.

For univariate outliers, Tabachnick and Fidell (2001) suggest changing the score(s) on the variable(s) for the outlying cases so they are deviant, but not as deviant as they originally were. For example, give the outlying case(s) a raw score on the specific variable that is one unit smaller (or larger) than the next most extreme score in the distribution. Thus, if the two largest scores in the distribution are 125 and 122, and the next largest score is 87, recode 122 as 88 and 125 as 89. This moves these outliers closer to the bulk of scores in the distribution. Sometimes, this conversion is all it takes to handle the problem of severe skewness in a distribution, discussed in Chapter 3.

The actual score a case has is somewhat arbitrary. What is important is that the case still retains its place in the distribution: If the case has the lowest score, it will still have the lowest score after being assigned a number closer to the bulk of scores in the distribution. Any changes in scores should be noted, along with the rationale, in the results section of the research report.

Further details on the treatment of outliers can be found in Mertler and Vannatta (2002), Tabachnick and Fidell (2001), and Hair et al. (1998). A researcher can also view these actual outliers as case material and adopt the advice of Skinner (1972), who advocates that when you encounter something interesting, study it.

MISSING DATA

One of the most pervasive problems in data analysis is what to do about missing data. Most studies have missing information for some variables for some cases. Missing data can occur at the subject and/or item level (Kneipp & McIntosh, 2001). Missing data at the subject level are usually found in longitudinal and repeated measures studies when one or more subjects are lost to follow-up or decide not to continue participation in the study. Missing data at the item level are quite common when one or more items on a survey or questionnaire are not answered by a respondent. Missing data are a problem because all standard statistical techniques presume that each case in a dataset has information on all the variables to be included in the particular analysis (Allison, 2001).

With missing data, the researcher faces three major tasks: to identify the pattern and amount of missing data, to assess why it is missing, and to determine what to do about it.

Pattern and Amount of Missing Data

There are two characteristic patterns of missing data: random and systematic. A *random pattern* consists of values missing in an unplanned or haphazard fashion throughout a dataset. A *systematic pattern* consists of values missing in a methodical, nonrandom way throughout the data. The pattern of missing data is more important than the amount of missing data (Tabachnick & Fidell, 2001). If only a few data values are absent in a random pattern from a large dataset, almost any procedure for handling missing values can be used. However, if many data are missing from a small or moderately sized dataset, serious problems can ensue unless the researcher takes steps to handle the problem. Missing data is such a serious problem in a dataset that all statistical packages have conventions for coding missing data for further study and for special treatment in statistical procedures. For example, SPSS has a "System Missing" category that shows up in the data spreadsheet and on the computer printout as a period.

Random missing data can be one of three categories (Allison, 2001; Little & Rubin, 1987): missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR data have the highest degree of randomness, displaying no underlying reason that would contribute to biased data (Musil et al., 2002). MCAR data are randomly distributed across all cases and are completely unrelated to other variables in the dataset (Hair et al., 1998). In contrast, MAR data display some randomness to the pattern of omission that can be traced or predicted from cases with no missing data. In other words, MAR occurs when the probability of a missing value is not dependent on the value itself but may rely on the values of other variables in the dataset (Allison, 2001). The third type of missing data, NMAR, occurs when the missing values are systematically different from those observed, even among respondents with other similar characteristics (Kneipp & McIntosh, 2001). Systematic missing data, even in a few cases, should always be treated seriously because they affect the generalizability of results.

If you are not sure whether the missing data are random or systematic, you can test for patterns with the following procedure. First, create a grouping variable with two levels (using the Recode, or comparable, command in your computer program), making 1 = cases with missing values on the variable and 0 = cases with no missing values on the variable. Then, perform a test of differences, such as the *t* test, between the two levels on the dependent variable(s). If there are no meaningful differences, how you handle the missing data is not so important. If serious differences are noted, then handling missing data is critical and care should be taken to preserve the missing cases for further analyses (Tabachnick & Fidell, 2001). A useful program for examining missing data is SPSS Missing Values Analysis, which permits you to analyze patterns of missing data and to replace them in the dataset using one of several methods.

Assessing Why Data Are Missing

It is important to understand what, if any, factors led up to missing data in a research study, because the researcher needs to grasp what may have happened to handle

the problem. Hair et al. (1998) define a missing data process as "any systematic event external to the respondent (such as data entry errors or data collection problems) or action on the part of the respondent (such as refusal to answer) that leads to missing data" (p. 46). If the missing data process is under the researcher's control and can be explicitly defined, these missing data can be ignored and no specific remedies are needed because allowances have been made for missing data in the technique used (Allison, 2001; Little & Rubin, 1987).

An example of ignorable missing data inherent in the technique used is the application of probability sampling to select respondents for a study. This sampling method permits the researcher to stipulate that the missing data process leading to the missing data points is random and that the missing data are explained as sampling error in the statistical procedures (Hair et al., 1998).

More often than not, however, the researcher has no idea why specific data are missing. Thus, examining the pattern of missing data becomes important. Are the respondents with missing data on some variables different than the respondents who provided information on these variables? Only by understanding to the greatest extent possible why missing data occurred can the researcher take appropriate steps to handle the impact that it can have on the analyses, the results, and the subsequent interpretation of the data.

Handling Missing Data

Missing data can be handled in several ways: Using complete-case (listwise deletion) and available-case (pairwise deletion) analysis (Kneipp & McIntosh, 2001); deleting cases or variables (Tabachnick & Fidell, 2001); weighting techniques (Patrician, 2002); and estimating missing data through imputation (Tabachnick & Fidell, 2001).

USING OBSERVATIONS: COMPLETE-CASE AND AVAILABLE-CASE ANALYSIS

The easiest and most direct method for dealing with missing data values is to analyze only those cases with complete data. This procedure, called *listwise deletion*, is the default procedure in most major statistical programs such as SPSS, SAS, BMDP, and Systat. As such, numerous cases can be deleted without the researcher's knowledge, resulting in a substantial loss of subjects. Thus, it is important to check the number of cases when running statistical analyses to ensure that all desired cases are used. Hair et al. (1998) suggest using this method if the amount of missing data is small, the sample is sufficiently large to permit deletion of cases with missing data, and the relationships in the data are so strong that they will not be influenced by any missing data process. With MCAR data, listwise deletion produces unbiased parameter estimates but larger standard errors due to the decrease in sample size and can lead to misleading results and decreased analytic power especially if a large number of cases is removed (Patrician, 2002).

Available-case analysis using only those cases that have available data on the variables for a specific analysis is a common research practice. This can be accomplished using *pairwise deletion* of cases with missing data, commonly available as an

option in most statistical packages. This method permits cases to be deleted only if the variables being used in the analysis have missing data. Both listwise and pairwise deletion procedures are ad hoc in nature, have no theoretical justification, and are designed solely to provide complete data for specific analyses. Pairwise deletion is often used for correlations, factor analysis, and linear regression (Allison, 2001).

DELETING CASES OR VARIABLES

Dropping the case or variable is another remedy for handling missing data. The researcher simply determines the extent of missing data on each case and variable in the dataset, then removes the cases or variables with excessive levels. Although there are no hard-and-fast rules for determining excessive levels of missing data, many researchers use a predetermined percentage of missing data as a cutoff for deciding whether to exclude a variable from analysis. It is not unusual to see a 5% or 10% cutoff being used. Usually, this planned cutoff level is based on theoretical and empirical reasons. In many situations, given a large enough sample, this is the most efficient solution. Once cases or variables with missing data are removed, the researcher may discover that the missing data were localized in a small set of cases or variables. Once excluded, the extent of missing data is considerably decreased (Hair et al., 1998).

WEIGHTING TECHNIQUES

Another, less common, way to handle missing data is to disregard missing values and assign a weight to cases with complete data. Little and Rubin (1987) believe that weighting those cases with no missing data higher than those with missing data decreases the bias from case-deletion methods as well as the sample variance, but makes calculating standard errors more difficult.

ESTIMATING MISSING DATA BY IMPUTATION

Imputation is the process of estimating missing data based on valid values of other variables or cases in the sample. The goal of imputation is to use known relationships that can be identified in the valid values of the sample to help estimate the missing data (Hair et al., 1998). Tabachnick and Fidell (2001) discuss five popular ways to estimate missing data: using prior knowledge, inserting mean values, using regression, expectation maximization (EM), and multiple imputation.

Prior knowledge involves replacing a missing value with a value based on an educated guess. This is a reasonable method if the researcher has a good working knowledge of the research domain, the sample is large, and the number of missing values is small. In such circumstances, the researcher is confident that the missing value would have been near the median, or other, value.

Mean replacement (or median replacement for skewed distributions) involves calculating mean values from available data on that variable and using them to replace missing values before analysis. This is a conservative

procedure because the distribution mean as a whole does not change and the researcher does not have to guess at missing values. Hair et al. (1998) cite three disadvantages to this approach: It invalidates the variance estimates derived from the standard variance formulas by understating the data's true variance; it distorts the actual distribution of values; and it depresses the observed correlation that this variable will have with other variables because all missing data have a single constant value, thus reducing variance.

Mean substitution, however, has the advantage of being easily implemented and provides all cases with complete data. A compromise procedure is to insert a group mean for the missing value. If, for example, the case with a missing value is a female patient with hypertension, the mean value for female patients with hypertension is computed and inserted in place of the missing value. This procedure is less conservative than inserting the overall mean value but not as liberal as using prior knowledge (Tabachnick & Fidell, 2001).

Using regression, a more sophisticated method for estimating missing values, involves using other variables in the dataset as independent variables to develop a regression equation for the variable with missing data serving as the dependent variable. Cases with complete data are used to generate the regression equation; the equation is then used to predict missing values for incomplete cases. More regressions are computed, using the predicted values from the previous regression to develop the next equation, until the predicted values from one step to the next are comparable. Predictions from the last regression are the ones used to replace missing values.

Hair et al. (1998) cite four disadvantages to using the regression approach: It reinforces the relationships already in the data, resulting in less generalizability; the variance of the distribution is reduced because the estimate is probably too close to the mean; it assumes that the variable with missing data is correlated substantially with the other variables in the dataset; and the regression procedure is not constrained in the estimates it makes. Thus, the predicted values may not fall in the valid ranges for variables—for instance, a value of 6 may be predicted for a 5-point scale. The main advantage of the regression approach is that it is more objective than the researcher's guess but not as blind as simply using the overall mean (Tabachnick & Fidell, 2001).

Expectation maximization (EM) method, available for randomly missing data, is an iterative process that proceeds in two discrete steps. First, in the Expectation (E) step, the conditional expected value of the complete data is computed and then given the observed values, such as correlations. Second, in the maximization (M) step, these expected values are then substituted for the missing data and maximum likelihood estimation is then computed as though there were no missing data. The procedure iterates until convergence is reached and the filled-in data are saved in the dataset. SPSS Missing Values

Analysis performs EM to produce imputed values and allows some specifications of other nonnormal distributions (Tabachnick & Fidell, 2001).

Multiple imputation (MI), similar to maximum likelihood estimation, produces several datasets and analyzes them separately. One set of parameters is then formed by averaging the resulting estimates and standard errors. The number of datasets to impute derives from the extent of missing data in the dataset, although most statisticians recommend 3 to 5 sets (Patrician, 2002). For an excellent discussion of the various types of multiple imputation, the reader is referred to Allison (2001) and Little and Rubin (2002).

Multiple imputation has a number of advantages:

- It makes no assumptions about whether data are randomly missing (Tabachnick & Fidell, 2001) but incorporates random error because it requires random variation in the imputation process (Patrician, 2002);
- It permits use of complete-data methods for data analysis and also includes the data collector's knowledge (Patrician, 2002);
- It permits estimates of nonlinear models (Allison, 2001);
- It simulates proper inference from data and increases efficiency of the estimates (Patrician, 2002) by minimizing standard errors (Rubin, 1987); and
- It is the method of choice for databases that are made available for analyses outside the agency that collected the data (Tabachnick & Fidell, 2001).

MI has the following major disadvantages:

- It requires computational intensiveness to carry out MI, including special software and model building (Kneipp & McIntosh, 2001), although this has become less so in recent years due to technological advances;
- It does not produce a unique answer because randomness is preserved in the MI process, making reproducibility of exact results problematic (Patrician, 2002); and
- It requires large amounts of data storage space that often exceeds space on personal computers' hard drives or the amounts allotted on university-shared drives, especially when national datasets with thousands of respondents are used (Kneipp & McIntosh, 2001).

When using imputation methods, Tabachnick and Fidell (2001) recommend repeating analyses with and without missing data to make sure that the results do not get distorted by imputed values. This can be particularly problematic if the dataset is small.

SUMMARY

Descriptive statistics based on the mean are best for distributions that are reasonably symmetrical and have a single peak. These measures include the mean, the SD, Pearson's skewness coefficient, and Fisher's measures of skewness and

kurtosis. For skewed distributions, the median and the IQR are less influenced by extreme scores. The range, the mode, Pearson's coefficient of skewness, and Fisher's measure of skewness are quick estimates. In addition, the mode is informative when a distribution has several peaks. The range is useful for locating the most extreme values. Outliers are extreme values that meet objective criteria, and researchers must consider carefully how to handle them in data analysis. Two charts that make use of summary statistics are the line chart and the box plot. A line chart uses statistics such as means at various time points to portray longitudinal trends. Box plots emphasize extreme values in a distribution and are handy for displaying outliers. Missing data are a fact of life in data analysis. The researcher must determine the pattern and amount of missing data, why it occurred, and what method is best for handling it. No one technique can solve all problems with missing data.

Application Exercises and Results

Exercises

1. Access the dataset named SURVEY03.SAV. Run frequencies and include statistics for all variables. Examine the output for outliers, marked skewness, unequal groups, and missing data. Decide what to do about the problems you encounter.
2. Construct a table that includes some of the categorical variables in this dataset. Write a description of the table.
3. Construct a table that includes some of the continuous variables in this dataset. Write a description of the table.
4. Construct a box plot (sometimes called a box-and-whiskers plot) for the variable EDUC by GENDER.

Results

1. Generally, when you first look at output, you will find invalid numbers—that is, a number that is not valid for a particular variable. With SURVEY03.SAV, we tried to remove all invalid numbers, so unless we missed one, you should not have found any.

   For an example of an outlier, find the frequencies for the variable AGE in the printout and examine them. You should see one 78-year-old, one 79-year-old, two 82-year-olds, one 83-year-old, and one 95-year-old. The next largest age is 74 years. We questioned whether the 95-year-old was a data entry error but were assured by the student who collected the data that the individual was indeed 95 years old. These five data points could be viewed as outliers because they are several units away from the next highest age score of 74 years. If we follow Tabachnick and Fidell's (2001) suggestion for univariate outliers, we would change the scores of the five outliers in the AGE variable to bring them closer to the bulk of the distribution's scores. In all versions of SPSS for Windows, we accomplish this by clicking on the Transform, Recode, and Into Different Variables menus. In this last window, highlight the AGE variable in the variable column and paste it in the Numeric Variable box. Give the output variable a new name such as RECAGE, then click Change. This will move the new variable name into the Numeric Variable window following the arrow.

Click on Old and New Values and specify how to recode the values. Under Old Value, type the number 78 in the Value box. Under New Value, type the number 74 in the Value box. Click Add to paste the conversion into the Old → New window. Repeat this procedure for the additional four values (79, 82, 83, and 95, making them 75, 76, 77, and 78, respectively). Then, in the Old Value column, click on All Other Values at the bottom, then click on the Copy All Value(s) in the New Value column and hit the Add button. (If you do not Copy All Values, your resulting variable will have only six cases because all of the other values were not moved to the new variable, RECAGE.) When you have completed these operations, you should have five value statements in the Old S New window. To complete the Recode procedure, click Continue, then OK. The transformation will then take place and the new variable RECAGE will appear as the last variable in the data file window. Next, return to the spreadsheet and switch to the Variable View. Scroll down the variable list to the last variable, RECAGE. Now, move horizontally to the column named Value and type in the following: Recoded Age Variable Making 78, 79, 82, 83, 95 into 74, 75, 76, 77, and 78. Finally, to make sure you do not lose this newly created variable when you exit the program, save the data file at this time.

Exercise Fig. 2-1 contains the descriptive statistics for several variables in the dataset. The recoded AGE variable (RECAGE) resulted in small changes in the mean, standard error of the mean, SD, variance, range, and minimum and maximum values. Skewness was reduced from 0.753 to 0.643; kurtosis was greatly changed from 0.796 to 0.287.

Skewness is often a problem in data analysis and violates the assumptions underlying parametric tests. Look at the variable HEALTH (overall state of health). It was scored from 1 = Very Sick to 10 = Very Healthy. Only 12.9% of the distribution falls between the scores of 1 and 5. Only one respondent had a rating of 1; no respondents rated themselves as 2. This is understandable because students are not likely to request data from people who are very ill. You can tell by looking at the distribution that it is negatively skewed (ie, the values tail off at the lower end). The value for skewness (−0.961) divided by the standard error of skewness (0.092) yields −10.44, indicating significant skewness beyond the 0.01 level (critical value = 2.58 SD from the mean). You might try to transform this variable and see if you can create a normally distributed variable. Remember, it is negatively skewed and would require "reflecting" before being transformed.

There are a number of examples of *unequal groups* in the dataset, SURVEY03.SAV. Males make up only 36.6% of the sample. Only 11.3% of the sample still smokes, and only 1.6% are routinely depressed. What other examples of unequal groups can you find?

Although there are quite a lot of *missing data* across the dataset, no variable has more than 5% missing data. This indicates a random rather than a systematic pattern of missing data. If you examine the frequency distributions, you will note that EDUC has 28 (4.0%) missing data points and SMOKE (Smoking History) has 4 (0.6%) missing data points. Are there any other variables with similar amounts of missing data?

We now need to determine what method we should use to handle the missing data problem. If we choose the most conservative method and use only the cases with no missing data, we might end up with just a few cases in our dataset. This is because there are missing data in 44 (86.3%) of the 51 variables in our dataset. So, we now consider the next most conservative approach for handling missing data: dropping the case or the variable with excessive levels of missing data. If we examine the frequencies for the variables in the dataset, we see that TOTAL, the total score for the IPPA scale, has missing data for 40 (5.7%) of cases. Why is this so high when the items that were summed to compute the total

(text continues on page 69)

Statistics

		AGE subject's age	RECAGE Recorded Age (78,79,82 83,95 into 75,76,77, 78,79)
N	Valid	684	684
	Missing	17	17
Mean		38.06	38.0073
Std. Error of Mean		.496	.48884
Median		37.00	37.0000
Mode		28	28.00
Std. Deviation		12.972	12.78479
Variance		168.266	163.45090
Skewness		.753	.643
Std. Error of Skewness		.093	.093
Kurtosis		.796	.287
Std. Error of Kurtosis		.796	.287
Range		80	64.00
Minimum		15	15.00
Maximum		95	79.00

Statistics

HEALTH overall state of health

N	Valid	700
	Missing	1
Mean		7.85
Std. Error of Mean		.066
Median		8.00
Mode		9
Std. Deviation		1.740
Variance		3.027
Skewness		−.961
Std. Error of Skewness		.092
Kurtosis		.555
Std. Error of Kurtosis		.185
Range		9
Minimum		1
Maximum		10

EXERCISE FIGURE 2–1. Descriptive statistics for study variables.

HEALTH overall state of health

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very Sick	1	.1	.1	.1
	3	17	2.4	2.4	2.6
	4	21	3.0	3.0	5.6
	5	51	7.3	7.3	12.9
	6	31	4.4	4.4	17.3
	7	117	16.7	16.7	34.0
	8	170	24.3	24.3	58.3
	9	182	26.0	26.0	84.3
	10 Very Healthy	110	15.7	15.7	100.0
	Total	700	99.9	100.0	
Missing	System	1	.1		
Total		701	100.0		

Frequency Table

GENDER gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 male	255	36.4	36.6	36.6
	1 female	441	62.9	63.4	100.0
	Total	696	99.3	100.0	
Missing	System	5	.7		
Total		701	100.0		

SMOKE Smoking History

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Never Smoked	433	61.8	62.1	62.1
	1 Quit Smoking	185	26.4	26.5	88.7
	2 Still Smoking	79	11.3	11.3	100.0
	Total	697	99.4	100.0	
Missing	System	4	.6		
Total		701	100.0		

EXERCISE FIGURE 2–1. (*Continued*).

DEPRESS depressed state of mind

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Rarely	361	51.5	51.6	51.6
	2 Sometimes	280	39.9	40.0	91.6
	3 Often	48	6.8	6.9	98.4
	4. Routinely	11	1.6	1.6	100.0
	Total	700	99.9	100.0	
Missing	System	1	.1		
Total		701	100.0		

Frequency Table

EDUC education in years

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7	1	.1	.1	.1
	8	12	1.7	1.8	1.9
	9	2	.3	.3	2.2
	10	6	.9	.9	3.1
	11	5	.7	.7	3.9
	12	70	10.0	10.4	14.3
	13	25	3.6	3.7	18.0
	14	3	.4	.4	18.4
	14	63	9.0	9.4	27.8
	15	28	4.0	4.2	31.9
	16	147	21.0	21.8	53.8
	17	40	5.7	5.9	59.7
	18	1	.1	.1	59.9
	18	84	12.0	12.5	72.4
	19	46	6.6	6.8	79.2
	20	1	.1	.1	79.3
	20	65	9.3	9.7	89.0
	21	27	3.9	4.0	93.0
	22	21	3.0	3.1	96.1
	23	7	1.0	1.0	97.2
	24	1	.1	.1	97.3
	24	7	1.0	1.0	98.4
	25	5	.7	.7	99.1
	26	2	.3	.3	99.4
	28	3	.4	.4	99.9
	30	1	.1	.1	100.0
	Total	673	96.0	100.0	
Missing	System	28	4.0		
Total		701	100.0		

EXERCISE FIGURE 2–1. *(Continued).*

SMOKE Smoking History

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Never Smoked	433	61.8	62.1	62.1
	1 Quit Smoking	185	26.4	26.5	88.7
	2 Still Smoking	79	11.3	11.3	100.0
	Total	697	99.4	100.0	
Missing	System	4	.6		
Total		701	100.0		

Statistics

		TOTAL	CONFID	LIFE	IPA1 energy level	IPA2 reaction to pressure	IPA3 characterization of life as a whole
N	Valid	661	681	676	700	695	693
	Missing	40	20	25	1	6	8
Mean		152.7035	62.6975	89.7544	5.01	4.05	5.20
Median		155.0000	64.0000	92.0000	5.00	4.00	5.00
Std. Deviation		28.07608	12.92818	17.08832	1.351	1.705	1.270

AUTHOR COMMENTS
TOTAL = IPPA TOTAL SCORE; PERCENT MISSING: 40/661 = 6.1%
CONFID = IPPA CONFIDENCE SCALE; PERCENT MISSING: 20/681 = 2.9%
LIFE = IPPA LIFE SCALE; PERCENT MISSING: 25/676 = 3.7%

MARITAL marital status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Never Married	230	32.8	33.4	33.4
	2 Married	346	49.4	50.3	83.7
	3 Living with Significant Other	44	6.3	6.4	90.1
	4 Separated	13	1.9	1.9	92.0
	5 Widowed	20	2.9	2.9	94.9
	6 Divorced	35	5.0	5.1	100.0
	Total	688	98.1	100.0	
Missing	System	13	1.9		
Total		701	100.0		

EXERCISE FIGURE 2–1. (*Continued*).

score had only one or two missing data points each? The large amount of missing data in the TOTAL IPPA scale score is due to the default option in the Compute command in SPSS for Windows. Only cases that have scores for each variable in the Compute statement will be used in the procedure. Any case that is missing a data point for a variable in the equation will be dropped, resulting in missing data in the computed scale score. Thus, if the missing data problem is not addressed **before** forming subscale and total scores, the resulting scores will have a fair amount of missing data. That is what happened to TOTAL (IPPA total score) as well as to the CONFID and LIFE IPPA subscale scores.

Although we are getting ahead of ourselves a bit, we can correct this problem in one of two ways. We can replace the missing data in each of the IPA1 through IPA30 items by substituting the mean or median (if the data are markedly skewed) on that variable for the missing data point using the Recode command. Once completed, we recompute the total score for TOTAL and for the CONFID and LIFE subscales. Then, we rerun the frequencies for these variables and compare them to the original frequencies. You should find that these recomputed variables have no missing data.

What we have just described is another way to handle missing data through substituting the mean or median of the distribution for missing data points. Given the relatively small amount of missing data throughout the SURVEY03.SAV dataset, this choice is most likely the best, and the easiest, way to handle the missing data problem. However, you must decide, based on the amount of skewness in the continuous variables, whether to use the mean or the median as the replacement value. For example, what value would you use for the AGE variable, or for the HEALTH variable?

For nominal level variables such as MARITAL (Marital Status) and POLAFF (Political Affiliation), it is usually best to use the modal value if there is only one mode and not a great deal of missing data. If there is more than one mode or lots of missing data, it might be best to consider using another method or methods. If all else fails, you may have to live with the fact that some cases will have some missing data and, in some analyses, you will use only those cases with complete data on the variables of interest.

2. Exercise Figure 2-2 contains a sample table of two categorical variables. Often, variables are combined into one table, but because SPSS for Windows 12.0 presents separate tables that can be copied into a manuscript, we have kept them as they appear in the SPSS output. Tables are used to present data clearly and succinctly. Not all the information is repeated in the text; generally just the highlights are presented. In our text description, we could describe these three variables by stating that over 62% of the sample is employed full time, and almost half list their political affiliation as independent. When choosing a winter vacation, the most popular choice was a beachfront condo in Hawaii, followed by a Caribbean cruise and a chalet in the Swiss Alps: about 86% of the respondents selected one of those choices. A trip to Disney World was the least popular choice.
3. Exercise Figure 2-3 contains a sample table of continuous variables for respondents' age, education, and total scores on the IPPA scale. We created the table using the descriptives program in SPSS. We could describe the table in the text by stating that respondents ranged in age from 15 to 95 years, with a mean age of 38.1 $\pm$ 13 years. They were a well-educated group, with an average of 16.6 $\pm$ 3.5 years of education. On the IPPA scale, in which scores can range from 30 to 210, respondents' actual scores ranged from 51 to 210.
4. Exercise Figure 2-4 contains the box plot that we ran in SPSS for Windows 12.0 by clicking on Graphics, then Box plot. The box encloses the data from the 25th to the 75th percentile (50% of the data) for the IPPA total score (TOTAL) by gender. The median is represented by the horizontal line within the box. In SPSS, extreme outlying values are defined as those that

WORK current work status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Unemployed	105	15.0	15.0	15.0
	1 Part Time	159	22.7	22.7	37.8
	2 Full Time	435	62.1	62.2	100.0
	Total	699	99.7	100.0	
Missing	System	2	.3		
Total		701	100.0		

POLAFF political affiliation

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Republican	137	19.5	19.8	19.8
	2 Democrat	225	32.1	32.5	52.3
	3 Independent	330	47.1	47.7	100.0
	Total	692	98.7	100.0	
Missing	System	9	1.3		
Total		701	100.0		

WINTER choosing a winter vacation

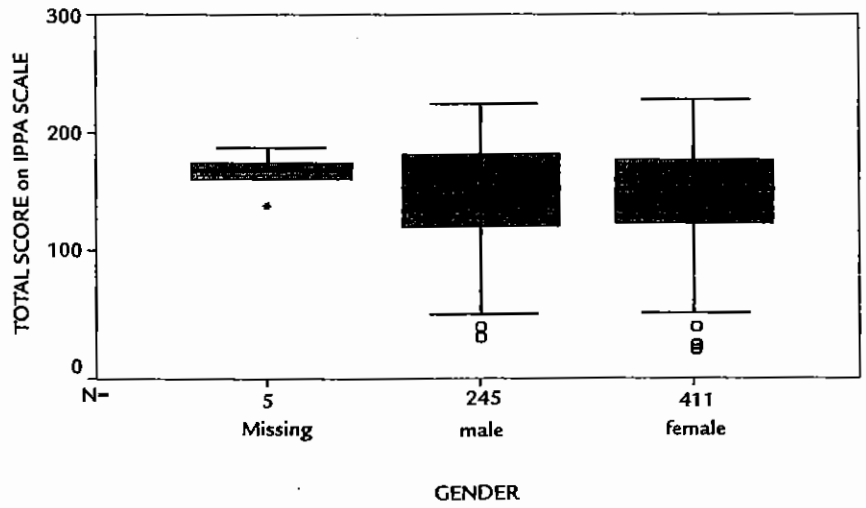
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 beachfront condo in Hawaii	294	41.9	42.2	42.2
	2 chalet in Swiss Alps	145	20.7	20.8	63.0
	3 luxury hotel at Disney World in Florida	97	13.8	13.9	76.9
	4 ocean cruise through Caribbean Islands	161	23.0	23.1	100.0
	Total	697	99.4	100.0	
Missing	System	4	.6		
Total		701	100.0		

EXERCISE FIGURE 2–2. Tables of categorical variables.

Statistics

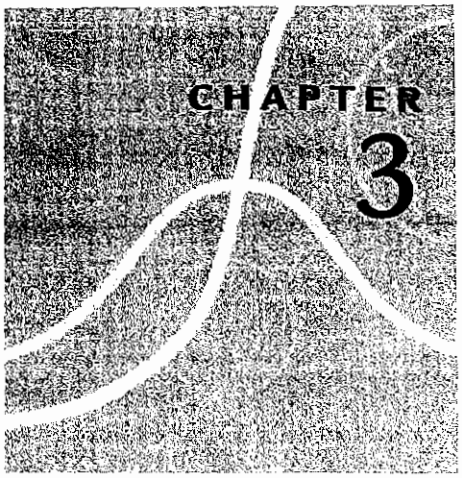
		AGE subject's age	EDUC education in years	TOTAL
N	Valid	684	673	661
	Missing	17	28	40
Mean		38.06	16.60	152.7035
Std. Error of Mean		.496	.133	1.09203
Median		37.00	16.00	155.0000
Mode		28	16	162.00
Std. Deviation		12.972	3.460	28.07608
Variance		168.266	11.973	788.26649
Skewness		.753	.153	-.612
Std. Error of Skewness		.093	.094	.095
Kurtosis		.796	.474	.472
Std. Error of Kurtosis		.187	.188	.190
Range		80	23	159.00
Minimum		15	7	51.00
Maximum		95	30	210.00
Sum		26034	11171	100937.00

EXERCISE FIGURE 2–3. A table describing continuous variables.



EXERCISE FIGURE 2-4. A box plot of the variable IPPA total score by gender.

are more than three box lengths from the upper or lower edge of the box and are designated by asterisks. In this figure, there are no extreme outlying scores. Cases with values between 1.5 and 3 box lengths from the edges of the box are called outliers and are designated by a circle. Although not seen on the box plot, SPSS will print next to the circle the code number(s) of the case(s) with the outlier value. We can then find that specific outlying cases on the variable of interest. The plot helps us determine quickly which subjects are associated with the outlying values.



Key Principles of Statistical Inference

Mary E. Duffy, Barbara Hazard Munro, and
Barbara S. Jacobsen

Objectives for Chapter 3

After reading this chapter, you should be able to do the following:

1. Describe the principles of statistical inference.
2. Describe the characteristics of a normal distribution.
3. Discuss the types of hypothesis testing.
4. Discuss type I and type II statistical errors.
5. Define sensitivity, specificity, predictive value, and efficiency.
6. Discuss tests of significance.
7. Interpret a confidence interval.
8. Examine the components of sample size estimation for study populations.

Statistical inference involves obtaining information from a sample of data about the population from which the sample is drawn and setting up a model to describe this population. For example, the average birth weight of all newborns in a hospital in 2002 (population) can be estimated using observations from a sample of those newborns. Suppose 810 infants were born in the hospital in 2002, and the birth weights of the first 81 newborns (starting January 1) were recorded and averaged. Would the average (mean) birth weight in that sample of 81 be a good estimate of the mean birth weight in the 810 (the population of interest)? It would not be if birth weight depends on time of year or if an effective prenatal nutrition program to improve birth weight had begun in the surrounding community near that time. How can a sample that is representative of that population of 810 be selected? One way is by random selection.

When a *random sample* is drawn from the population of interest, every member of the population has the same probability (chance) of being selected in the sample. If the population is a finite one in which every person in the population can be listed,

a table of random numbers can then be used to select a random sample of any size. Prior to using the World Wide Web (WWW), most people would use a table of random numbers to draw the sample. Now, using the WWW, it is very easy to generate a table of random numbers for many different purposes. Two websites that researchers have used extensively to generate random number tables are *Research Randomizer* (http://www.randomizer.org) and *Random.org* (http://www.random.org). Both sites also permit downloading of randomly generated numbers in a variety of formats, including Microsoft Excel. It is very worthwhile to use one of these sites to accomplish random number generation for research purposes.

Random samples have a high likelihood of being representative of the population from which they were drawn. In contrast, nonprobability samples, or samples nonrandomly selected, are very likely not to represent the populations from which they were selected. Suppose a 10% random sample was selected from the population of 810 newborns. It is very unlikely that the resulting random sample would be the first 81 infants born in 2002.

Random samples are likely to represent the target population because they are based on the principle that each unit in the population has an equal chance of being chosen for the sample. Thus, random samples are considered unbiased in that the process of random sampling produces samples that theoretically represent the population. Most important, the statistical theory on which this book is based assumes random sampling.

Statistical inference is of two types: parameter estimation and hypothesis testing. *Parameter estimation* takes two forms: point estimation and interval estimation. When an estimate of the population parameter is given as a single number, it is called a *point estimate*. The sample mean, median, variance, and standard deviation would all be considered point estimates. Thus, the average birth weight for the random sample of 81 newborns would be a point estimate. In contrast, *interval estimation* of a parameter involves more than one point; it consists of a range of values within which the population parameter is thought to be. A common type of interval estimation is the construction of a confidence interval (CI) and the upper and lower limits of the range of values, called confidence limits. Both point and CI estimates are types of statistical estimates that let us *infer* the true value of an unknown population parameter using information from a *random sample* of that population.

Hypothesis testing, the second and more common type of parameter estimation, will be discussed later in this chapter.

NORMAL CURVE

The *normal curve*, also called the Gaussian curve, is a theoretically perfect frequency polygon in which the mean, median, and mode all coincide in the center, and it takes the form of a symmetrical bell-shaped curve (Fig. 3-1). De Moivre, a French mathematician, developed the notion of the normal curve based on his observations of games of chance. Many human traits, such as intelligence, attitudes,

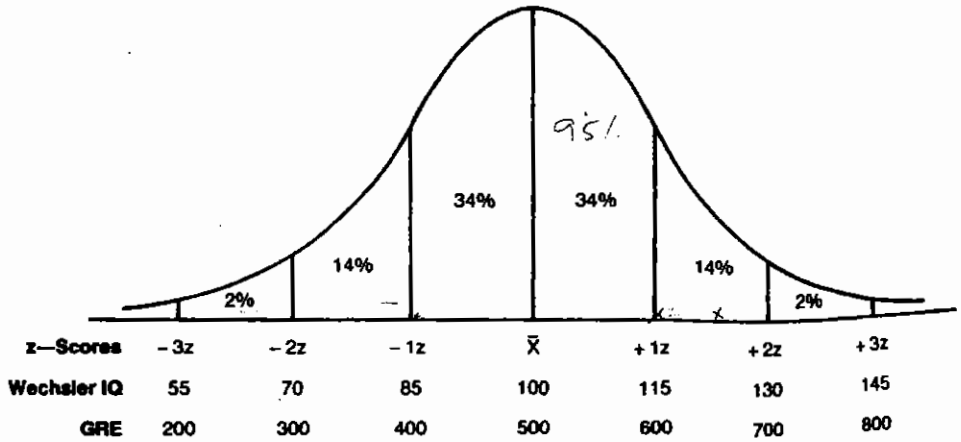


FIGURE 3-1. The normal curve.

and personality, are distributed among the population in a fairly normal way; that is, if you measure something, such as an intelligence test, in a representative sample of sufficient size, the resulting scores will assume a distribution that is similar to the normal curve. Most scores will fall around the mean (an IQ of 100), and there will be relatively few extreme scores, such as an IQ below 55 or above 145.

The normal curve is the most important distribution in statistics for three reasons (Vaughan, 1998). First, although most distributions are not exactly normal, most variables tend to have approximately normal distributions. Second, many inferential statistics assume that the populations are distributed normally. Third, the normal curve is a probability distribution and is used to answer questions about the likelihood of getting various particular outcomes when sampling from a population. For example, when we discuss hypothesis testing, we will talk about the probability (or the likelihood) that a given difference or relationship could have occurred by chance alone. Understanding the normal curve prepares you for understanding the concepts underlying hypothesis testing.

The baseline of the normal curve is measured off in standard deviation (SD) units. These are indicated by the lowercase letter z in Fig. 3-1. A score that is 1 SD above the mean is symbolized by $+1z$, and $-1z$ indicates a score that is 1 SD below the mean. For example, the Wechsler IQ test has a mean of 100 and an SD of 15. Thus, 1 SD above the mean ($+1z$) is determined by adding the SD to the mean ($15 + 100 = 115$), and 1 SD below the mean ($-1z$) is found by subtracting the SD from the mean ($100 - 15 = 85$). A score 2 SD above the mean is $15 + 15 + 100 = 130$; a score 2 SD below the mean is $100 - (15 + 15) = 70$.

When a variable's mean and SD are known, any set of scores can be transformed into z-scores, which have a mean of 0 and an SD of 1. Thus, the z-score tells how many SD a given score is above or below the mean of the distribution. The general formula for converting a score into a z-score is:

$$z = (\text{Score} - \text{M})/\text{SD}$$

However, do not assume that converting variable raw scores to z-scores will result in a normal distribution: A distribution of z-scores has exactly the same distribution as the original distribution. Thus, if the original distribution was positively skewed, the resulting z-score distribution will be positively skewed.

In a normal distribution, approximately 34% of the scores fall between the mean and 1 SD above the mean. Because the curve is symmetrical, 34% also fall between the mean and 1 SD below the mean. Therefore, 68% of scores fall between $-1z$ and $+1z$. With the Wechsler IQ test, this means that 68%, or approximately two thirds of the scores, will fall between 85 and 115. Of the one third of the scores remaining, one sixth will fall below 85, and one sixth will be above 115.

Of the total distribution, 28% fall between 1 and 2 SD from the mean, 14% fall between 1 and 2 SDs above the mean, and 14% fall between 1 and 2 SD below the mean. Thus, 96% of the scores (14 + 34 + 34 + 14) fall between ±2 SD from the mean. For the Wechsler IQ test, this means that 96% of the population receive scores between 70 and 130. Most of the last 4% fall between 2 and 3 SD from the mean, 2% on each side. Thus, 99.7% of those taking the Wechsler IQ test score between 55 and 145.

Two other z-scores are important because we use them when constructing confidence intervals. They are $z = \pm 1.96$ and $z = \pm 2.58$. Of the scores in a distribution, 95% fall between $\pm 1.96z$, and 99% fall between $\pm 2.58z$. For additional practice with the normal curve, look at the Graduate Record Examination (GRE) scores in Fig. 3-1. Each section of the GRE was scaled to have a mean of 500 and an SD of 100. A person who scored 600 on this test would be 1 SD above the mean, or at the 84th percentile. (The 50th percentile is the mean, and 34% above the mean equals the 84th percentile.)

PERCENTILES

In Chapter 2, we pointed out that percentiles allow us to describe a given score in relation to other scores in a distribution. A ***percentile*** tells us the relative position of a given score and allows us to compare scores on tests that have different means and SDs. A percentile is calculated as

$$\frac{\text{number of scores less than a given score}}{\text{total number of scores}} \times 100$$

Suppose you received a score of 90 on a test given to a class of 50 people. Of your classmates, 40 had scores lower than 90. Your percentile rank would be:

$$(40/50) \times 100 = 80$$

You achieved a higher score than 80% of the people who took the test, which also means that almost 20% who took the test did better than you.

As mentioned in Chapter 2, the 25th percentile is called the *first quartile*, the 50th percentile, the *second quartile* or more commonly the *median*; and the 75th

percentile, the *third quartile*. The quartiles are points, not ranges like the interquartile range (IQR). Therefore, the third quartile is not from 50 to 75; it is just the 75th percentile. A score is not said to fall within a quartile, because the quartile is only one point.

As demonstrated with the GRE score of 600, we also can determine percentile rank by using the normal curve. For another example, in Fig. 3-1, the IQ score of 85 exceeds the score of 16% of the population, so a score of 85 is equal to a percentile rank of 16. To test your understanding, determine the percentile rank of a GRE score of 700. Remember that a percentile rank is not a percentile. The *percentile rank* is the *percentage of observations* below a certain score value; a *percentile* is a *score value* below which a certain number of observations in a distribution falls.

Tables make it possible to determine the proportion of the normal curve found between various points along the baseline. They are set up as in Appendix A. To understand how to read the table, go down the first column until you come to 1.0. Note that the percentage of area under the normal curve between the mean and a standard score (z-score) of 1.00 is 34.13. This is how the 34% was determined in Fig. 3-1. Moving down the row to the right, note that the area under the curve between the mean and 1.01 is 34.38, between the mean and 1.02 is 34.61, and so forth.

Suppose you have a standard score of +1.39 (the next section discusses how to calculate the z-scores). Finding this score in the table, we see that the percentage of the curve between the mean and 1.39 is 41.77. A plus z-score is above the mean, so 50% of the curve is on the minus z side, and another 41.77% is between the mean and +1.39; the percentile rank is 91.77 (50 + 41.77). If the z-score were −1.39, the score would fall below the mean, and the percentile rank would be 8.23 (50 − 41.77).

In summary, to calculate a percentile when you have the standard score, you first look up the score in the table (Appendix A) to determine the percentage of the normal curve that falls between the mean and the given score. Then, if the sign is positive, you add the percentage to 50. If the sign is negative, you subtract the percentage from 50.

When using percentiles to determine relative position, it is important to remember the following points:

1. Because so many scores are located near the mean and so few at the ends, the distance along the baseline in terms of percentiles varies a great deal.
2. The distance between the 50th percentile and the 55th percentile is much smaller than the distance between the 90th and the 95th.

What this means in practical terms is that if you raise your score on a test, there will be more impact on your percentile rank if you are near the mean than if you are near the ends of the distribution. For example, suppose three people again took the GRE quantitative examination in hopes of raising their scores and thus their percentile ranks (Table 3-1). All three subjects raised their score by 10 points. For subject 1, who was right at the mean, that meant an increase of 4 points in percentile rank, whereas for subject 3, who was originally 2 SD above the mean, the percentile rank went up only half a point.

TABLE 3-1 *Relationship of Scores to Percentiles at Varying Distances from the Mean*

Subject	*Scores*	*GRE-Q*	*Percentile*
1	1st score	500	50
	2nd score	510	54
2	1st score	600	84
	2nd score	610	86
3	1st score	700	97.7
	2nd score	710	98.2

STANDARD SCORES

Standard scores are a way of expressing a score in terms of its relative distance from the mean. A *z*-score is one such standard score. The meaning of an ordinary score varies depending on the mean and the SD of the distribution from which it was drawn. In research, standard scores are used more often than percentiles. Thus far, we have used examples when the *z*-score was easy to calculate. The GRE score of 600 is 1 SD above the mean, so the *z*-score is +1. The formula used to calculate *z*-scores is:

$$z = \frac{X - M}{SD}$$

The numerator is a measure of the deviation of the score from the mean of the distribution. The following calculation is for the GRE example:

$$z = (600 - 500)/100 = 100/100 = 1$$

As another example, suppose a person obtained a score of 50 on a test in which the mean was 36 and the SD was 4.

$$z = (50 - 36)/4 = 14/4 = 3.5$$

Using the table in Appendix A, we find that 49.98% of the curve is contained between the mean and 3.5 SD above the mean, so the percentile rank for this score would be 99.98 (50 + 49.98).

Suppose the national mean weight for a particular group is 130 pounds, and the SD is 8 pounds. An individual from the group, Jane, weighs 110 pounds. What is Jane's *z*-score and percentile rank?

$$z = (110 - 130)/8 = -20/8 = -2.5$$

Jane's percentile rank is 50 − 49.38, or 0.62.

If all the raw scores in a distribution are converted to *z*-scores, the resulting distribution will have a mean of zero and an SD of 1. If several distributions are converted to *z*-scores, the *z*-scores for the various measures can be compared directly. Although

each new distribution has a new SD and mean (1 and 0), the shape of the distribution is not altered.

Transformed Standard Scores

Because calculating *z*-scores results in decimals and negative numbers, some people prefer to transform them into other distributions. A widely used distribution is one with a mean of 50 and an SD of 10. Such *transformed standard scores* are generally called *T-scores*, although some authors call them *Z*-scores. Some standardized test results are given in *T*-scores. To convert a *z*-score to a *T*-score, use the following formula:

$$T = 10z + 50$$

For example, with a *z*-score of 2.5, the *T*-score would be:

$$T = (10)(2.5) + 50$$
$$T = 25 + 50$$
$$T = 75$$

In the new distribution, the mean is 50 and the SD is 10, so a score of 75 is still 2.5 SD above the mean.

In the same way, other distributions can be established. This is the technique used to transform *z*-scores into GRE scores with a mean of 500 and an SD of 100. The basic formula for transforming *z*-scores is to multiply the *z*-scores by the desired SD and add the desired mean:

$$\text{transformed } z\text{-scores} = (\text{new SD})(z\text{-score}) + (\text{new mean})$$

Suppose you wanted to transform your *z*-scores into a scale with a mean of 70 and an SD of 10. Then your formula would be $10z + 70$. Transforming scores in this way does not change the original distribution of the scores. In some circumstances, however, a researcher may want to change the distribution of a set of data. This might occur when you have a set of data that is not normally distributed.

CORRECTING FAILURES IN NORMALITY THROUGH DATA TRANSFORMATIONS

Many statistical techniques assume that data are normally distributed in the population being studied. Even though many methods will work just as well when this assumption is violated (Glass & Hopkins, 1996), data transformation is often recommended to convert original scores to another metric that approximates normality. Such transformations, however, should be approached with caution because they make interpretation of the results more difficult. The transformed scales are not in the same metric as the original; thus, measures of central tendency and dispersion are not clear in relation to the original measure.

As discussed in Chapter 2, Tabachnick and Fidell (2001) recommend procedures for handling skewness problems. Because the impact of skewness on data analysis, results, and interpretation is often overlooked, this information bears repeating.

First, determine the direction of the deviation. Positive skewness, with the long tail to the right, can be handled in a straightforward manner. However, if a variable has negative skewness, with the long tail to the left, it is best to make it positive by "reflecting" it before transformation. (Reflecting a variable is the same as reverse-coding of all scores in the distribution using a RECODE command so that what was once the lowest score becomes the highest score and so on for all values in the distribution.) This is done as follows:

1. Find the largest value in the distribution and add one to it to form a constant that is larger than any score in the distribution. For example, if the largest score in a distribution is 24, adding one forms a constant of 25 (24 + 1 = 25).
2. Create a new variable by subtracting each score in the distribution from this constant. The new variable, which originally was negatively skewed, now has a positive skewness.

The reflected variable's interpretation changes in the opposite direction as well. If high scores on a variable before it was reflected indicated a large amount of a characteristic, these scores, after reflection, signify a small amount of that characteristic.

Once the direction of the skew has been addressed, use

- A square root transformation for moderate skewness,
- A log transformation for severe skewness,
- An inverse transformation for very severe, or J-shaped, skewness (Tabachnick & Fidell, 2001).

After each attempt at correcting the skew, recalculate the measure of skewness to determine whether the variable is normally, or nearly normally, distributed after transformation. If the transformed variable has a more normal distribution, then use it in subsequent data analyses. Report your use of the transformed variable in subsequent tables and in the narrative of the research report. If transformations are not successful, consider creating a categorical (nominal) variable in place of the continuous variable.

CENTRAL LIMIT THEOREM

If you draw a sample from a population and calculate its mean, how close have you come to knowing the mean of the population? Statisticians have provided us with formulas that allow us to determine just how close the mean of our sample is to the mean of the population.

When many samples are drawn from a population, the means of these samples tend to be normally distributed; that is, when they are charted along a baseline, they tend to form the normal curve. The larger the number of samples, the more the distribution approximates the normal curve. Also, if the average of the means of the samples is calculated (the mean of the means), this average (or mean) is very close to the actual mean of the population. Again, the larger the number of samples, the closer this overall mean is to the population mean.

If the means form a normal distribution, we can then use the percentages under the normal curve to determine the probability statements about individual means. For example, we would know that the probability of a given mean falling between +1 and −1 SD from the mean of the population is 68%.

To calculate the standard scores necessary to determine position under the normal curve, we need to know the SD of the distribution. You could calculate the SD of the distribution of means by treating each mean as a raw score and applying the regular formula. This new SD of the means is called the *standard error of the mean.* The term *error* is used to indicate the fact that due to sampling error, each sample mean is likely to deviate somewhat from the true population mean.

Fortunately, statisticians have used these techniques on samples drawn from known populations and have demonstrated relationships that allow us to estimate the mean and SD of a population given the data from only one sample. They have established that there is a constant relationship between the SD of a distribution of sample means (the standard error of the mean), the SD of the population from which the samples were drawn, and the size of the samples. We do not usually know the SD of the population. If we had measured the whole population, we would have no need to infer its parameters from measures taken from samples. The formula for the standard error of the mean can be written as:

$$\frac{\text{standard deviation}}{\text{square root of } n}$$

The formula indicates that we are estimating the standard error given the SD of a sample of n size. A sample of 30 (Vaughan, 1998) is enough to estimate the population mean with reasonable accuracy. Given the SD of a sample and the size of the sample, we can estimate the standard error of the mean. For example, given a sample of 100 and an SD of 40, we would estimate the standard error of the mean to be:

$$40/\sqrt{100} = 40/10 = 4$$

Two factors influence the standard error of the mean: the SD of the sample and the sample size. The sample size has a large impact on the size of the error because the square root of n is used in the denominator. As the size of n increases, the size of the error decreases. Suppose we had the same SD as just demonstrated, but a sample size of 1,000 instead of 100. Now we have $40/\sqrt{1{,}000} = 40/31.62 = 1.26$ a much smaller standard error. This shows that the larger the sample, the less the error. If there is less error, we can estimate more precisely the parameters of the population.

If there is more variability in the sample, the standard error increases. If there is much variability, it is harder to draw a sample that is representative of the population. Given wide variability, we need larger samples. Note the effect of variability (SD) on the standard error of the mean.

$$20/\sqrt{100} = 20/10 = 2$$
$$40/\sqrt{100} = 40/10 = 4$$

As is shown later in this chapter, the standard error of the mean underlies the calculation of the confidence interval.

PROBABILITY

Ideas about probability are of primary importance to health care researchers. The use of data in making decisions is a hallmark of our information world, and probability provides a means for translating observed data into decisions about the nature of our world (Kotz & Stroup, 1983). For example, probability helps us to evaluate the accuracy of a statistic and to test a hypothesis. Thus, research findings in journals often are stated in terms of probabilities and are often communicated to patients using this language. The approach to probability in this chapter is practical, with special attention given to concepts that are important later in this text. Probability also underlies the use of logistic regression, presented in Chapter 13.

In the life of a health care professional, questions about probability frequently occur in connection with a patient's future. For example, suppose patient X's mammogram revealed a cluster of five calcifications with no other signs of breast abnormality. The patient is told she should have a breast biopsy based solely on these X-ray findings. If the patient asks about the probability that the biopsy will reveal a malignancy, health care professionals refer to the literature. Powell, McSweeney, and Wilson (1983) studied 251 patients who underwent a breast biopsy with mammographic calcifications as the only reason for the biopsy. Everyone in the sample had at least five microcalcifications in a well-defined cluster. Cancer was found in 45 of these patients (17.9%). Consequently, the patient might be told that the probability of cancer was 17.9%.

What has really been stated? The health care practitioner has presumably imagined that the names of the 251 patients in the study were placed in a hat, and one name was drawn by chance. The odds of drawing the name of one of the 45 patients with cancer are 45 in 251, or 17.9%. Patient X, of course, is not one of the 251 names in the hat, but the practitioner is thinking along the lines of, "What if she were?" This way of thinking, although hypothetical, is reasonable if patient X is similar to the group of 251 patients in the study. Powell et al. (1983) described the sample of 251 patients as consecutively chosen over a period of 18 years from the practice of one surgeon at one hospital. Although this information implies a fairly broad sample, it does not provide any breakdown by prognostic variables.

An outcome may vary according to membership in a certain subset of a total group. For example, in 1982, a male patient was diagnosed with a rare form of abdominal cancer (Gould, 1985). The patient read that the median mortality was 8 months after diagnosis; therefore, he reasoned that his chances of living longer than 8 months were 50%. On reading further, he decided that his chances of being in that 50% who lived longer were good: He was young, the disease was discovered early, and he was receiving the best treatment. He also realized that the survival distribution was undoubtedly skewed to the right, indicating that some patients lived for years with the disease. Therefore, if he was in the upper 50%, his chances of living a lot longer than 8 months were very good.

Self-Evident Truths (Axioms) about Probabilities

All probabilities are between 0% and 100%, as illustrated in Fig. 3-2. There are no negative probabilities. If the probability of something happening is 0%, then it is

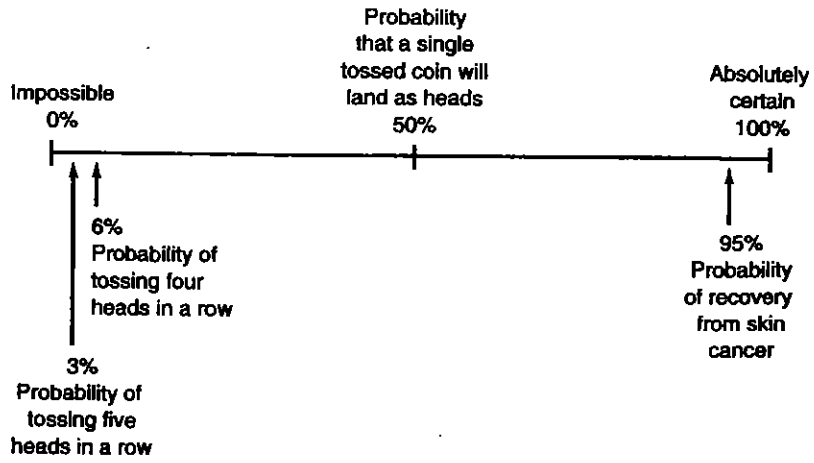


FIGURE 3-2. Diagram of the scale of probabilities.

impossible. Although one must be careful about assigning a probability of 0% to an event, it is highly likely that the probability is 0% that a 98-year-old woman would give birth to a newborn. If the probability of an event is 100%, then we are certain that it will occur. The eventual death of a person has a probability of 100%.

The probability of an event is 100% minus the probability of the opposite event. Perhaps a different health care worker would have preferred to tell patient X there was an 82.1% chance that the breast microcalcifications would *not* be malignant. This would be accurate because 100% − 17.9% = 82.1%.

Table 3-2 lists the four possibilities for the sample of 45 women who were diagnosed with cancer after a biopsy based on a suspicious mammogram, as given by Powell et al. (1983). The sum of all the possibilities for an event is 100%. The sum

TABLE 3-2 *Malignant Pathology of X-Ray Calcifications*

Pathology	***N***
Duct cancer, in situ	25
Lobular cancer, in situ	9
Duct, invasive	9
Lobular, invasive	2
	45
	(100%)

Data from Powell, R. W., McSweeney, M. B., & Wilson, C. E. (1983). X-ray calcifications as the only basis for breast biopsy. *Annals of Surgery, 197,* 555–559. Modified slightly to represent individual patients rather than 47 breasts from 45 patients.

of the four outcomes in Table 3-2 is 100%, indicating that it is certain that one of these possibilities will occur.

Definitions of Probability

FREQUENCY PROBABILITY

Most health care professionals, however, think of probability in the sense of a frequency or statistical probability—that is, they think of probability as a percentage based on empirical observation, which allows them to make an intelligent guess about the future. Their definition for such a probability, based on observations from a sample, is:

$$\text{sample probability} = \frac{\text{number of times the event occurred}}{\text{total number of people in the sample}} \times 100$$

For patient X with the suspicious mammogram findings, the health care worker would substitute as follows:

$$\text{probability of cancer} = (45/251) \times 100 = 17.9\%$$

Patient X was not a member of the group of 251 patients, but the hypothetical type of thinking, "What if patient X were from that group?" is at least reasonable as a practical type of probability. It helps to be able to argue logically that patient X might have been a member of the total group of 251 patients.

In mathematical theory, however, probabilities are meaningful only in the context of chance. We also have to imagine that patient X was chosen "by chance" from the total sample, which implies a random choice. There are two criteria for a random process. First, every item must have an equal chance of being chosen. In the case of drawing from a hat, this means that attention must be given to details, such as whether each name was written on the same-size slip of paper, whether the slips were well mixed, whether the person who drew from the hat was blindfolded, and so forth.[1] Second, each choice must be independent of every other choice. This means that we must not be able to predict whose name will be drawn after patient X.

These criteria for a random process also are important when we consider the larger question of whether the 17.9% probability of cancer would still be the same if more patients were followed. The mathematical definition for a frequency probability invokes the law of averages; that is, we must think of drawing more patients at random. As the sample becomes larger and larger, the percentage will converge to the true or population value.

$$\text{population probability} = \frac{\text{total number of times the event occurred}}{\text{total number of people in the population}} \times 100$$

Thus, the sample probability is an estimate of the population probability. A random sample provides, in theory, a better estimate of the population probability.

[1]Health care researchers who wish to draw a random sample avoid having to deal with such details by using a random number table or Research Randomizer.

In a brief discussion section following the article by Powell et al. (1983), Letton reported on a second sample of 269 patients collected for 10 years. A mammogram indicated calcium deposits, and subsequent biopsies revealed that 46 patients (17.1%) had cancer. Thus, a second study, again with a nonrandom sample, produced remarkably similar results to those of Powell et al.

Random sampling is infrequently used in health care research (Burns & Grove, 2001; Jacobsen & Meininger, 1985; McLaughlin & Marascuilo, 1990). Patients who arrive for care become the sample, and health care researchers take all they can get rather than drawing random samples. These sample probabilities, although not based on a chance process, remain as our only estimates of the true or population probabilities.

Frequency probabilities are based on empirical observations and can thus be termed objective. However, not all probabilities that can be considered objective are determined empirically. For example, when tossing a fair coin, the probabilities of heads or tails can be deduced logically without ever actually tossing the coin. These are called a priori (before the fact), or prior, probabilities. Derdiarian and Lewis (1986) provided an illustration of how a priori probabilities could be used in health care research. Each of three raters was asked to code an item from an interview transcript as belonging to category 1 or category 2. There are eight possible outcomes, as listed in Table 3-3. All three raters could agree that the item belonged in category 1 (1-1-1), or they could disagree, for example, with the first rater coding the item as category 1 and the other two coding the item as category 2 (1-2-2). If all of these outcomes are equally probable by chance, then each will have a probability of 1 in 8. Derdiarian and Lewis (1986) showed how comparing actual results to the tabled probabilities can provide a measure of inter-rater agreement.

TABLE 3-3 *Probability of Eight Possible Outcomes for Three Raters Coding an Item Into Dichotomous Categories (1 or 2) by Chance*

Outcome			
Rater #1	**Rater #2**	**Rater #3**	**Probability**
1	1	1	1/8
1	1	2	1/8
1	2	1	1/8
2	1	1	1/8
1	2	2	1/8
2	1	2	1/8
2	2	1	1/8
2	2	2	1/8

Derdiarian, A. K., & Lewis, S. (1986). The D-L test of agreement: A stronger measure of interrater reliability. *Nursing Research, 35*, 375–378.

SUBJECTIVE PROBABILITY

Another definition for probability is a percentage that expresses our personal, subjective belief that an event will occur. Fisher and van Belle (1993) emphasize that these judgments are not whimsical or irrational; they are based on empiric evidence chosen for some personal reason. In the example of patient X with the suspicious mammogram findings, what is the health care professional's opinion about the probability of 17.9% that the calcifications are cancer? If the patient was told that the probability was close to zero that the biopsy findings would be malignant, then we would be surprised if it turned out to be cancer. On the other hand, most health care professionals would not view a probability of 17.9% as "close to zero." A practitioner would not be very surprised if the calcifications turned out to be cancer, and that is why patient X with five or more calcifications in a cluster was recommended for biopsy.

When testing hypotheses, which is discussed later in this text, researchers focus on probabilities (often called p values) that fall at the lower end of the continuum in Fig. 3-2. Generally, probabilities that are 5% or less are considered unusual in research. The reasons for this are partly intuitive and partly historic. For example, as part of a statistics class, the professor would toss a coin and arrange for it to turn up heads all the time. Intuitively, students begin to laugh and become skeptical after seeing four or five heads in a row. The probability of four heads in a row by chance is approximately 6%, and the probability of five heads in a row is approximately 3%. Note that 5% falls between the two.

The historic reasons for the 5% cutoff are partly based on the preference of Sir Ronald Fisher. Moore (1991) quotes Fisher as writing in 1926 that he preferred the 5% point for marking off the probable from the improbable. Because Fisher was an extremely influential statistician, others adopted this rule too. Moreover, the past inconveniences of calculating have influenced the choice of the 5% mark. Before the computer age, the tables for probabilities for various distributions in textbooks were constructed with handy columns such as 20%, 10%, 5%, and 1%—presumably because we have five fingers and our number system is based on 10. Today, these tables and the use of the 5% level are "almost obsolete" (Freedman et al., 1991, p. 494) because the computer can produce an exact probability based on a mathematical equation. Many researchers and editors of journals, however, persist in using the 5% mark as a cutoff for "unusual" simply because it is convenient to have some general standard that is easy to grasp.

Instead of using the 5% criterion, however, researchers often adopt probability cutoffs that are more generous (eg, 10%) or more strict (eg, 1%) based on their own intuition or the purposes and design of their research. Oftentimes when the researcher is interested in exploring relationships among variables and not hypothesis-testing, a less-stringent cutoff, such as 10% or 20%, might be used. In contrast, a researcher might set the cutoff level at 1% because of testing several hypotheses using one dataset.

At the higher end of the probability continuum, researchers consider probabilities of 95% or more as evidence for reporting potential events that they are confident

will occur. The oft-quoted probability of 95% for recovery from skin cancer is empirically derived, and most health care workers would be surprised if recovery did not occur. Likewise, probabilities near the upper end of the probability scale frequently are used to express confidence in a statistic. For example, a poll reported that 38% of a pre-election random sample favored candidate A. The margin of error was given as 3%, with 95% confidence.

HYPOTHESIS TESTING

Given an underlying theoretical structure, a representative sample, and an appropriate research design, the researcher can test hypotheses. We test to see whether the data support our hypothesis. We do not claim to prove that our hypothesis is true, because one study can never prove anything; it is always possible that some error has distorted the findings.

Null Hypothesis and Alternative Hypothesis

Hypothesis testing is a predominant feature of quantitative health care research. Hypotheses originate from the theory that underpins the research. When a hypothesis relates to the characteristics of a population, such as population parameters, statistical methods can be used with sample data to test its soundness.

There are two types of hypotheses: null and alternative. The ***null hypothesis*** proposes no difference or relationship between the variables of interest. Often written as H_0, the null hypothesis is the foundation of the statistical test. When you statistically test a hypothesis, you assume that H_0 correctly describes the state of affairs between the two variables of interest. If a significant difference or relationship is found, the null hypothesis is rejected; if no difference or relationship is found, H_0 is accepted.

The *alternative hypothesis*, represented by H_a, is a hypothesis that contradicts H_0. The alternative hypothesis can indicate the direction of the difference or relationship that you expect. Thus, the alternative hypothesis is often called the *research hypothesis*, represented by H_r (Agresti & Finlay, 1997).

Types of Error

When we sample, we select cases from a predetermined population. Due to chance variations in choosing the sample's few cases from the population's many possible cases, the sample will deviate from the defined population's true nature by a certain amount. This deviation is called *sampling error*. Thus, inferences from samples to populations are always *probabilistic*, meaning we can never be 100% certain that our inference was correct.

Drawing the wrong conclusion is called an *error of inference*. There are two types of errors of inference, defined in terms of the null hypothesis: *type I* and *type II*.

Before describing these errors, the possibilities related to decisions about the null hypothesis are presented using the following diagram:

	Null Hypothesis	
Decision	*True*	*False*
Accept H_o	OK	Type II
Reject H_o	Type I	OK

If H_o is true and we accept that hypothesis, we have responded correctly. The incorrect response would be to reject a true null hypothesis (type I error). If H_o is false and we reject it, we have responded correctly. The wrong response would be to accept a false null hypothesis (type II error).

Suppose you compared two groups of patients with diabetes taught by different methods (A and B) on how to care for themselves at home, and the data indicated that group A scored significantly higher than group B. You would then reject H_o. Suppose, however, that group A had more diabetics with knowledge about caring for themselves at home and that the method actually did not matter at all. Rejecting the null hypothesis is a type I error.

The probability of making a type I error is called alpha (a) and can be *decreased* by altering the significance level. In other words, you could set the p at 0.01 instead of 0.05; then there is only 1 chance in 100 (1%) that the result termed significant could occur by chance alone. If you do that, however, you will make it more difficult to find a significant result; that is, you will decrease the *power* of the test and increase the risk of a type II error.

A *type II error* is accepting a *false* null hypothesis. If the data showed no significant results, the researcher would accept the null hypothesis. If there were significant differences, a type II error would have been made. To avoid a type II error, you could make the level of significance less extreme. There is a greater chance of finding significant results if you are willing to risk 10 chances in 100 that you are wrong ($p = 0.10$) than there is if you are willing to risk only 5 chances in 100 ($p = 0.05$). Other ways to decrease the likelihood of a type II error are to increase the sample size, decrease sources of extraneous variation, and increase the effect size. The *effect size* is the impact made by the independent variable. For example, if group A scored 10 points higher on the diabetic self-care knowledge scale than group B, the effect size would be 10 divided by the SD of the measure (Cohen, 1988).

There is a trade-off, however, because there is an inverse relationship between type I and type II error. Decreasing the likelihood of a type II error increases the chance of a type I error. If decreasing the probability of one type of error increases the probability of the other type, the question arises: Which type of error are you willing to risk? As you might expect, that depends on the study. An example would be a test for a particular genetic defect. If the defect exists and is diagnosed early, it can be successfully treated; however, if it is not diagnosed and treated, the child will

become severely retarded. On the other hand, if a child is erroneously diagnosed as having the defect and treated, no physical damage is done.

In terms of the types of errors, a type I error would be diagnosing the defect when it does not exist. In that case, the child would be treated but not harmed by the treatment. In contrast, a type II error would be declaring the child to be normal when he or she is not. In that case, irreversible damage would be done. In such a situation, it is obvious that you would make every attempt to avoid the type II error.

Suppose a national study was conducted to determine whether a particular approach to preschool preparation of underprivileged children leads to increased success in school. This approach would cost a great deal of money to implement nationwide. Those responsible for deciding whether to implement this approach would certainly want to be sure that a type I error had not been made. They would not want to institute a costly new program if it did not really have any effect on success in school.

Type I and II errors are hard for some people to grasp, so here are a few examples to help you understand the concept. Let's hypothesize that the two diabetic groups are equal in their knowledge of taking care of themselves. Has an error been made, and if so, what type of error, if the researcher does the following?

1. Accepts the null hypothesis when the groups are really equal in diabetic self-care knowledge.
2. Rejects the null hypothesis when the groups are really equal in diabetic self-care knowledge.
3. Rejects the null hypothesis when the groups are really different in their diabetic self-care knowledge.
4. Accepts the null hypothesis when one group has much more diabetic self-care knowledge than the other.

These four examples summarize the possibilities surrounding these errors. First, if we are given a situation in which H_0 is true—that is, there is no difference—we can either accept it and make the correct decision (#1), or reject it and make an incorrect decision, or a type I error (#2). Second, if H_0 is false, we can reject it, making a correct decision (#3), or accept it and make an incorrect decision, or a type II error (#4).

Sensitivity, Specificity, Predictive Value, and Efficiency

When thinking about types of errors, there is an analogy that can be drawn to diagnostic testing for specific diseases. Clinicians routinely order tests to screen patients for the presence or absence of disease. There are four possible outcomes to diagnosing and testing a particular patient: True Positive (TP), where both diagnosis and test are positive for the disease; True Negative (TN), where both diagnosis and test are negative; False Positive (FP), where the diagnosis is positive and the test is negative for the disease; and False Negative (FN) where the diagnosis is negative for the disease and the test is positive for it (Kraemer, 1992). Essex-Sorlie (1995) notes that a type I error resembles a *false positive outcome,* occurring when a clinical test result incorrectly indicates disease presence. A type II error is comparable to a *false*

negative (FN) *outcome,* indicating a test result incorrectly points to disease absence. The following 2 × 2 table is often used as a way to depict the relationship between the various outcomes.

	Condition Present	***Condition Absent***
Test Positive	True Positive (TP)	False Positive (FP)
Test Negative	False Negative (FN)	True Negative (TN)

The terms used to define the clinical performance of a screening test are sensitivity, specificity, positive predictive value, negative predictive value, and efficiency. Test *sensitivity* (Sn) is defined as the probability that the test is positive when given to a group of patients who have the disease. It is determined by the formula Sn = (TP/(TP + FN)) × 100. In other words, sensitivity can be viewed as, **1 − the false negative rate**, expressed as a percent.

For example, Harvey and colleagues (1992) undertook a study to assess the use of plasma D-dimer levels for diagnosing deep venous thrombosis (DVT) in 105 patients hospitalized for stroke rehabilitation. Plasma samples were drawn from patients within 24 hours of a venous ultrasound screening for DVT. Of the 105 patients in the study, 14 had DVTs identified by ultrasound. The optimal cutoff for predicting DVT was a D-dimer > 1,591 ng/ml. Test results showed the following:

	Positive Ultrasound	***Negative Ultrasound***
d-Dimer > 1,591 ng/ml.	13 (TP)	19 (FP)
d-Dimer ≤ 1,591 ng/ml.	1 (FN)	72 (TN)
	14 (TP + FN)	91 (FP + TN)

Using the above formula, Sn = (TP/(TP + FN)) × 100 = 13/14 × 100 = 93, the sensitivity for the D-dimer test for diagnosing DVTs is 93%. The larger the sensitivity, the more likely the test is to confirm the disease. The D-dimer's test for diagnosing the presence of DVT is accurate 93% of the time.

The *specificity* (Sp) of a screening test is defined as the probability that the test will be negative among patients who do not have the disease. Its formula is Sp = (TN/(TN + FP)) × 100 and can be understood as **1 − the false positive rate**, expressed as a percent.

In the same example, the specificity for the D-dimer test was 79% (Sp = (72/(72 + 19)) × 100 = 72/91 × 100 = 79%). A large Sp means that a positive test can rule out the disease. The D-dimer's specificity of 79% indicates that that test is fairly good in ruling out the presence of DVTs in rehabilitation stroke patients.

The *positive predictive value* (PPV) of a test is the probability that a patient who tested positive for the disease actually has the disease. The formula for PPV is PPV = (TP/(TP + FP)) × 100. Again using the D-dimer test for predicting DVT, its PPV is

calculated as PPV = (13/(13 + 19)) × 100 = 13/32 × 100 = 40.6 or 41%. This means that only 41 out of every 100 screened patients is likely to be a correctly diagnosed and 59 out of 100 are likely to be false positives.

The *negative predictive value* (NPV) of a test is the probability that a patient who tested negative for a disease will not have the disease. It is calculated as NPV = (TN/(TN + FN)) × 100. Using this formula in the above D-dimer test example, NPV = (72/(72 + 1)) × 100 = 72/73 × 100 = 98.6 or 99%. This value indicates that 99 out of 100 patients screened are likely to be true negatives. Thus, the D-dimer test is outstanding at ruling out DVTs in rehabilitation stroke patients who test negative for their presence.

The *efficiency* (EFF) of a test is the probability that the test result and the diagnosis agree (Kraemer, 1992) and is calculated as EFF = ((TP + TN)/(TP + TN + FP + FN)) × 100. In the D-dimer test example, EFF = ((13 + 72)/(13 + 72 + 19 + 1)) × 100 = 85/105 × 100 = 80.9%. Thus, the efficiency of this test in diagnosing rehabilitation stroke patients with DVTs is almost 81%.

SUMMARY

Sensitivity, specificity, predictive values, and efficiency of outcome measures are often reported in health care research studies. Sensitivity depends solely on how positive and negative test results are distributed within a diseased population whereas specificity depends only on how results are distributed in a nondiseased population. Positive predictive values are related to sensitivity and negative predictive values are associated with specificity. Efficiency is the overall accuracy of the test in measuring true findings divided by all of the test results.

In addition to the above calculations, clinical researchers may compute likelihood ratios and relative risks (discussed in later chapters in this book) and receiver operator characteristic (ROC) curve analysis, which graphically portrays a series of sensitivities and specificities for a given test. Kraemer (1992) provides a full treatment of ROC curve analysis. An excellent example of the use of ROC curve analysis in instrument validation can be found in Curley et al. (2003).

Significance Level (*p* Value)

In significance testing, we evaluate differences between what we expect on the basis of our hypothesis and what we observe, but only in relation to one criterion, the probability (p) that these differences could have happened by chance (Elwood, 1998; Henkel, 1986). Chance, or random, factors are those associated with the manner in which the observations used to test the hypothesis were chosen.

In significance testing, we have these two assumptions: H_0 is true and only chance factors could produce results different from what was hypothesized. We then obtain a distribution of possible outcomes, their relative frequency of occurrence, and the likelihood (or probability) that any particular observation would occur. The *p value* is the chief reported result of a significance test and enables us to judge the extent of the evidence against H_0. The p value, which ranges from 0.00 to 1.0, summarizes the evidence in the data about H_0. A large p value, such as 0.53 or 0.78, indicates that the observed data would not be unusual if H_0 were true. A small p value, such as 0.001, denotes that these data would be very doubtful if H_0 were true. This provides strong evidence

against H_o. In such instances, results are said to be *significant at the 0.001 level*, indicating that getting a result of this size might occur only 1 out of 1,000 times.

The alpha level for a statistical test, usually chosen before analyzing data, reflects how careful the researcher wishes to be. The smaller the alpha level, the stronger the evidence must be to reject H_o.

In older studies, hypotheses were usually stated in null form; however, this is not done as often today. When you hypothesize, you state that you believe there is a difference or a relationship between the variables of interest (nondirectional relationship). It is stronger if you state what differences or relationships you expect (directional relationship), rather than write a string of null hypotheses.

It is important to understand the null hypothesis, however, because without it, there is no significance test. Suppose you stated that there is no significant difference between breast-fed and bottle-fed babies in terms of weight gain. If you really had no idea about this issue, it is more common not to state a hypothesis but simply to ask the research question: Is there a difference in weight gain between breast-fed and bottle-fed babies? Even though there is no explicit hypothesis, the null hypothesis (of no difference in weight gain between breast-fed and bottle-fed babies) is implied. If you had rationale for a hypothesis, you might state a directional research hypothesis (H_r) such as: Breast-fed babies gain more weight in the first week of life than bottle-fed babies.

Testing a Statistical Hypothesis

Statistical hypotheses are assumed to be true or false. When we use inferential statistics, we make a decision within a certain margin of error about whether to accept (H_o is true) or reject (H_o is false) the statistical hypothesis. By using the sampling distribution of the test statistic, we compute the probability, labeled p, that the values of the statistic like the one observed would occur if H_o were true.

Testing a statistical hypothesis involves several sequential steps (Glass & Hopkins, 1996; Henkel, 1986):

Step 1. State the statistical hypothesis to be tested; for example, H_o: population mean = 50.
Step 2. Choose the appropriate statistic to test H_o.
Step 3. Define the degree of risk of incorrectly concluding that H_o is false when it is true (type I error). This risk, commonly called alpha, is stated as the probability of a type I error (discussed in the next section). Unless otherwise indicated, alpha ≤ 0.05.
Step 4. Calculate the statistic from a set of randomly sampled observations.
Step 5. Decide whether to reject H_o on the basis of the sample statistic. For example, if p from Step 3 ≤ 0.05, H_o is rejected and we conclude the population mean is not 50. If $p > 0.05$, H_o is not rejected and we conclude the population mean = 50.

Power of a Test

The *power* of a test is the probability of detecting a difference or relationship if such a difference or relationship really exists. Anything that decreases the probability of

a type II error increases power, and vice versa (Vaughan, 1998). A more powerful test is one that is more likely to reject H_0; that is, it is more likely to indicate a statistically significant result when such a difference or relationship exists in the population. The level of significance (probability level) and the power of the test are important factors to consider.

One-Tailed and Two-Tailed Tests

The "tails" refer to the ends of the normal curve. When we test for statistical significance, we want to know whether the difference or relationship is so extreme, so far out in the tail of the distribution, that it is unlikely to have occurred by chance alone. When we hypothesize the direction of the difference or relationship, we state in which tail of the distribution we expect to find the difference or relationship.

Although there is controversy about this, the practice among many researchers is to use a *one-tailed* test of significance when a directional hypothesis is stated and a *two-tailed* test in all other situations. The advantage of using the one-tailed test is that it is more powerful, because the value yielded by the statistical test does not have to be so large to be significant at a given p level. To gain this advantage, however, you must have a sound theoretical basis for the directional hypothesis; you cannot base it on a hunch.

The normal curve is used to demonstrate the difference between one-tailed and two-tailed tests (Fig. 3-3). Recall from our discussion of the normal curve that 95% of the distribution falls between ±1.96 SD from the mean. Thus, only 5% falls beyond these two points: 2.5% of the distribution falls below a z-score of −1.96, and

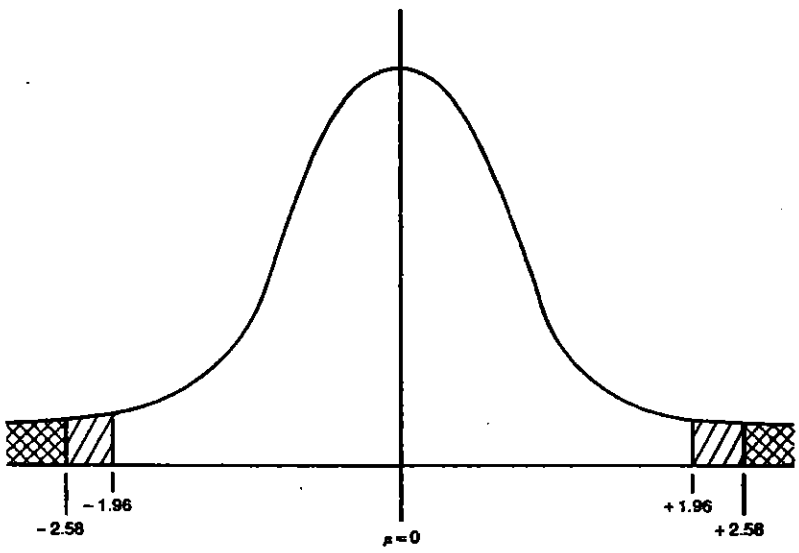


FIGURE 3-3. Two-tailed test of significance using the normal curve.

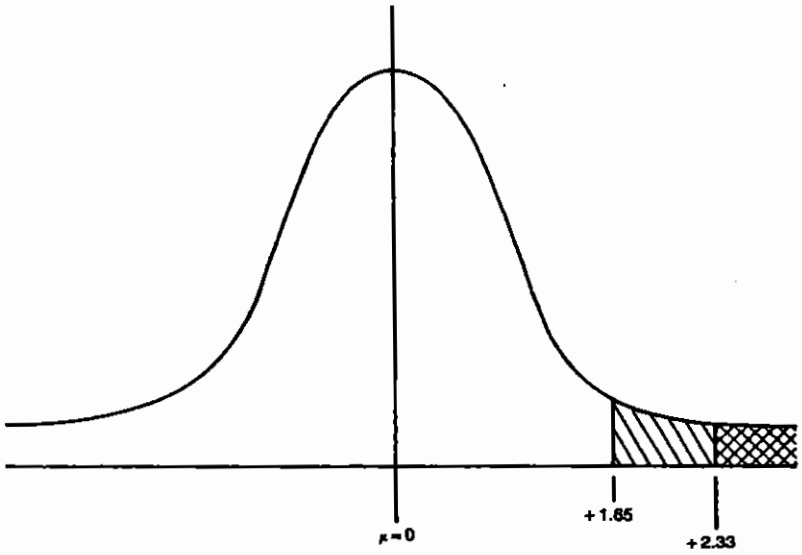


FIGURE 3-4. One-tailed test of significance using the normal curve.

2.5% falls above $+1.96z$. To be so "rare" as to occur only 5% of the time, a z-score would have to be $-1.96z$ or less or $+1.96z$ or greater. Note that we are using both tails of the distribution. Because 99% of the distribution falls between ± 2.58 SD from the mean of the normal curve, a score would have to be -2.58 or less or $+2.58$ or more to be declared significant at the 0.01 level.

Figure 3-4 shows what occurs when a directional hypothesis is stated. We examine only one tail of the distribution. In this example, we look at the positive side of the distribution. Fifty percent of the distribution falls below the mean and 45% falls between the mean and a z-score of $+1.65$ (see Appendix A). Thus, 95% (50 + 45) of the distribution falls below $+1.65z$. To score in the upper 5% would require a score of $+1.65$ or greater. Given a one-tailed test of significance, you would need a score of $+1.65z$ to be significant at the 0.05 level, whereas with a two-tailed test, you needed a score of $\pm 1.96z$. This is an example of the concept of power. With an a priori directional hypothesis, a lower z-score would be considered significant.

For the 0.01 level of significance and a one-tailed test, a z-score of $+2.33$ or greater is needed for significance. This is because 49% of the distribution falls between the mean and $+2.33$, and another 50% falls below the mean.

Degrees of Freedom

The effects of degrees of freedom (*df*) were included in the discussion of the denominator in the computation of the SD. In the sample formula, the denominator is $n - 1$, thus correcting for the possible underestimation of the population

parameter. When describing the calculation of various statistics, we discuss dividing by the *df* and looking up levels of significance in tables using *df*s. Because this is sometimes a confusing concept, a simple example follows.

Degrees of freedom are related to the number of scores, items, or other units in a dataset and to the idea of freedom to vary. Given three scores (1, 5, 6), we have three degrees of freedom, one for each independent item. Each score is free to vary; that is, before collecting the data, we do not know what any of these scores will be. Once we calculate the mean, however, we lose one *df*. The mean of these three scores is four. Once you know the mean and two of the three scores, you can figure out what the third score is; it is no longer free to vary. In calculating the variance or SD, you are calculating how much the scores vary around the sample mean. Because the sample mean is known, one *df* is lost, and the *df*s become $n - 1$, the number of items in the set less one.

Confidence Intervals

When the means (point estimates) are normally distributed, we can use the standard error of the mean to calculate interval estimates. Typically, the 95% and 99% intervals are used. Recall that 95% of the curve is contained between ±1.96 SD from the mean, and that 99% of the curve is contained between ±2.58 SD from the mean. The term *confidence interval* (CI) refers to the degree of confidence, expressed as a percent, that the interval contains the population mean (or proportion), and for which we have an estimate calculated from our sample data (Newton & Rudestam, 1999).

The following formulas are used to calculate the CIs for the population means when the sample size is adequate (generally greater than 30). (For small samples, the *t* distribution may be used to calculate CIs.)

$$95\% = M \pm 1.96 \text{ (standard error)}$$
$$99\% = M \pm 2.58 \text{ (standard error)}$$

The following hypothetical examples are designed to illustrate point estimates and CI estimates derived from a random sample. Suppose that a random sample of 81 newborn infants from a hospital in a poor neighborhood during the last year had a mean birth weight of 100 oz, with an SD of 27 oz.

1. What is the point estimate for the unknown true value of the average (mean) birth weight of all infants born in that hospital in the last year (called the population parameter)?

   Answer: The mean value of 100 oz (computed from the 81 observations) is the best single number estimate (the point estimate) of the unknown value (parameter) for the population of interest. Another random sample of 81 would have given a sample mean different from 100 oz, so the mean value depends on the particular sample that was taken. The difference between the sample mean of 100 oz and the unknown population mean (which it estimates) is the sampling error.

Because the point estimate, 100 oz, is a single number, it gives no indication of its sampling error. CIs computed from random samples enable us to measure sampling error in numeric terms.

2. What is the value of the 95% CI estimate for mean birth weight?

   Answer: First, we must calculate the standard error using the following formula:

$$SD/\sqrt{n}$$

   For our example, this is:

$$27/\sqrt{81} = 27/9 = 3$$

   Next, we calculate the 95% CI:

$$\bar{X} \pm 1.96 \text{ (standard error)}$$
$$100 \pm (1.96)\ (3)$$
$$100 \pm 5.88$$
$$94.12 \text{ and } 105.88$$

   The 95% CI ranges from 94.12 to 105.88. It is a range or interval of estimates for the unknown true value. Thus, a CI consists of an entire interval of estimates for the population parameter.

3. How do we interpret the 95% CI?

   Answer: First, another sample of 81 would almost surely yield a different point estimate. The width of the 95% CI reflects the sampling error resulting from using an estimate based on a random sample of 81 rather than the entire population. In other words, the width of the 95% CI indicates the range of variation for point estimates that may be expected by chance differences from one random sample of the hospital population to another. It is a 95% CI because about 95% of such CIs (obtained from different random samples of that size) will include the true mean value of hospital birth weights. Because that parameter value is usually unknown, we use statistical estimates, the point estimate and the CI estimate, to approximate it. However, if the parameter value (the true mean birth weight for all newborn infants born in that hospital during the last year) were known, approximately 95% of the 95% CIs computed from different random samples of 81 would include that true value.

   The value of the point estimate and the CI estimate depends on the birth weights in the particular sample that was taken, and the estimates will vary from sample to sample. Therefore, we may *not* conclude that the probability is 95% that the mean hospital birth weight is between 94 and 106 oz.

   Either the parameter (the mean of all birth weights in the hospital during 2002) is between 94 and 106 oz or it is not; we do not know which. The 95% denotes the typical accuracy in computing the varying CIs, and not the one CI calculated (Hahn & Meeker, 1991). However, the width of the CI provides useful information about the sampling error or uncertainty of the point estimate unavailable from the point estimate itself. To interpret the specific CI we computed from our sample (here, 95% CI, 94.12 to 105.88), it is necessary to understand the relationship between CIs and significance tests.

Relationship between Confidence Intervals and Significance Tests

To help explain the relationship between CIs and the levels of significance (p values) derived from statistical tests, the following four questions might be asked in relation to our sample mean:

1. Is the mean birth weight in this hospital sample (100 oz) statistically significantly different from 88 oz (5.5 lb, the definition of low birth weight)?
2. Is the mean birth weight in this sample statistically significantly different from 106 oz (6.6 lb, the mean birth weight in that city)?
3. Is the mean birth weight in this sample statistically significantly different from a birth weight of 103 oz?
4. Is the mean birth weight in this sample statistically significantly different from 100 oz, the sample estimate itself?

To test the null hypothesis that there is no statistically significant difference between the mean of 100 oz and each of the other values, we apply the t test. The results follow:

Question	*Null Hypothesis*	*Difference between Values*	p *Value*
1	88	12 oz	0.0006
2	106	6 oz	0.0456
3	103	3 oz	0.3174
4	100	0 oz	1.0000

For the first question, H_o is rejected. The observed mean of 100 oz is statistically significantly higher than the hypothesized value of 88 oz; that is, the 12-oz difference is a significant difference. The p value indicates that a difference that large would occur by chance alone only 6 times in 10,000. H_o is also rejected for question 2; that is, the hospital mean of 100 oz is statistically significantly lower than the city mean of 106 oz. A difference that large would occur by chance alone only 4.6 times in 100 random samples of equal size. For questions 3 and 4, H_o is not rejected; that is, the observed mean of 100 oz is not statistically significantly different from the values of 103 or 100 oz. In the case of the 3-oz difference (question 3), if there really was no difference between the population means, a difference at least that large could be expected to occur by chance in 32% of the random samples. In question 4, the point estimate and H_o are numerically indistinguishable (both 100 oz) and also statistically indistinguishable ($p = 1.0$), because the difference between the two values being compared is zero.

From the chart of the p values, you can see that the further a particular H_o is from the point estimate (100 oz), the lower the p value. In other words, hypotheses become less compatible with the mean of the observed values (here, 100 oz), the larger the difference between the point estimate and the hypothesized or comparison score becomes.

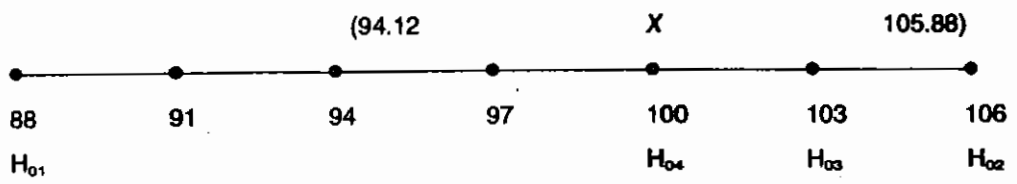


FIGURE 3-5. Relationship of confidence intervals to hypothesis testing.

Figure 3-5 summarizes our results. The null hypotheses are numbered and indicated by H_0. For example, our first H_0 compared our mean of 100 with a value of 88. The CI of 94.12 to 105.88 is included in the figure.

What can be said about the p values for the null hypotheses that fall outside the 95% CI? The two that fall outside of the CI are 88 and 106 from questions 1 and 2, and in both cases the p was less than 0.05. Notice that the second null hypothesis of 106 is just outside the 95% CI, and its p value is barely below 0.05. If H_0 falls at either end of a 95% CI, $p = 0.05$.

Because all numbers outside of the CI have p values less than 0.05, we would expect that all numbers within the CI would have p values greater than 0.05. This leads to a characterization of a 95% CI in terms of p values. A 95% CI contains all the (H_0) values for which $p \geq 0.05$. In other words, a 95% CI contains values (hypotheses) that are statistically compatible (will not be rejected at the 0.05 level) with the point estimate (observed value).

Consistency Checks for Evaluating Research Reports

The relationships between point estimates, CI estimates, and significance tests make it possible to uncover inconsistencies in research reports. The point estimate cannot be outside of the CI. A value for H_0 within the 95% CI should have a p value greater than 0.05, and one outside of the 95% CI should have a p value less than 0.05.

Value of Confidence Intervals

Levels of significance (p values) determine whether a particular hypothesis is statistically compatible with the observed sample value, whereas 95% CIs specify all the population values that are statistically indistinguishable from the observed sample value. Smithson (2003) states that confidence intervals seem clearly superior to the traditional significance testing approach because they display the entire range of hypothetical values of a parameter that cannot be rejected compared to the significance test that focuses solely on one null hypothesis value. In addition, confidence intervals help researchers move toward cumulative knowledge because they enhance comparisons between research replications.

A Word of Caution

Statistical tests and statistical estimates assume random sampling. When using either significance tests or CIs, clear-cut conclusions regarding the entire population apply

only when the study sample is a random sample of that population. Because study patients are rarely random samples from a population, we should be wary about making statistical inferences. If the sample appears to represent some population (but not a random sample), the width of the CI is often viewed as a lower bound (minimum) for the uncertainty in the point estimate. However, clinical judgment must supplement statistical analysis whenever nonrandom samples are used to generalize to individuals not studied (Riegelman, 1981).

When reading a research report, it is essential to determine whether there is an explicitly defined population of interest and whether and how the study sample was selected. Although a representative (nonrandom) sample from an explicitly defined population falls short of a random sample, it is superior to a nonrepresentative sample or to a situation in which the population or the study sample is not clearly defined. When inferences about the population are drawn using statistical tests or CIs in such situations, the reader should beware. Descriptions of the study sample (eg, using point estimates) provide useful information in all situations. When the population is ill defined, the study sample is unrepresentative, or the relation of the study sample to the population is unclear, point estimates and other statistics describing the sample may provide the only reliable information.

Statistical Significance Versus Meaningful Significance

A common mistake in research is to confuse statistical significance with substantive meaningfulness (Ingelfinger et al., 1994; Pedhazur & Schmelkin, 1991). A statistically significant result simply means that if H_0 were true, the observed results would be very unusual. Given a sufficiently large sample (eg, $n \geq 100$), even the tiniest relationship can be statistically significant (Knapp, 1998; Piantadosi, 1997). Statistically significant results tell you nothing about the clinical importance or meaningful significance of the results.

The major task facing the health care researcher is not determining how statistically significant results are, but how meaningful, or substantively important, they are. Because statistical programs are widespread, readily accessible, and easy to use, it is simple to perform tests of statistical significance for various hypotheses. In contrast, it requires a good deal of knowledge and critical thinking skills to determine whether a finding is substantively meaningful. Perhaps this is why researchers still do not refrain from statistical "sanctification" of data (Pedhazur & Schmelkin, 1991; Tukey, 1969), despite numerous writings to this effect.

SAMPLE SIZE DETERMINATION

When planning research, the question always arises as to how large a sample is needed. Determining sample size involves ethical and statistical considerations. If the sample size is too small to detect significant differences or relationships or includes far more subjects than necessary, the cost to subjects and researchers cannot be justified.

In this section, the basic elements of sample size are addressed as they relate to the specific statistics covered in the rest of this book. Jacob Cohen (1988) made a

major contribution to sample size determination. His book provides tables that help us determine the appropriate sample size for a particular statistical test.

Determining the right sample size for a specific study depends on several factors: power, effect size, and significance level. *Power* is defined as the likelihood of rejecting H_0 (ie, avoiding a type II error). An 80% level is generally viewed as an adequate level. *Effect size* is the degree to which H_0 is false; that is, the magnitude of the effect of an independent variable on the dependent variable. This magnitude must be known or estimated in order to determine the minimum sample size needed to achieve a statistical analysis with a power $\geq .80$. For example, in the absence of actual knowledge, for the *t* test, which compares the means of two groups, Cohen (1988) defines a small effect as 0.2 of an SD, a moderate effect as 0.5 SD, and a large effect as 0.8 SD. In relation to GRE scores with an SD of 100, a small effect would be 20 points (100×0.2), a moderate effect 50 points, and a large effect 80 points. The *significance level* is the probability of rejecting a true H_0 (making a type I error); it is called *alpha* and is often set at 0.05.

Given three of these parameters, the fourth can be determined. Cohen's book has both power and sample size tables for most statistical procedures. If we know the sample size, effect size, and significance level, we can determine the power of the analysis. This can be particularly helpful when critiquing research because nonsignificant results may be related to an inadequate sample size, and significant results may be related to a very large sample rather than to a meaningful result.

When planning a study, the desired power, acceptable significance level, and expected effect size are determined, and these three parameters are used to determine the necessary sample size. In addition to Cohen's book, there are many other resources to help you determine appropriate sample sizes for different types of studies and related statistical techniques. These include books by Kraemer and Thiemann (1987), Maxwell and Delaney (1990), and Murphy and Myers (1998). There are also several stand-alone software programs that can be purchased. These include: Power and Precision, developed by Borenstein, Rothstein, and Cohen (1997) and also marketed as SamplePower, 2.0 (SPSS, 2002), and PASS (NCSS, 2002).

There are also many Web-based applications available to assist you with determining sample size. An excellent source of these websites is found at http://members.aol.com/johnp71/javastat.html. Another quick way to locate these websites is to use your favorite search engine, such as Google (http://www.google.com) and type in the terms "power analysis websites." Then, visit the found websites until you find the one calculator that will be most useful in calculating power or sample size. You will find discussions about power and sample size issues relevant to specific statistical tests in subsequent chapters.

SUMMARY

Topics covered in this chapter are basic to understanding the use of the specific statistical techniques contained in subsequent chapters of this book. Please be sure you understand these topics before proceeding.

Application Exercises and Results

Exercises

1. Scores on a particular test are normally distributed with a mean of 70 and an SD of 15. Between what two scores would you expect
   a. 68% of the scores to fall between: _____ and _____?
   b. 96% of the scores to fall between: _____ and _____?
2. In a positively skewed distribution, the "tail" extends toward the _____ (right/left) or toward _____ (higher/lower) scores of the distribution.
3. When raw scores are converted to standard scores, the resulting distribution has a mean equal to _____ and an SD equal to _____.
4. A distribution of scores has a mean of 70 and an SD of 5. The following four scores were drawn from that distribution: 58, 65, 73, and 82.
   a. Transform the raw scores to standard scores and *T*-scores.
   b. Calculate the percentile for each score.
   c. Use the standard scores that you have calculated for the four scores, and transform them into scores from a distribution with a mean of 100 and an SD of 25.
5. Look at your frequencies for the variables AGE and EDUC. Determine whether the variables are significantly skewed. If they are skewed, perform the appropriate transformations and then run descriptives on the new variables to determine whether the transformations were successful.
6. At your hospital, there were 1,500 deliveries last year; 364 of the women had cesarean sections. What is the probability of having a cesarean section at your hospital?
7. You are testing for significant differences between the mean scores of two groups. You set the level of significance at 0.05. If the mean difference is so large that it would occur by chance 1% of the time, would you accept or reject the null hypothesis?
8. When you make a prediction about the direction of mean differences between an experimental and a control group, would you use a one-tailed or a two-tailed test of significance?
9. Which is the more powerful test, one tailed or two tailed?
10. You hypothesize that there is no significant difference in weight between infants in newborn nursery A and those in newborn nursery B. In each of the following, determine whether an error has been made and, if so, what type of error.
    a. Infants in newborn nursery A really weigh significantly more than infants in newborn nursery B, and you accept the null hypothesis.
    b. Infants in newborn nursery B really weigh more than infants in newborn nursery A, and you reject the null hypothesis.
    c. Infants in both newborn nurseries really do weigh the same, and you accept the null hypothesis.
    d. Infants in both newborn nurseries really do weigh the same, and you reject the null hypothesis.
11. You have measured 120 subjects on a particular scale. The mean is 75 and the SD is 6.
    a. What is the standard error of the mean?

b. Set up the 95% confidence interval for the mean.
c. Set up the 99% confidence interval for the mean.

12. You are reading a review paper discussing the use of serum ferritin as a diagnostic test for iron deficiency anemia, with the results summarized as follows:

		Anemia Present	*Anemia Absent*	*Total*
Serum Ferritin	+ (Positive)	731	270	1001
Test Result	− (Negative)	78	1500	1578
	Total	809	1770	2579

a. Calculate the sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and efficiency (EFF).
b. Describe the clinical performance of the serum ferritin test as a diagnostic tool.

Results

1. With a mean of 70 and an SD of 15:
   a. 68% of scores = ±1 SD; therefore, 68% fall between 55 and 85.
   b. 96% of scores = ±2 SD; therefore, 96% fall between 40 and 100.
2. In a positively skewed distribution, the tail is to the **right** or **higher** scores of the distribution.
3. A standard score distribution has a **mean of 0** and an **SD of 1**.
4. a. Standard scores and *T*-scores

Raw Score	*Standard Score*	*T-Score*
	$z = (X - X/SD$	$T = 10z + 50$
58	$z = (58 - 70)/5 = -2.4$	$T = (10)(-2.4) + 50 = 26$
65	$z = (65 - 70)/5 = -1.0$	$T = (10)(-1.0) + 50 = 40$
73	$z = (73 - 70)/5 = 0.6$	$T = (10)(0.6) + 50 = 56$
82	$z = (82 - 70)/5 = 2.4$	$T = (10)(2.4) + 50 = 74$

b. Percentiles: Areas between mean and *z*-score (Appendix A)

Raw Score		*z-Score*	*Tabled Values Percentiles*
58	−2.4	49.18	50 − 49.18 = 0.82
65	−1.0	34.13	50 − 34.13 = 15.87
73	0.6	22.57	50 + 22.57 = 72.57
82	2.4	49.18	50 + 49.18 = 99.18

c. New distribution: Transformed z-scores = (new SD)(z) + (new X)
$= 25z + 100$

z-Scores	*Transformed Scores*
−2.4	25(−2.4) + 100 = 40
−1.0	25(−1.0) + 100 = 75
0.6	25(0.6) + 100 = 115
2.4	25(2.4) + 100 = 160

5. Here are the relevant values for the variables AGE and EDUC. To determine the degree of skewness using Fisher's measure of skewness formula, we divide the measure of skewness by its standard error. Values greater than 1.96 are significant at the 0.05 level and values greater than 2.58 are significant at the 0.01 level.

AGE = 0.753/0.093 = 8.096 = 8.10
EDUC = 0.153/0.094 = 1.627 = 1.62

The AGE variable is significantly skewed ($p < 0.01$), but the EDUC variable is not. Therefore, we can leave the EDUC variable alone, but we need to consider what method might be best to transform the AGE variable. We will use three methods for handling skewness that are recommended by Tabachnick and Fidell (2001). Each method requires us to create a new variable from the AGE variable. In the first method, we create the new variable, RECAGE, by using the Recode into a Different Variable approach, where we recode the outlier scores of 78, 79, 82, 83, and 95 to 75, 76, 77, 78, and 79, respectively, to make them closer to the bulk of scores in the distribution. In the second method, we use the Compute command to create a new variable, called SQRTAGE, which consists of the square root of every subject's age. In the third method, we again use the Compute command and create a new variable, called LG10AGE, which is a log transformation of the AGE variable. Here are the commands that were used to create these three new variables:

```
RECODE AGE (78=75)(79=76)(82=77)(83=78)(95=79)(ELSE=Copy)* INTO RECAGE.
COMPUTE SQRTAGE = SQRT(AGE).
COMPUTE LG10AGE = LG10(AGE).
EXECUTE.
```

(*If you don't use the ELSE command to bring over the rest of the scores in this type of Recode, the resulting variable will have only those scores that were created, in this case, an n = 6.)

After creating the three new variables, we check to see what, if any, transformation had corrected the skewness. We do this by computing descriptive statistics for the new variables and then calculating Fisher's measure of skewness.

RECAGE = 0.643/0.093 = 6.91
SQRTAGE = 0.268/0.093 = 2.89
LG10AGE = −0.190/0.093 = −2.04

Despite transformation, both RECAGE and SQRTAGE remained markedly skewed ($p < 0.01$). The log transformation, however, reduced the skewness level considerably but not less than the desired 1.96 SD unit level, indicating that the variable LG10AGE was

almost normally transformed. Even though normal transformation was not achieved completely, we could use the log-transformed age variable LG10AGE in subsequent analyses as long as the sample size is not too small. Skewness tends to be more influential in small samples.

6. $100 \times 364/1{,}500 = 24.3\%$.
7. The researcher would reject H_o.
8. Use a one-tailed test of significance.
9. A one-tailed test is more powerful.
10. a. Type II error
    b. No error made
    c. No error made
    d. Type I error
11. a. S_x = SD/square root of n
    = 6/square root of 120
    = 6/10.95 = 0.547 = 0.55
    b. 95% = $\times \pm 1.96\ s_x$
    = 75 ± (1.96)(0.55)
    = 75 ± 1.08
    = 73.92 to 76.08
    c. 99% = $\times \pm 2.58\ s_x$
    = 75 ± (2.58)(0.55)
    = 75 ± 1.42
    = 73.58 to 76.42
12. a. The sensitivity is calculated as follows:

$$\begin{aligned} Sn &= (TP/(TP + FN)) \times 100 \\ &= (731/(731 + 78)) \times 100 \\ &= 731/809 \times 100 \\ &= 90.4\% \text{ or } 90\% \end{aligned}$$

The specificity is calculated as follows:

$$\begin{aligned} Sp &= (TN/(TN + FP)) \times 100 \\ &= (1500/(1500 + 270)) \times 100 \\ &= 1500/1770 \times 100 \\ &= 84.7\% \text{ or } 85\% \end{aligned}$$

Positive predictive value is calculated as follows:

$$\begin{aligned} PPV &= (TP/(TP + FP)) \times 100 \\ &= (731/(731 + 270)) \times 100 \\ &= 731/1001 \times 100 \\ &= 73.0\% \end{aligned}$$

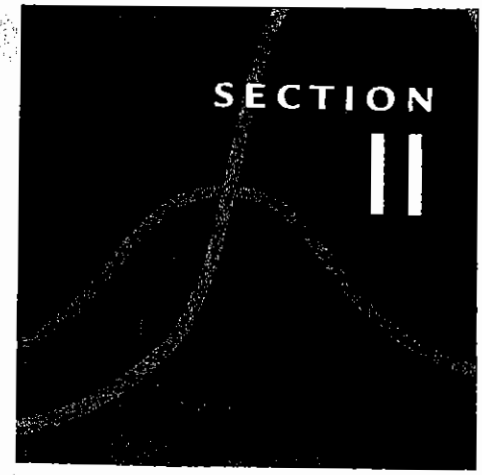
Negative predictive value is calculated as follows:

$$\begin{aligned} NPV &= (TN/(TN + FN)) \times 100 \\ &= (1500/(1500 + 78)) \times 100 \\ &= 1500/1578 \times 100 \\ &= 95.1\% \text{ or } 95\% \end{aligned}$$

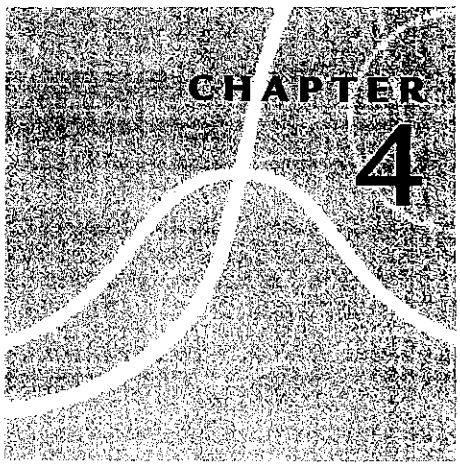
Efficiency is calculated as follows:

$$\begin{aligned}\text{Eff} &= ((\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})) \times 100 \\ &= ((731 + 1500)/(731 + 1500 + 270 + 78)) \times 100 \\ &= 2231/2579 \times 100 \\ &= 86.5\% \text{ or } 86\%\end{aligned}$$

b. These results indicate that 90% of patients with iron deficiency anemia have a positive serum ferritin level test result (Sn), and 85% of patients who do not have the disorder test negative (Sp). Only 27 out of every 100 patients tested will be incorrectly diagnosed with the disorder (PPV) and only 5 out of every 100 patients will be incorrectly classified as not having the disorder when they in fact do have iron deficiency anemia (NPV). The serum ferritin level test is 86% accurate (EFF) in diagnosing patients with iron deficiency anemia.



Specific Statistical Techniques



Selected Nonparametric Techniques

Barbara Hazard Munro

Objectives for Chapter 4

After reading this chapter, you should be able to do the following:

1. Identify situations in which the use of nonparametric techniques is appropriate.
2. Interpret computer printouts containing specified nonparametric analyses.
3. Relate the results of the analysis to the research question posed.

RESEARCH QUESTION

Nonparametric tests can be used to answer research questions ranging from whether a relationship exists between two variables to whether groups differ on an outcome measure. The focus of this chapter is on comparing groups of subjects on outcome measures. The nonparametric techniques to be covered in this chapter are listed in Table 4-1. This chapter is intended to discuss only the commonly used nonparametrics. The parametric analogs covered in later chapters also are included.

TYPE OF DATA REQUIRED

Parametric Versus Nonparametric Tests

When we use *parametric* tests of significance, we are estimating at least one population parameter from our sample statistics. To be able to make such an estimation, we must make certain assumptions; the most important one is that the variable we have measured in the sample is normally distributed in the population to which we plan to generalize our findings. With *nonparametric* tests, there is no assumption

TABLE 4-1 *Nonparametric Tests and Corresponding Parametric Analogs*

	Nonparametric Tests		
	Nominal Data	*Ordinal Data*	***Parametric Analog***
One-group case	Chi-square goodness of fit	—	—
Two-group case	Chi-square	Mann-Whitney U	*t* test
k-group case	Chi-square	Kruskal-Wallis H	One-way ANOVA
Dependent groups (repeated measures)	McNemar test for significance of change	Wilcoxon matched-pairs signed rank test Friedman matched samples	Paired *t* tests Repeated measures ANOVA

about the distribution of the variable in the population. For that reason, nonparametric tests often are called *distribution free.*

At one time, level of measurement was considered a critical element in deciding whether to use parametric or nonparametric tests. It was believed that parametric tests should be reserved for use with interval- and ratio-level data. However, it has been shown that the use of parametric techniques with ordinal data rarely distorts the results.

Parametric techniques have several advantages. Other things being equal, they are more powerful and more flexible than nonparametric techniques. They not only allow the researcher to study the effect of many independent variables on the dependent variable, but they also make possible the study of their interaction. Nonparametric techniques are much easier to calculate by hand than parametric techniques, but that advantage has been eliminated by the use of computers. Small samples and serious distortions of the data should lead one to explore nonparametric techniques. As discussed in Chapter 2, when the data are significantly skewed, thus failing the assumption of normal distribution, one might transform them to achieve a normal distribution. Rather than transform such variables, nonparametric techniques might be used. There are no clear rules for when one approach is preferred. An advantage of the nonparametric approach is that the data retain their original values, thus making interpretation easier. A disadvantage of nonparametrics is their inability to handle multivariate questions.

CHI-SQUARE

Research Question

Chi-square is the most commonly reported nonparametric statistic. It can be used with one or more groups. It compares the actual number (or frequency) in each group with the expected number. The expected number can be based on theory,

experience, or comparison groups. The question is whether the expected number differs significantly from the actual number.

Type of Data Required

Chi-square is used when the data are nominal (categorical). In later chapters we discuss how the chi-square is used to test the fit of models in techniques such as logistic regression and path analysis.

Youngblut and colleagues (2001) compared family characteristics for their two groups of subjects, preterm and full-term deliveries. They used *t* tests to compare the groups on continuous variables and chi-square to compare them on categorical variables. Table 4-2 describes their results. They used an asterisk to indicate significant results. As you can see, none of the chi-square (χ^2) results are significant. The two groups (preterm and full term) did not differ on mother's race, mother's education, family income, mother's employment status, or child's sex. There were differences between the two groups. For example, 98% of the full-term group had family incomes of less than $20,000, whereas only 85% of the preterm group was in that category. This difference, however, was not greater than could occur by chance alone.

Assumptions Underlying Chi-Square

There are four assumptions underlying the chi-square:

1. Frequency data
2. Adequate sample size
3. Measures independent of each other
4. Theoretical basis for the categorization of the variables

The **first assumption** is that the data are frequency data, that is, a count of the number of subjects in each condition under analysis. The chi-square cannot be used to analyze the difference between scores or their means. If data are not categorical, they must be categorized before being used. Whether to categorize depends on the data and the question to be answered.

If the data are not normally distributed and violate the assumptions underlying the appropriate parametric technique, then categorization might be appropriate. The categories developed must adequately represent the data and must be based on sound rationale. If you had the ages of subjects, you could categorize them as 20 to 29, 30 to 39, and 40 to 49. However, you have treated all people within one of your three categories as being equal in age. Does a 29-year-old belong in the same group as a 20-year-old, or is he or she more like a 30-year-old? Specificity and variability are decreased through this categorization, and as a result the analysis will be less powerful.

The question addressed affects the categorization of subjects. Suppose the researcher was interested in whether being in school affects some categorical outcome measure. Then grouping the children as preschool and in school would make

TABLE 4-2 *Comparison of Families with Preterm and Full-Term Preschoolers*

Characteristic	*Preterm M (SD)*	*Full-Term M (SD)*	*Statistic*
Mother's age	29.90 (6.86)	29.20 (6.17)	$t = .58$
Proportion child's life employed	.27 (.37)	.22 (.32)	$t = .78$
Discrepancy	20.80 (12.94)	21.00 (14.59)	$t = .08$
Number of children	2.50 (1.55)	2.50 (1.34)	$t = .13$
Child's age (months)	48.70 (9.92)	48.40 (9.96)	$t = .18$
Birth weight (grams)	1444.10 (527.21)	3331.30 (514.18)	$t = 19.93^*$
Gestational age at birth (weeks)	30.50 (3.17)	39.60 (1.60)	$t = 19.98^*$
Proportion child's life single	.89 (.26)	.88 (.24)	$t = .23$
Mother's race	*N* (%)	*N* (%)	
White	16 (13%)	23 (19%)	$\chi^2 = 4.05$
Black	44 (36%)	36 (30%)	
Hispanic	0 (0%)	2 (2%)	
Mother's education			
<High school	12 (10%)	16 (13%)	$\chi^2 = 1.15$
High school grad	20 (16%)	22 (18%)	
>High school	28 (23%)	23 (19%)	
Family income			
<$20,000	51 (85%)	60 (98%)	$\chi^2 = 5.23$
$20,000–39,999	6 (10%)	1 (2%)	
≥$40,000	3 (5%)	0 (0%)	
Mother's employment status			
Employed	17 (14%)	17 (14%)	$\chi^2 = .003$
Nonemployed	43 (35%)	44 (36%)	
Child's sex			
Female	32 (26%)	25 (21%)	$\chi^2 = 1.85$
Male	28 (23%)	36 (30%)	

$^*p < .01$.

From Youngblut, J. M., Brooten, D., Singer, L. T., Standing, T., Lee, H., and Rodgers, W. L. (2001). Effects of maternal employment and prematurity on child outcomes in single parent families. *Nursing Research, 50*(6), 349.

sense, rather than using their actual ages. When categories have clinical relevance, statistical analyses that preserve these categories are more likely to provide useful interpretations. They are less likely to provide "differences that do not make a difference."

The **second assumption** is that the sample size is adequate. In cross-tabulation procedures, cells are formed by the combination of measures. None of the cells should be empty. Expected frequencies of less than five in 2×2 tables present

problems. In larger tables, many researchers use the rule of thumb that not more than 20% of the cells should have frequencies of less than five (SPSS, 1999b, p. 67). If the cells do not contain adequate numbers, then the variables should be restructured to have fewer categories. It is very important to look at the frequencies of variables before running analyses to ascertain whether adequate numbers of subjects exist. Even with that, however, low numbers in particular cells may not be obvious until the cross-tabulation is run. Most statistical programs print a warning when cell sizes are inadequate. If the cell sizes are problematic, then the researcher should consider restructuring the variable to have less categories.

The **third assumption** is that the measures are independent of each other. This means that the categories created are mutually exclusive; that is, no subject can be in more than one cell in the design, and no subject can be used more than once. It also means that the response of one subject cannot influence the response of another. This seems relatively straightforward, but difficulties arise in clinical research situations when data are collected for a period of time. If you are testing subjects in a hospital or clinic, you must be sure that a person who is readmitted is not enrolled in the study for a second time. You also must be sure that subjects in one condition are not communicating with subjects in their own or different conditions in such a way that responses are contaminated.

The **fourth assumption** is that there is some theoretical reason for the categories. This ensures that the analysis will be meaningful and prevents "fishing expeditions." The latter would occur if the researcher kept recategorizing subjects, hoping to find some relationship between the variables. Research questions and methods for analysis are established before data collection. Although these may be modified to suit the data actually obtained, the basic theoretical structure remains.

Power

Power must be considered when planning sample size. If you have 40 subjects (10 in each of the four cells in a 2×2 design), set your probability level at 0.05 and expect a moderate effect, your power is only 0.47 (Cohen, 1987, p. 235). You have less than a 50% chance of finding a significant relationship between the two variables. Under the same conditions, a sample of 80 results in a power of .76, and a sample of 90 in a power of .81. After the description of the computer printout, an example of power is given.

Example for Computer Analysis

The research question is whether socioeconomic status (SES) is related to the abuse of women. The data were gathered in an AREA grant funded by NINR (Hawkins et al., 1996). Another way to state the question is whether women with low SES differ from women with high SES in their reports of abuse. The researchers used insurance as one way of measuring SES. They categorized the subjects into those who had private insurance against those who did not. They asked the women whether they had ever been emotionally or physically abused. Using the Hawkins et al. (1996) data

ever emotionally or physically abused * SES as risk factor Cross-tabulation

Count

		SES as risk factor		
		Has private insurance	Medicaid mass health or none	Total
Ever emotionally or physically abused	no	1,011	954	1,965
	yes	104	294	398
Total		1,115	1,248	2,363

FIGURE 4-1. Data for chi-square analysis.

and the SPSS program Crosstabs, we produced Fig. 4-1. All figures associated with this analysis were produced by SPSS for Windows version 12.0. Some have been edited slightly. Author comments have been added and appear in shaded boxes.

First, look at the totals for the columns (SES) and rows (abuse). Overall, there are 2,363 subjects, 1,115 with private insurance and 1,248 without private insurance. Fortunately, the group that reports being abused (n = 398) is much smaller than the group that reports no abuse (n = 1,965). Look closely at the figure. Do you think that SES as measured by insurance is related to abuse? The null hypothesis is that there is no difference between the two abuse groups in frequency of abuse.

Because the subjects are not divided equally between those with or without insurance or between those who have been abused and those who have not, adding percentages to the table is helpful in clarifying the results (Fig. 4-2). This was done by requesting row, column, and total percents. Requesting all of the possible percents results in a "busy" table, so take a moment to get comfortable with the figure. Generally, for publication, one uses the independent variable as the column variable and the dependent variable (outcome measure) as the row variable. Then, just the column percents are enough to interpret the results.

In each cell (box) the top number is the count, the second number is the row percent, the third is the column percent, and the bottom number is the total percent. Look at the top box on the left. The count is 1,011; that is 1,011 subjects had private insurance and said they had not ever been emotionally or physically abused. Abuse is the row variable. Here 51.5% (1,011/1,965) of those who had not been emotionally or physically abused had private insurance. SES is the column variable. In this box we see that 90.7% of those who had private insurance said they had not been abused (1,011/1,115). The bottom number in the box indicates that of all the subjects 42.8% had private insurance and had not been abused (1,011/2,363).

Looking at the totals for the row variable, abuse (right-hand column), we see that 1,965 or 83.2% of the women said they were never abused and 398 or 16.8%

ever emotionally or physically abused * SES as risk factor Cross-tabulation

			SES as risk factor		
			Has private insurance	Medicaid mass health or none	Total
Ever emotionally or physically abused	no	Count	1011	954	1965
		% within ever emotionally or physically abused	51.5%	48.5%	100.0%
		% within SES as risk factor	90.7%	76.4%	83.2%
		% of Total	42.8%	40.4%	83.2%
	yes	Count	104	294	398
		% within ever emotionally or physically abused	26.1%	73.9%	100.0%
		% within SES as risk factor	9.3%	23.6%	16.8%
		% of Total	4.4%	12.4%	16.8%
Total		Count	1115	1248	2363
		% within ever emotionally or physically abused	47.2%	52.8%	100.0%
		% within SES as risk factor	100.0%	100.0%	100.0%
		% of Total	47.2%	52.8%	100.0%

AUTHOR COMMENTS
Within a given cell, the percents are as follows:
Frequency (count)
Row %
Column %
Total %

FIGURE 4-2. Frequencies and all percents.

ever emotionally or physically abused *SES as risk factor Cross-tabulation

			SES as risk factor		
			Has private insurance	Medicaid mass health or none	Total
Ever emotionally or physically abused	no	Count	1,011	954	1,965
		Expected Count	927.2	1037.8	1965.0
	yes	Count	104	294	398
		Expected Count	187.8	210.2	398.0
Total		Count	1,115	1,248	2,363
		Expected Count	1115.0	1248.0	2363.0

FIGURE 4-3. Actual and expected frequencies.

said they had been abused. The totals for the column variable, SES, show that 1,115 women or 47.2% had private insurance and 1,248 or 52.8% did not. If abuse were not related to SES, then we would expect that for each level of SES, the rate of abuse would be the same. For the entire sample the rate of abuse is 16.8%. Thus, if there were no differences between the groups, we would expect that within each insurance group, 16.8% would have been abused and 83.2% would not.

These expectations become the *expected frequencies* in the calculation of the chi-square. For those with private insurance (n = 1,115), the expected frequencies would equal 187.32 for the abused group (1,115 × .168 = 187.32). For those without private insurance, the expected frequencies for abuse would equal 209.664 (1248 × .168 = 209.664). Figure 4-3 contains the actual (observed) and expected frequencies. The slight discrepancies come from rounding errors. Actually the percent of those who were abused is 16.84304. If that number is used instead of 16.8, you will get the same expected counts as in Fig. 4-3.

Compare the observed and expected frequencies. Given a rate of 16.8%, we "expect" that 188 (187.8) of those with private insurance would report abuse, but only 104 actually reported abuse. For those without private insurance, more women reported abuse (294) than expected (210). There is a difference in reported abuse between the two groups. In the insurance group, 9.3% report abuse, whereas in the no private insurance group, 23.6% report abuse (see Fig. 4-4). The statistical test tells us whether or not such a difference could have happened by chance alone.

Computer Output for Chi-Square Analysis

Figure 4-4 contains the computer printout of this analysis. Under chi-square tests, we see four different values, with their degrees of freedom (*df*) and significance

ever emotionally or physically abused * SES as risk factor Cross-tabulation

			SES as risk factor		
			Has private insurance	Medicaid mass health or none	Total
Ever emotionally or physically abused	no	Count	1,011	954	1,965
		% within SES as risk factor	90.7%	76.4%	83.2%
	yes	Count	104	294	398
		% within SES as risk factor	9.3%	23.6%	16.8%
Total		Count	1,115	1,248	2,363
		% within SES as risk factor	100.0%	100.0%	100.0%

AUTHOR COMMENTS
The percents are column percents.

Chi-Square Tests

	Value	*df*	Asymp. sig. (2 sided)	Exact sig. (2 sided)	Exact sig. (2 sided)
Pearson chi-square	85.141(b)	1	.000		
Continuity correction (a)	84.128	1	.000		
Likelihood ratio	88.671	1	.000		
Fisher's exact test				.000	.000
Linear-by-linear association	85.105	1	.000		
N of valid cases	2363				

a Computed only for a 2 × 2 table
b 0 cells (.0%) have expected count less than 5. The minimum expected count is 187.80.

AUTHOR COMMENTS
The Pearson value is the usual chi-square value. The other values are described in the text.

FIGURE 4-4. Computer output of chi-square analysis.

Symmetric Measures

		Value	Approx. sig.
Nominal by nominal	Phi	.190	.000
	Cramer's V	.190	.000
	Contingency Coefficient	.186	.000
N of valid cases		2,363	

Not assuming the null hypothesis.
Using the asymptotic standard error assuming the null hypothesis.

AUTHOR COMMENTS
Symmetric measures are only reported when chi-square is significant.
Phi is a shortcut method of calculating a correlation coefficient that can be used when both variables are dichotomous (have only two levels).
Cramer's V is a modified version of Phi that can be used with larger tables.

FIGURE 4-4. (*Continued*)

levels. The *Pearson value* is what you would get if you did this by hand using the usual formula. It is based on the differences between the observed and expected frequencies. For example, the actual (or observed) number of abused women with private insurance is 104, but the expected number (based on an overall rate of 16.8%) is 187.8. The difference between these two values is 83.8. The chi-square value based on the differences between observed and expected frequencies in each of the four cells in our design is 85.141. There is one *df*, and the significance level is .000 (which is at least less than .001). Therefore, since the significance level is less than .05, we would say that the null hypothesis of no difference in abuse between the two insurance groups has not been supported. There is a significant difference between insured and uninsured women in their reported levels of abuse, with uninsured women reporting significantly more abuse.

Since the differences were quite large, this is probably what you expected.

In Chapter 3, the concept of degrees of freedom is defined as the extent to which values are free to vary given a specific number of subjects and a total score. In chi-square analysis, however, frequencies are used rather than scores. The number of cells that are free to vary depends on the number of cells found in the table. How many cell frequencies would we need to know to derive the others? The answer to that question will be equal to the *df*. Given the row and column totals, we only need to know one cell value in a 2 × 2 table to be able to calculate the rest by simple subtraction. Therefore, only one cell is free to vary; the others are

dependent on that value. The *df* for a 2 × 2 chi-square analysis is always 1, regardless of sample size. The formula for calculating the *df* for any size table in a chi-square analysis is:

$$df = (r - 1)(c - 1)$$

For our 2 × 2 table, this becomes $df = (2 - 1)(2 - 1) = 1$.

The *continuity correction* is often referred to as the Yates' correction. Although nominal data are used to calculate a chi-square, chi-square values have a distribution (see Appendix B). The distribution is continuous, but when the expected frequency in any of the cells in a 2 × 2 table is less than 5, the sampling distribution of chi-square for that analysis may depart substantially from normal (Hinkle, Wiersma, & Jurs, 1998, p. 590). In those cases, the continuity correction is recommended. The correction consists of subtracting .5 from the difference between each pair of observed and expected frequencies. In our example, the difference of 83.8 would be reduced to 83.3 by subtracting .5, which results in an overall lower chi-square value. On the output we see that the Pearson value is 85.141, but with the continuity correction, this drops to 84.128. Thus, applying the correction reduces the power of the analysis.

In our example, the *minimum expected count* is 187.80; therefore, we would report the Pearson result. If the minimum expected count (or frequency) had been less than 5, the continuity correction value or Fisher's exact test should have been reported.

The *likelihood ratio* chi-square is an alternative to the Pearson chi-square used for log-linear models. When the sample is large, the likelihood ratio is very similar to the Pearson (SPSS, 1999b).

Fisher's exact test is an alternative to Pearson's chi-square for the 2 × 2 table. It assumes that the marginal counts remain fixed at the observed values and calculates exact probabilities of obtaining the observed results if the two variables are independent (SPSS, 1999b). It is most useful when sample sizes and expected frequencies are small. If the minimum expected value is less than 5, in a 2 × 2 table, Fisher's exact is more appropriate than Pearson's chi-square.

The *linear-by-linear association chi-square*, although printed when chi-square is requested, is not always appropriate because it is based on the relationship between the two variables as measured by the Pearson correlation coefficient. The Pearson correlation coefficient assumes normally distributed data, and this is not usually the case with nominal data, especially with two dichotomous variables, as in a 2 × 2 table (SPSS, 1999b).

Two measures are listed in the table in Figure 4-4 titled Symmetric Measures. They are Phi and Cramer's V.

Phi is a shortcut used for calculating a correlation coefficient. It can be used when both variables are dichotomous (have only two levels). It is appropriate only when the chi-square value is significant. It is interpreted as a measure of association; that is, in this example, the correlation between these two variables is .190. It allows us to interpret the strength of the relationship. It is most useful with 2 × 2 tables in

TABLE 4-3 *Examples of Power*

	Insured	*Uninsured*	*Difference between Groups*
Null hypothesis	16.8%	16.8%	0%
Actual effect	9.3%	23.6%	14.3%
Small effect	10.0%	20.0%	10.0%
Moderate effect	8.0%	38.0%	30.0%
Large effect	5.0%	55.0%	50.0%

which the values of phi range from 0 to 1. In tables with more cells, the value can be greater than 1, decreasing its usefulness. It is complementary to chi-square because it is less sensitive to sample size. It could be used to compare the strength of the relationship across studies.

Cramer's V is a slightly modified version that can be used with larger tables. Phi is adjusted for the number of rows and columns. Thus, given a significant chi-square, report Phi for 2 × 2 tables and Cramer's V for larger tables.

Example of Power Analysis

Cohen (1987) defines the effect sizes related to the chi-square as small = 0.1, moderate = 0.3, and large = 0.5. Using our example, what do these mean? Table 4-3 demonstrates these effect sizes for our example. The null hypothesis in our example is based on the fact that overall, 16.8% of the women reported abuse. Thus, if the null hypothesis is true, 16.8% of the insured and 16.8% of the uninsured will report abuse. Our actual effect was a 14.3% difference between the groups. By Cohen's definition, a small effect would be a 10% difference between the two groups, such as 10% of the insured women being abused versus 20% of the noninsured. A moderate effect would be a 30% difference between the two groups, and a large effect would be a 50% difference. Look at the table to see examples of what those effects could look like.

Example from the Literature

Champion and colleagues (2001) compared genitourinary symptoms between abused and nonabused women. Table 4-4 contains the results. All of the comparisons are significant. Abused women reported significantly more vaginal discharge, abdominal pain, abnormal menses, and dyspareunia than did nonabused women. Look at the percentages, as well as the p values to see what the effects are. By Cohen's definition, the effects would be considered "small," since most are close to 10%.

TABLE 4-4 *Comparisons of Genitourinary Symptomatology of Abused and Nonabused Women*

Variable	*Abused n = 194*	*Nonabused n = 418*	*p*[a]
Vaginal discharge	71.6%	60.0%	<.01
Abdominal pain	46.9%	38.0%	<.05
Abnormal menses	48.2%	36.4%	<.01
Dyspareunia	21.1%	12.0%	<.01

[a]*p* values from comparisons of abused and nonabused groups using chi-square analysis.
From Champion, J. D., Piper, J., Shain, R. N., Perdue, S. T., & Newton, E. R. (2001). Minority women with sexually transmitted diseases: Sexual abuse and risk for pelvic inflammatory disease. *Research in Nursing & Health, 24*(1), p. 27.

Calculation of Chi-Square

When calculating chi-square, the expected and observed frequencies in each cell are compared. Using the expected and observed frequencies in Fig. 4-3, we demonstrate the use of the chi-square formula. In each cell the expected frequency is subtracted from the observed frequency, and that result is squared and then divided through by the expected frequency. The sum of these calculations is the chi-square.

Chi-Square Formula

$$\sum \frac{(1011 - 927.2)^2}{927.2} + \frac{(954 - 1037.8)^2}{1037.8}$$

$$+ \frac{(104 - 187.8)^2}{187.8} + \frac{(294 - 210.2)^2}{210.2} = 85.141$$

Summary for Chi-Square

Chi-square is the appropriate technique when variables are measured at the nominal level. It may be used with one or more groups. In the *one-group* case comparison, data may be provided from a theoretical perspective, norms, or past experience. Suppose a hospital had a cesarean section rate of 30%. This percentage could be compared with reported rates (locally, regionally, or nationally) through the use of chi-square.

Although only a 2 × 2 design has been used as an example, this *two-group* case with two levels in each group can be extended to larger designs. The groups in a chi-square analysis must be mutually exclusive. However, an adaptation of chi-square is the *McNemar test* for use with repeated measures at the nominal level.

McNemar Test

lumps pretest & lumps posttest

Lumps pretest	Lumps posttest	
	0	1
0	199	100
1	15	112

Test Statistics (b)

	Lumps pretest & lumps posttest
N	426
Chi-square (a)	61.357
Asymp. Sig.	.000

a Continuity Corrected
b McNemar Test

AUTHOR COMMENTS
Row Totals—lumps found on pretest
299 women (199 + 100) found 0 to 3 lumps (0)
127 women (15 + 112) found 4 to 8 lumps (1)
Column Totals—lumps found on posttest
214 women (199 + 15) found 0 to 3 lumps (0)
212 women (100 + 112) found 4 to 8 lumps (1)
Cells
Reflecting no change
199 women found 0 to 3 lumps both times (0)
112 women found 4 to 8 lumps both times (1)
Reflecting change
100 women who found 0 to 3 lumps at pretest, found 4 to 8 at posttest
15 women who found 4 to 8 lumps at pretest, found 0 to 3 at posttest

FIGURE 4-5. Computer output of McNemar test.

NOMINAL-LEVEL DATA, DEPENDENT MEASURES

The McNemar test can be used with two dichotomous measures on the same subjects. It is used to measure change. Figure 4-5 contains an example of a computer printout produced by SPSS for Windows, using data collected by Wood (1997). In this example, we are interested in subjects' ability to identify lumps in models of breasts before and after training. There were 8 lumps in the model. Those who detected 0–3 lumps were scored 0, and those who detected 4–8 lumps were scored 1. Looking at the cells, we see that 311 people did not change in their ability to detect the lumps, 199 scored low both times (0), and 112 scored high (1) both times.

Among those who changed, 100 who scored low on the pretest, scored high on the posttest, whereas only 15 people who scored high on the pretest, scored low on the posttest. Thus, for those who changed their scores, more moved from low to high (100) than from high to low (15). This change is statistically significant at the .000 level. This indicates that the training provided to these women improved their ability to detect lumps in a model of a breast.

Summary for McNemar

The McNemar test uses an adaptation of the chi-square formula to test the direction of change. Only the two cells that include changes are included in the analysis; therefore, $df = 1$.

ORDINAL DATA, INDEPENDENT GROUPS

Two commonly used techniques are the *Mann-Whitney U*, which is used to compare two groups and is thus analogous to the *t* test, and the *Kruskal-Wallis H*, which is used to compare two or more groups and is thus analogous to the parametric technique *analysis of variance*. In these techniques, scores for subjects are converted into ranks, and the analyses compare the mean ranks in each group. Using data collected by Wood (1997), we seek to answer the question, Is type of living quarters related to knowledge about breast self-examination? The three types of living quarters are private home, apartment, and elder housing. The knowledge score was significantly skewed, thus making the nonparametric test appropriate. Figures 4-6 and 4-7 contain the computer printouts.

Kruskal-Wallis Test

Ranks

	Type of living quarters	N	Mean Rank
Knowledge score, time 2	Private home	199	245.35
	Apartment	87	206.43
	Elder housing	141	174.43
	Total	427	

Test Statistics (a, b)

	Knowledge score, time 2
Chi-square	28.240
df	2
Asymp. sig.	.000

a Kruskal Wallis Test
b Grouping variable: Type of living quarters

FIGURE 4-6. Computer output, Kruskal-Wallis.

Mann-Whitney Test

Ranks

	What type of living quarters?	N	Mean rank	Sum of ranks
Knowledge score, time 2	Private home	199	151.66	30180.00
	Apartment	87	124.84	10861.00
	Total	286		

Test Statistics (a)

	Knowledge score, time 2
Mann-Whitney U	7033.000
Wilcoxon W	10861.000
Z	−2.557
Asymp. Sig. (2 tailed)	.011

a Grouping Variable: What type of living quarters?

Mann-Whitney Test

Ranks

	What type of living quarters?	N	Mean rank	Sum of ranks
Knowledge score, time 2	Private home	199	168.74	33579.00
	Elder housing	103	118.19	12174.00
	Total	302		

Test Statistics (a)

	Knowledge score, time 2
Mann-Whitney U	6818.000
Wilcoxon W	12174.000
Z	−4.825
Asymp. Sig. (2 tailed)	.000

a Grouping Variable: What type of living quarters?

FIGURE 4-7. Computer output, Mann-Whitney U.

Mann-Whitney Test

Ranks

	What type of living quarters?	N	Mean rank	Sum of ranks
Knowledge score, time 2	Apartment	87	103.79	9029.50
	Elder housing	103	88.50	9115.50
	Total	190		

Test Statistics (a)

	Knowledge score, time 2
Mann-Whitney U	3759.500
Wilcoxon W	9115.500
Z	−1.926
Asymp. Sig. (2 tailed)	.054

a Grouping Variable: What type of living quarters?

FIGURE 4-7. *(Continued)*

The Kruskal-Wallis test (see Fig. 4-6), with a significance level of .000, indicates that the three groups differ significantly on their knowledge of breast self-examination. Looking at the mean ranks, we can see that the group living in private homes scored highest (245.35), followed by those living in apartments (206.43). Those living in elder housing scored lowest (174.43). While we know that there is an overall difference across the three groups, we do not know if each pairwise comparison is significant.

For pairwise comparisons, we use the Mann-Whitney test (see Fig. 4-7), and make all the possible pairwise comparisons. Because we will be making three pairwise comparisons, we need to consider the chance of a type I error. To protect against that error, we can use a *Bonferroni correction*. This involves dividing the desired level of significance by the number of comparisons we are making (.05/3 = .0167). For a comparison to be considered significant, it must have a significance level of .0167, not 0.05. The first test compares those living in private homes with those living in apartments. The significance level of .011 indicates that these two groups are significantly different from each other. Specifically, those in private homes scored significantly higher on the knowledge test than did those living in apartments. The comparison of those living in private homes with those living in elder housing is also significant at the .000 level; that is, those living in private homes scored significantly higher than those living in elder housing. The third comparison between those in apartments and those in elder housing ($p = .044$) is not significant when we use the Bonferroni correction. Thus, these data indicate that women living in private homes score significantly higher on a test of knowlege of breast self-examination than those living in apartments or elder housing. There is no significant difference between those living in apartments and those living in elder housing.

TABLE 4-5 *Variables Contributing to Significant Differences between Frequent and Infrequent TSE Performers (N = 191)*

Variable	*Mann-Whitney U*	*Z*	*p*
Ethnic background	3082.50	−3.851	.000*
Education	3371.00	−2.449	.014*
Family problems	3570.50	−1.874	.050*
Social support	3390.50	−1.756	.035*

*$p < .05$ (two-tailed test)
From Wynd, C. A. (2002). Testicular self-examination in young adult men. *Journal of Nursing Scholarship, 34*(3), p. 254.

Example from the Literature

The Mann-Whitney test was used by Wynd (2002) to study factors related to the practice of testicular self-examination (TSE) among young adult men. Table 4-5 contains a table from her study. She compared frequent and infrequent TSE performers. Those two groups differed significantly in ethnic background, education, family problems, and social support. Additional analyses indicated that African American and Hispanic men practiced TSE less frequently than men from other ethnic groups. Men without a high school education were less likely to practice TSE. Those who reported more family problems and those who had less social support were less likely to practice TSE.

Summary of Kruskal-Wallis and Mann-Whitney U

The Kruskal-Wallis test is the nonparametric analog of the one-way analysis of variance and the Mann-Whitney U test is the nonparametric analog of the *t* test. They may be used when the data violate the assumptions underlying the parametric tests, especially when the data are not normally distributed.

ORDINAL DATA, DEPENDENT GROUPS

The last two nonparametric techniques to be presented are the *Wilcoxon matched-pairs signed rank test* and the *Friedman matched samples*. The Wilcoxon matched-pairs test is analogous to the parametric paired *t* test, and the Friedman matched samples is analogous to a repeated measures analysis of variance. They are used in within-subjects designs when subjects serve as their own controls or the outcome variables are measured more than once.

We will start with the Friedman to demonstrate once more how initial analysis and posthoc tests might be done using nonparametric techniques. Dr. Robin Wood (1997) tested her subjects on their ability to find lumps in models of breasts. She

Friedman Test

Ranks

	Mean rank
Lumps correct, time 1	2.93
Lumps incorrect, time 1	1.77
Lumps correct, time 2	3.47
Lumps incorrect, time 2	1.83

Test Statistics (a)

N	407
Chi-Square	749.367
df	3
Asymp. Sig.	.000

a Friedman Test

FIGURE 4-8. Computer output, Friedman.

counted the number of correct and the number of incorrect lumps they found at two points in time. These variables were not normally distributed, thus the use of nonparametrics is appropriate.

Each subject has a score on each of these variables. The question is whether the subjects differed significantly in their ability to find correct, versus incorrect lumps, and whether this ability changed over time. Figures 4-8 and 4-9 contain the results. The mean ranks for the three variables are given first. The ranks vary from a high of 3.47 for their ability to find correct lumps at time 2 to a low of 1.77 for the number of incorrect lumps they found at time 1. The chi-square has a significance level of .000. Because the initial analysis is significant, we will conduct comparisons of pairs of ranks. While six pairwise comparisons are possible, only four are of interest. (We are not interested in comparing correct lumps at time 1 with incorrect lumps at time 2 or vice versa) The Wilcoxon matched-pairs is used for the four comparisons, and the Bonferroni correction is .05/4 or .0125.

In the first comparison, the numbers of correct and incorrect lumps detected at time 1 are compared. The subjects found significantly more correct than incorrect lumps at time 1 ($p = .000$). They also found more correct than incorrect lumps at time 2 (second comparison, $p = .000$). In the third comparison we see that they found more correct lumps at time 2 versus time 1, which indicates that the training was effective ($p = .000$). They found more incorrect lumps at time 2 than time 1, but this difference was not statistically significant when the Bonferroni correction is used ($p = .023$).

The superscript letters on the printout can be confusing. Look at the first one in Fig. 4-9. This is saying that in 273 of the cases, the subjects rated the number of incorrect lumps lower than the number of correct lumps. Only five subjects found

Wilcoxon Signed Ranks Test

Ranks

		N	Mean rank	Sum of ranks
Lumps Incorrect, time 1; lumps correct, time 1	Negative ranks	273(a)	141.45	38614.50
	Positive ranks	5(b)	33.30	166.50
	Ties	140(c)		
	Total	418		
Lumps incorrect, time 2; lumps correct, time 2	Negative ranks	330(d)	170.09	56130.00
	Positive ranks	5(e)	30.00	150.00
	Ties	85(f)		
	Total	420		
Lumps correct, time 2; lumps correct, time 1	Negative ranks	51(g)	104.18	5313.00
	Positive ranks	226(h)	146.86	33190.00
	Ties	149(i)		
	Total	426		
Lumps incorrect, time 2; lumps incorrect, time 1	Negative ranks	28(j)	39.45	1104.50
	Positive ranks	50(k)	39.53	1976.50
	Ties	330(l)		
	Total	408		

a Lumps incorrect, time 1 < lumps correct, time 1
b Lumps incorrect, time 1 > lumps correct, time 1
c Lumps incorrect, time 1 = lumps correct, time 1
d Lumps incorrect, time 2 < lumps correct, time 2
e Lumps incorrect, time 2 > lumps correct, time 2
f Lumps incorrect, time 2 = lumps correct, time 2
g Lumps correct, time 2 < lumps correct, time 1
h Lumps correct, time 2 > lumps correct, time 1
i Lumps correct, time 2 = lumps correct, time 1
j Lumps incorrect, time 2 < lumps incorrect, time 1
k Lumps incorrect, time 2 > lumps incorrect, time 1
l Lumps incorrect, time 2 = lumps incorrect, time 1

FIGURE 4-9. Computer output, Wilcoxon.

more incorrect lumps than correct lumps. One hundred and forty subjects found an equal number of correct and incorrect lumps. Take a few minutes to look at the remaining superscript letters to be sure you understand their use.

EXAMPLE FROM THE LITERATURE

Tombes and Gallucci (1993) used subjects as their own controls in a study of the effects of hydrogen peroxide rinses on the normal oral mucosa. There were three

"rinse" groups: normal saline, quarter-strength hydrogen peroxide, and half-strength hydrogen peroxide. The Friedman test was used to compare the groups. In the hydrogen peroxide groups, significant mucosal abnormalities occurred over time.

Summary of Friedman and Wilcoxon

The Friedman and Wilcoxon techniques are the nonparametric analogs of the repeated measures analysis of variance and the paired *t* test.

SUMMARY

A few of the more commonly reported nonparametric techniques have been presented. It is important for investigators to examine their data before analysis to determine which techniques are appropriate.

Application Exercises and Results

Exercises

Conduct the appropriate nonparametric analyses to answer the research questions. Write a description of your results as it might appear in a manuscript.

1. Do men and women differ in their political affiliation?
2. Does current satisfaction with weight differ significantly from satisfaction with weight at age 18? To answer this question, first use the Recode procedure to create two new variables. Recode both SATCURWT and SATWT18 into new variables where the values of $1 - 5 = 0$, and $6 - 10 = 1$. This will create two dichotomous variables. Conduct your analysis on the dichotomous variables.
3. Does smoking status affect quality of life in the past month?
4. Do the respondents to this survey differ significantly on productivity (IPA9), goals (IPA13), or worry about the future (IPA29)?

Results

1. Exercise Fig. 4-1 contains the results of this analysis. We would report that chi-square was used to answer the research question. Men and women differed significantly in their political affiliation ($p = .025$). More men (24.8%) are Republicans than women (16.9%), and more women (35.1%) than men (28.3%) are Democrats. Men (46.9%) and women (48.0%) are fairly evenly represented in the Independent category.
2. Exercise Fig. 4-2 contains the results. McNemar was used to answer the research question. There is a significant difference between ratings of satisfaction with weight currently and at age 18 ($p = .000$). Zero equals a low level of satisfaction and one a high level. Of the 697 individuals included in the analysis, 168 were dissatisfied at both times (rating = 0), and 324 were satisfied at both times. Of those who changed their ratings over time, 157 who

political affiliation * gender Cross-tabulation

			Gender		
			Male	Female	Total
Political affiliation	Republican	Count	63	73	136
		% within gender	24.8%	16.9%	19.8%
	Democrat	Count	72	152	224
		% within gender	28.3%	35.1%	32.6%
	Independent	Count	119	208	327
		% within gender	46.9%	48.0%	47.6%
Total		Count	254	433	687
		% within gender	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	*df*	Asymp. sig. (2 sided)
Pearson chi-square	7.393(a)	2	.025
Likelihood ratio	7.298	2	.026
Linear-by-linear association	2.234	1	.135
N of valid cases	687		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 50.28.

EXERCISE FIGURE 4-1. Results for Exercise 1, chi-square.

were satisfied with their weight at age 18 are no longer satisfied, and 48 people who were not satisfied at age 18 are currently satisfied. Therefore, significantly more people reported satisfaction with their weight at age 18 than with their current weight.

3. Exercise Fig. 4-3 contains the results of the analysis. Kruskal-Wallis was used to answer the research question. Smoking status is significantly related to quality of life in the past month ($p = .030$). To test pairwise differences, Mann-Whitney U was used. Because three comparisons were made, the Bonferroni correction was used (0.05/3); thus a p value of .0167 was considered significant. There was no significant difference in quality of life between those who never smoked and those who quit smoking ($p = .971$). Subjects who never smoked rated their quality of life significantly higher than subjects who were still smoking ($p = .010$). Subjects who had quit smoking did not rate their quality of life significantly higher than those who were still smoking ($p = .018$).

McNemar Test

Satisfaction with current weight recoded & satisfaction with weight at age 18 recoded

Satisfaction with current weight recoded	Satisfaction with weight at age 18 recoded	
	0	1
0	168	157
1	48	324

Test Statistics (b)

	Satisfaction with current weight recoded & satisfaction with weight at age 18 recoded
N	697
Chi-Square (a)	56.898
Asymp. Sig.	.000

a Continuity Corrected
b McNemar Test

EXERCISE FIGURE 4-2. Results for Exercise 2, McNemar.

Kruskal-Wallis Test

Ranks

	Smoking history	N	Mean rank
Quality of life in past month	Never Smoked	432	355.00
	Quit smoking	185	354.51
	Still smoking	78	293.81
	Total	695	

EXERCISE FIGURE 4-3. Results for Exercise 3, Kruskal-Wallis and Mann-Whitney U.

Test Statistics (a, b)

	Quality of life in past month
Chi-Square	7.021
df	2
Asymp. Sig.	.030

a Kruskal Wallis Test
b Grouping variable: Smoking history

Mann-Whitney Test

Ranks

	Smoking history	N	Mean rank	Sum of ranks
Quality of life in past month	Never Smoked	432	309.17	133559.50
	Quit smoking	185	308.61	57093.50
	Total	617		

Test Statistics (a)

	Quality of life in past month
Mann-Whitney U	39888.500
Wilcoxon W	57093.500
Z	−.037
Asymp. Sig. (2-tailed)	.971

b Grouping variable: Smoking history

EXERCISE FIGURE 4-3. (*Continued*)

Mann-Whitney Test

Ranks

	Smoking history	N	Mean rank	Sum of ranks
Quality of life in past month	Never Smoked	432	262.33	113326.50
	Still smoking	78	217.67	16978.50
	Total	510		

Test Statistics (a)

	Quality of life in past month
Mann-Whitney U	13897.500
Wilcoxon W	16978.500
Z	−2.576
Asymp. Sig. (2 tailed)	.010

a Grouping variable: Smoking history

Mann-Whitney Test

Ranks

	Smoking history	N	Mean rank	Sum of ranks
Quality of life in past month	Quit smoking	185	138.90	25696.50
	Still smoking	78	115.63	9019.50
	Total	263		

Test Statistics (a)

	Quality of life in past month
Mann-Whitney U	5938.500
Wilcoxon W	9019.500
Z	−2.374
Asymp. Sig. (2-tailed)	.018

a Grouping variable: Smoking history

EXERCISE FIGURE 4-3. (*Continued*)

Friedman Test

Ranks

	Mean rank
Productivity of life	2.50
Defined goals for life	2.04
Worry about future	1.46

Test Statistics (a)

N	698
Chi-Square	472.592
df	2
Asymp. Sig.	.000

a Friedman Test

Wilcoxon Signed Ranks Test

Ranks

		N	Mean rank	Sum of ranks
Defined goals for life— Productivity of life	Negative ranks	341(a)	240.71	82081.00
	Positive ranks	114(b)	189.99	21659.00
	Ties	245(c)		
	Total	700		
Worry about future— Productivity of life	Negative ranks	531(d)	306.75	162882.00
	Positive ranks	61(e)	207.31	12646.00
	Ties	106(f)		
	Total	698		
Worry about future—Definded goals of life	Negative ranks	418(g)	286.91	119928.00
	Positive ranks	131(h)	237.00	31047.00
	Ties	150(i)		
	Total	699		

a defined goals for life < productivity of life
b defined goals for life > productivity of life
c defined goals for life = productivity of life
d worry about future < productivity of life
e worry about future > productivity of life
f worry about future = productivity of life
g worry about future < defined goals for life
h worry about future > defined goals for life
i worry about future = defined goals for life

EXERCISE FIGURE 4-4. Results of Exercise 4, Friedman and Wilcoxon.

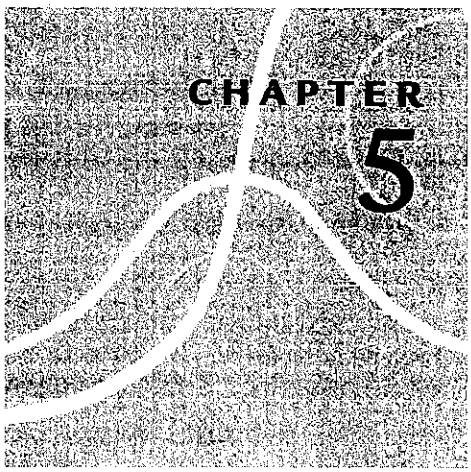
Test Statistics (b)

	Defined goals for life—productivity of life	Worry about future—productivity of life	Worry about future—defined goals for of life
Z	−11.070(a)	−18.205(a)	−12.085(a)
Asymp. Sig. (2-tailed)	.000	.000	.000

a Based on positive ranks
b Wilcoxon Signed Ranks Test

EXERCISE FIGURE 4-4. (*Continued*)

4. Exercise Fig. 4-4 contains the results. Friedman was used to answer the main question. There was an overall significant result ($p = .000$) in the comparison of the following ratings: productivity, goals, and worry about the future. Wilcoxon tests with a Bonferroni correction were conducted to test the pairwise comparisons. A p value of 0.0167 was considered significant. All of the pairwise comparisons were significant. Subjects rated their productivity significantly higher than their goals ($p = .000$) and significantly higher than their worry about the future ($p = .000$). They also rated their goals significantly higher than their worry about the future ($p = .000$).



t Tests: Measuring the Differences Between Group Means

Barbara Hazard Munro

Objectives for Chapter 5

After reading this chapter, you should be able to do the following:

1. Determine when the *t* test is appropriate to use.
2. Discuss how mean difference, group variability, and sample size are related to the statistical significance of the *t* statistic.
3. Discuss how the results of the homogeneity of variance test are related to choice of *t* test formula.
4. Select the appropriate *t* test formula (separate, pooled, or correlated) for a given situation.
5. Interpret computer printouts of *t* test analyses.

Many research projects are designed to test the differences between two groups. The *t* test involves an evaluation of means and distributions of each group. The *t* test, or Student's *t* test, is named after its inventor, William Gosset, who published under the pseudonym of Student. Gosset invented the *t* test as a more precise method of comparing groups. He described a set of distributions of means of randomly drawn samples from a normally distributed population. These distributions are the *t* distributions and are detailed in Appendix C.

The shape of the distributions varies depending on the size of the samples drawn from the populations. However, all the *t* distributions have a normal distribution with a mean equal to the mean of the population. Unlike the z distributions, which are based on the normal curve and estimate the theoretical population parameters, the *t* distributions are based on sample size and vary according to the degrees of freedom (*df*). Theoretically, when an infinite number of samples of equal size are drawn from a normally distributed population, the mean of the sampling distribution will equal

the mean of the population. If the sample sizes were large enough, the shape of the sampling distribution would approximate the normal curve.

RESEARCH QUESTION

When we compare two groups on a particular characteristic, we are asking whether the groups are different. The statistical question asks how different the groups are; that is, is the difference we find greater than that which could occur by chance alone? The null hypothesis for the *t* test states that any difference that occurs between the means of two groups is a difference in the sampling distribution. The means are different not because the groups are drawn from two different theoretical populations, but because of different random distributions of the samples from such a population. The null hypothesis is represented by the *t* distributions constructed by the random sampling of one population. When we use the *t* test to interpret the significance of the difference between groups, we are asking the statistical question, "What is the probability of getting a difference of this magnitude in groups this size if we were comparing random samples drawn from the same population?" In other words, "What is the probability of getting a difference this large by chance alone?"

An example of the *t* test used to compare two groups is the study of Appel, Harrell, and Deng (2002), who compared African American and White southern rural women on physiological variables (Table 5-1). The two groups of women differed significantly on two of the four physiological variables, weight and body mass index (BMI). African American women were significantly heavier and had significantly higher BMIs. The groups of women did not differ on age ($p = .064$) or height ($p = .0931$).

To answer research questions through use of the *t* test, we compare the difference we obtained between our means with the sampling distribution of such differences. In general, the larger the difference between our two means, the more likely it is that the *t* test will be significant. However, two other factors are taken into

TABLE 5-1 *Physiological Variables by Race (n = 1,110)*

	African American (n = 300)			***White (n = 810)***				
Variables	*M*	*(SD)*	*Range*	*M*	*(SD)*	*Range*	*t*	*p*
Age	37.3	(6.9)	24–1	38.2	(6.1)	22–68	−1.85	.64
Weight (kg)	78.5	(17.5)	45.4–158.2	69.0	(15.3)	40.4–168.1	8.05	.001
Height (cm)	163.5	(7.3)	123.1–187.9	164.3	(6.7)	134.6–195.5	−1.6	.0931
BMI (kg/m^2)	29.5	(7.0)	15.6–61.7	25.5	(5.6)	14.2–65.6	8.32	<.0001

Note. M = mean; SD = standard deviation; BMI = body mass index.
From Appel, S. J., Harrell, J. S., & Deng, S. (2002). Racial and socioeconomic differences in risk factors for cardiovascular disease among southern rural women. *Nursing Research, 51*(3), 144.

account: the variability and the sample size. An increase in variability leads to an increase in error, and an increase in sample size leads to a decrease in error.

Given the same mean difference, groups with less variability will be more likely to be significantly different than groups with wide variability. This is because in groups with more variability, the error term will be larger. If the groups have scores that vary widely, there is likely to be considerable overlap between the two groups; thus, it will be difficult to ascertain whether a difference exists. Groups with less variability and a real mean difference will have distributions more clearly distinct from each other; that is, there will be less overlap between their respective distributions. With more variability (thus, larger error), we need a larger difference to be reasonably "sure" that a real difference exists.

TYPE OF DATA REQUIRED

For the *t* test, we need one nominal level variable, with two levels as the independent variable. A simpler way to say this is that we must have two groups. The dependent variable should be continuous.

Some people have criticized the use of the term *continuous* rather than specifying the level of measurement of the variable (ordinal, interval, ratio). However, even when data are measured at the ordinal level, they may be appropriate for use in parametric analyses if they approximate the data required to meet the assumptions of a given analysis. Nunnally and Bernstein (1994) consider any measure that can assume 11 or more dichotomous levels as continuous and state that with multicategory items, "somewhat fewer items are needed to qualify" (p. 570). Scales with fewer items are considered discrete. For ease of expression, we use the term continuous to describe scale scores.

The *t* test has been commonly used to compare two groups. The mathematics involved are simpler than those required for analysis of variance, which is discussed in Chapter 6. However, when comparing two groups, it does not matter whether one uses a *t* test or a one-way analysis of variance: The results will be mathematically identical. The *t* statistic (derived from the *t* test formula) is equal to the square root of the *F* statistic (derived from the one-way analysis of variance), or $t^2 = F$.

With the use of the computer, ease of calculation is not an issue, so some people use analysis of variance to compare two groups. Either way is correct. The typical *t* test table has the advantage of clearly presenting the means being compared in the analysis.

ASSUMPTIONS

The three assumptions underlying the *t* test concern the type of data used in the test and the characteristics of the distribution of the variables:

1. The independent variable is categorical and contains two levels; that is, you have two mutually exclusive groups of subjects. Mutually exclusive means that a

subject can contribute just one score to one of the two groups. This is the assumption of independence. When this assumption is violated, as when subjects are measured twice, a correlated or paired *t* test may be appropriate.
2. The distribution of the dependent variable is normal. If the distribution is seriously skewed, the *t* test may be invalid.
3. The variances of the dependent variable for the two groups are similar. This is related to the assumption implied by the null hypothesis that the groups are from a single population. This assumption is called the *requirement of homogeneity of variance.*

Meeting this last assumption protects against type II errors (incorrectly accepting the null hypothesis). When the variances are unequal—that is, when the variation in one sample is significantly greater than the variation in the other—we are less likely to find a significant *t* value. Therefore, we might incorrectly conclude that the groups were drawn from the same population when they were not.

What if the variances are significantly different? Occasionally, groups that we want to compare do not have equal variances. Fortunately, a statistical method approximates the *t* test and can be interpreted in the same way using a different calculation for the standard error.

Actually, three different formulas based on the *t* distribution can be used to compare two groups:

1. The *basic formula*, sometimes called the *pooled formula*, is used to compare two groups when the three assumptions for the *t* test are met.
2. When the variances are unequal, the *separate formula* is used. This takes into account the fact that the variances are not alike and is a more conservative measure.
3. When the two sets of scores are not independent (assumption 1)—that is, there is correlation between the data taken from the two groups—adjustment must be made for that relationship. That formula often is called the *correlated t test* or the *paired t test.* Comparing a group of subjects on their pretest and posttest scores is an example of when this technique would be used. Because these are not two independent groups, but rather one group measured twice, the scores will most likely be correlated. Another example is when the two groups consist of matched pairs. If the pairs are carefully matched, their scores will correlate, and the standard *t* test would not be appropriate.

When checking the assumptions, the first step is to be sure that each subject contributes only one score to one of the groups. In a randomized clinical trial with angioplasty patients (Sulzbach, Munro, & Hirshfeld, 1995), some patients who had been enrolled in the study in one of the two groups returned to the hospital for a second angioplasty. They could not be reentered into the study without violating the principles underlying the notion of mutual exclusivity.

Next, examine the frequency distribution of the dependent variable. Is it normally distributed? Remember that you divide the skewness by the standard error of the skewness to make this determination. Values that are greater than ± 1.96 are

considered skewed. Alternatives for dealing with skewed data include data transformations, categorizing the variable, or using a nonparametric (distribution free) test. If data transformation is selected and successful, then the *t* test can still be used. If the variable is categorized, then a chi-square would be appropriate. The Mann-Whitney U is an appropriate nonparametric test.

For the test of homogeneity of variance, the analysis is run and the results are examined. The computer produces a test of the assumption, the results of two *t* tests, the pooled or equal variance formula, and the separate or unequal variance formula. An example is given after the discussion of sample size considerations.

SAMPLE SIZE CONSIDERATIONS AND POWER

How many subjects do you need for a *t* test? Cohen (1987) provides tables for determining sample size based on power and effect size determinations, or a computerized program can be used. To enter the tables, we must first decide whether we will be conducting a one- or two-tailed test and what our alpha or probability level will be. If there is sufficient theoretical rationale and we can hypothesize that one group will score significantly higher than the other, we will be using a one-tailed test. If we simply want to answer a question such as, "Is there a difference between the experimental and control groups on the outcome measure?" then we will use a two-tailed test. When planning a study, the sample size is set based on the planned analysis that will require the highest number of subjects. If you were going to run three *t* tests and one would be two tailed, you would base your sample on that, because it requires more subjects than the one-tailed tests.

The power of the test of the null hypothesis is "the probability that it will lead to the rejection of the null hypothesis" (Cohen, 1987, p. 4). A power of 0.80 means, therefore, that there is an 80% chance of rejecting the null hypothesis. The higher the desired power, the more subjects required. Cohen (1987) suggests that for the behavioral scientist, a power of 0.80 is reasonable, given no other basis for selecting the desired level.

The effect size should be based on previous work, if it exists, rather than simply picking a "moderate" effect from the Cohen (1987) tables. The effect size for the *t* test is simply the difference between the means of the two groups divided by the standard deviation for the measure. Cohen's moderate effect size is set at 0.5, which means half of a standard deviation unit. As an example, the graduate record examinations (GRE) have a mean of 500 and a standard deviation of 100. Half of a standard deviation unit on that measure would be 50 (100/2). Thus, a moderate effect would be a difference of 50 points on the GRE between two groups.

In a test of the model of transitional nursing care (Brooten et al., 1995), the LaMonica-Oberst Patient Satisfaction Scale was used. A 17-point difference was found between the experimental and control groups. The standard deviation on the scale was 24. If we were going to use that scale again in a similar experiment, what

TABLE 5-2 n *to Detect d by* t *Test*

	$a_2 = .05$ ($a_1 = .025$)										
Power	*.10*	*.20*	*.30*	*.40*	*.50*	*.60*	*.70*	*.80*	*1.00*	*1.20*	*1.40*
.25	332	84	38	22	14	10	8	6	5	4	3
.50	769	193	86	49	32	22	17	13	9	7	5
.60	981	246	110	62	40	28	21	16	11	8	6
2/3	1144	287	128	73	47	33	24	19	12	9	7
.70	1235	310	138	78	50	35	26	20	13	10	7
.75	1389	348	155	88	57	40	29	23	15	11	8
.80	1571	393	175	99	64	45	33	26	17	12	9
.85	1797	450	201	113	73	51	38	29	19	14	10
.90	2102	526	234	132	85	59	44	34	22	16	12
.95	2600	651	290	163	105	73	54	42	27	19	14
.99	3675	920	409	231	148	103	76	58	38	27	20
	$a_1 = .05$ ($a_2 = .10$)										
	d										
Power	*.10*	*.20*	*.30*	*.40*	*.50*	*.60*	*.70*	*.80*	*1.00*	*1.20*	*1.40*
.25	189	48	21	12	8	6	5	4	3	2	2
.50	542	136	61	35	22	16	12	9	6	5	4
.60	721	181	81	46	30	21	15	12	8	6	5
2/3	862	216	96	55	35	25	18	14	9	7	5
.70	942	236	105	60	38	27	20	15	10	7	6
.75	1076	270	120	68	44	31	23	18	11	8	6
.80	1237	310	138	78	50	35	26	20	13	9	7
.85	1438	360	160	91	58	41	30	23	15	11	8
.90	1713	429	191	108	69	48	36	27	18	13	10
.95	2165	542	241	136	87	61	45	35	22	16	12
.99	3155	789	351	198	127	88	65	50	32	23	17

From Cohen, J. (1987). *Statistical power analysis for the behavior sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Assoc. pp. 54–55.

would our expected effect size be? Divide the difference between the means of 17 by the standard deviation of 24 (17/24), which gives an effect size of .71.

Table 5-2 gives a section of Cohen's tables. The top section has the table for a two-tailed test (a_2) at the .05 level (or a one-tailed test at the .025 level). Given an effect size of .70 (numbers across top of table) and a power of .80 (numbers down

the left side of the table), we would need 33 subjects in each of our groups. If we had used the moderate effect (defined by Cohen as .50), we would need 64 subjects in each of our groups at the same power level. The larger effect size indicates a larger difference between the mean scores and can be detected by fewer subjects.

Now look at the lower section, which includes a one-tailed test at the 0.05 level ($a_1 = .05$). Given an effect size of .70 and a power of .80, we would need 26 subjects per group. Thus, we can see that a one-tailed test is more powerful; that is, we need fewer subjects to detect a significant difference.

To summarize, for sample size with the *t* test, you must determine:

- One tailed versus two tailed
- Alpha level
- Effect size
- Power

You must also estimate how many subjects will be "lost" during data collection and oversample to be sure of having the appropriate numbers for analysis.

COMPUTER ANALYSIS

Dr. Robin Wood (1997) collected data from women in Massachusetts and Georgia. To compare these two groups of women on their years of education, a *t* test was used. Figure 5-1, produced by SPSS for Windows, contains the results. Author comments have been added and appear in a shaded box. The first table contains the group statistics. We see that 99 women were from Massachusetts and 333 from Georgia. The Massachusetts group on average completed 12.6 years of school, whereas the Georgia group averaged 10 years.

The second table, titled Independent Samples Test, contains Levene's test for equality of variances. This is a test of the equality of variance assumption. The test is not significant ($p = .787$), indicating that the variances are equal. We can see in the table of Group Statistics that the standard deviations for the two groups were 4.19946 and 4.19624. The Levene's test tells us that these numbers are equivalent, and that the equal variance or pooled formula *t* test is appropriate.

The computer produces the equal (pooled) variance formula and the unequal (equal variances not assumed or separate) variance formula. Always look first at Levene's test. If the significance level exceeds .05, report the equal variance (pooled) results; if the significance level is less than .05, report the unequal (separate) variance results. In this example, we report the equal variance formula.

The Independent Samples Test table contains both results. Since the variances were equal, we would report a *t* of 5.435, $df = 430$, and $p = .000$. Our analysis indicates that the women from Massachusetts had significantly more years of education (mean = 12.6) than did the women from Georgia (mean = 10.0). The computer printed out the two-tailed significance. For a one-tailed significance, simply divide the *p* value by 2.

T-Test

Group Statistics

	State	N	Mean	Std. deviation	Std. error mean
What is the highest grade or year of school that you completed?	MA	99	12.6111	4.19946	.42206
	GA	333	10.0000	4.19624	.22995

Independent Samples Test

		Levene's test for equality of variances		*t* test for equality of means						
									95% confidence interval of the difference	
		F	Sig.	t	df	Sig. (2 tailed)	Mean difference	Std. error difference	Lower	Upper
What is the highest grade or year of school that you completed?	Equal variances assumed	.073	.787	5.435	430	.000	2.6111	.48044	1.66681	3.55541
	Equal variances not assumed			5.433	160.638	.000	2.6111	.48064	1.66192	3.56030

AUTHOR COMMENTS
If Levene's test for equality of variances is not significant ($p > .05$) report the equal variance results. If Levene's test is significant ($p < .05$) report the separate ("equal variances not assumed") results.
For a one-tailed t test, divide the two-tailed significance by 2.

FIGURE 5-1. Computer output, independent *t* tests.

TABLE 5-3 *Racial Differences in Sample Characteristics by Chi-Square and t Test*

Variable (Categorical)	*Whites (n = 598) N (%)*	*Blacks (n = 44) N (%)*	*p*	*Odds Ratio (95% CI)*
Insured by Medicare, Medicaid, city welfare, or none	76 (12.8)	11 (25.0)	.041	2.27 (1.10, 4.68)
Diabetes	174 (29.1)	23 (52.3)	.001	2.67 (1.44, 4.95)
Chronic renal failure	64 (10.7)	12 (27.3)	.002	3.13 (1.54, 6.38)
Current smoking	180 (31.0)	21 (48.8)	.016	2.12 (1.14, 3.96)
Pulmonary edema	100 (16.7)	17 (38.6)	.001	3.14 (1.65, 5.97)
Nonwhite physician	25 (4.2)	12 (27.3)	.000	8.58 (3.95, 18.62)
Noncardiac physician	192 (32.1)	26 (59.1)	.001	3.05 (1.64, 5.71)
Variable (Continuous)	*Whites (n = 598) Mean (SD)*	*Blacks (n = 44) Mean (SD)*	*p*	
Age	66.99 (12.65)	61.39 (13.40)	.005	

Note. CI = confidence interval.
From Funk, M., Ostfeld, A. M., Chang, V. M. & Lee, F. A. (2002). Racial differences in the use of cardiac procedures in patients with acute myocardial infarction. *Nursing Research, 51*(3), p. 149.

EXAMPLE FROM A PUBLISHED STUDY

Funk and colleagues (2002) studied the racial differences in the use of cardiac procedures in patients with acute myocardial infarctions. First, they compared the two racial groups on various demographic characteristics. They used chi-square for the categorical variables and *t* test for the continuous variable. Table 5-3 contains the results of their comparison of the ages of the two groups. There was a significant difference in age between the Black and Whites in their study ($p = .005$). Whites were significantly older (mean age = 67) than Blacks (mean age = 61).

CORRELATED OR PAIRED *t* TEST

If the two groups being compared are matched or paired on some basis, the scores are likely to be similar. The chance differences between the two groups will not be as large as when they are drawn independently. In the correlated *t* test, a correction is made that has the effect of increasing *t*, thus making it more likely to find a significant difference if one exists.

Figure 5-2 contains a computer printout produced by SPSS for Windows, using data collected by Wood (1997). Subjects were tested on their knowledge of breast self-examination at two points in time. There was a significant change over time with

T-Test

Paired Samples Statistics

		Mean	N	Std. deviation	Std. error mean
Pair 1	Knowledge score time 1	50.0683	439	20.53497	.98008
	Knowledge score time 2	71.0706	439	22.83537	1.08987

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Knowledge score time 1 & % of total Knowledge score, time 2	439	.441	.000

Paired Samples Test

		Paired differences							
					95% confidence interval of the difference				
		Mean	Std. deviation	Std. error mean	Lower	Upper	*t*	*df*	Sig. (2 tailed)
Pair 1	Knowledge score time 1 & % of total Knowledge score, time 2	−21.0023	23.01275	1.09834	−23.1609	−18.8436	−19.122	438	.000

FIGURE 5-2. Computer output, paired *t* test.

subjects scoring significantly higher (p = .000) at the second testing (mean = 71) than at the first testing (mean = 50). The correlation between the two scores, presented in the second table, was .441, significant at the .000 level. The Paired Samples Test table shows that the means differed by 21.0023. The t value of −19.122, with df = 438, has a two-tailed significance of .000 (at least less than .001).

SUMMARY

The t test is a statistical method for comparing differences between two groups. The test requires a continuous dependent variable on which the groups are being compared. The test assumes that the variable is normally distributed in the populations from which the samples are drawn and that the samples have equivalent variances. The t test is particularly useful in experimental and quasi-experimental designs in which an experimental and a control group are compared.

Application Exercises and Results

Exercises

Answer the following research questions by running the appropriate t tests and writing up the results.

1. Do people who have never smoked differ significantly from those who are still smoking on positive psychological attitudes (total IPPA score)?
2. Do current ratings of quality of life differ significantly from ratings of quality of life at age 18?

Results

1. An independent t test should be used to answer this question. On the SMOKE variable, only two groups are selected, those who never smoked (0) and those who are still smoking (2). Exercise Fig. 5-1 contains the results.

   Which t test should be reported? Because Levene's test for equality of variances is significant (p = .000), the equal variances not assumed (separate) formula should be reported. We would report that people who never smoked scored significantly higher (p = .004) on the total IPPA score (mean = 154.67) than did those who are still smoking (mean = 141.35). It should be noted that there was a large difference in the size of the two groups.
2. There was no significant difference (p = .251) between the ratings of quality of life in the past month (mean = 4.27) and at age 18 (mean = 4.22). The correlation between the two measures was .331 (p = .000). Exercise Fig. 5-2 contains the printout.

Test

Group Statistics

	Smoking history	N	Mean	Std. deviation	Std. error mean
OTAL	Never smoked	410	154.6683	26.41131	1.30436
	Still smoking	71	141.3521	36.57013	4.34008

Independent Samples Test

		Levene's test for equality of variances		*t* test for equality of means					
		F	Sig.	*t*	*df*	Sig. (2 tailed)	Mean difference	Std. error difference	95% conf interval differe Lower
TOTAL	Equal variances assumed	14.086	.000	3.683	479	.000	13.3162	3.61540	6.21219
	Equal variances not assumed			2.938	83.100	.004	13.3162	4.53184	4.30268

EXERCISE FIGURE 5-1. Results of Exercise 1, independent samples *t* test.

T-Test

Paired Samples Statistics

		Mean	N	Std. deviation	Std. error mean
Pair 1	Quality of life in past month	4.27	697	1.043	.039
	Quality of life at age 18	4.22	697	1.123	.043

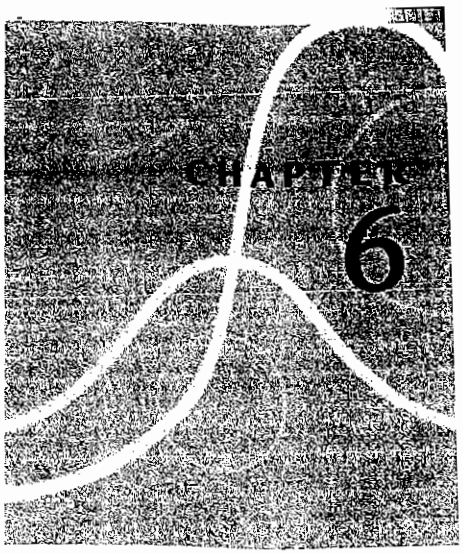
Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Quality of life in past month & quality of life at age 18	697	.331	.000

Paired Samples Test

		Paired differences					t	df	Sig. (2-tailed)
		Mean	Std. deviation	Std. error mean	95% confidence interval of the difference Lower	Upper			
Pair 1	Quality of life in past month—quality of life at age 18	0.5	1.254	0.47	−.04	.15	1.148	696	.251

EXERCISE FIGURE 5-2. Results of Exercise 2, paired samples *t* test.



Differences Among Group Means: One-Way Analysis of Variance

Barbara Hazard Munro

Objectives for Chapter 6

After reading this chapter, you should be able to do the following:

1. Determine when analysis of variance is appropriate to use.
2. Interpret a computer printout of a one-way analysis of variance.
3. Describe between-group, within-group, and total variance.
4. Explain the use of posthoc tests and a priori comparisons.
5. Report the results of one-way analysis of variance in a summary table.

Many times, a clinical research question involves a comparison of several groups on a particular measure. In Chapter 5, we discussed the *t* test as a method for examining the difference between two groups. The basic *t* test compares two means in relation to the distribution of the differences between pairs of means drawn from a random sample. When we have more than two groups and are interested in the differences among the set of groups, we are dealing with different combinations of pairs of means. If we choose to analyze the differences by *t* test analysis, we would need to do a number of *t* tests. Suppose that we had four different groups—A, B, C, and D—that we wanted to compare on a particular variable. If we were interested in the differences among the four groups, we would need to do a *t* test for each of the possible pairs that exist in the four groups. We would have A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D. In all, we would have six separate comparisons, each requiring a separate analysis.

The problem with conducting such multiple-group comparisons relates to the underlying concept of statistical analysis. Each test is based on the probability that the null hypothesis is true. Therefore, each time we conduct a test, we are running the risk of a type I error. The probability level we set as the point at which we reject

the null hypothesis also is the level of risk with which we are comfortable. If that level is 0.05, we are accepting the risk that 5 of 100 times, our rejection of the null hypothesis will be in error. However, when we calculate multiple *t* tests on independent samples that are being measured on the same variable, the rate of error increases exponentially by the number of tests conducted. For example, with our four-group problem, the error rate increases to 18 of 100 times, a substantial increase. The calculation of the rate of type I errors is determined by the following formula:

$$1 - (1 - a)^t$$

where a = the level of significance for the tests and t = the number of test comparisons used. In our example, the calculation would give us:

$$1 - (1 - 0.05)^4 = 0.18.$$

Instead of using a series of individual comparisons, we examine the differences among the groups through an analysis that considers the variation across all groups at once. This test is the *analysis of variance* (ANOVA).

STATISTICAL QUESTION IN ANALYSIS OF VARIANCE

The statistical question using ANOVA is based on the null hypothesis: the assumption that all groups are equal and drawn from the same population. Any difference comes from a random sampling difference. The question answered by the ANOVA test is whether group means differ from each other.

TYPE OF DATA REQUIRED

With ANOVA, the independent variable(s) are at the nominal level. A one-way ANOVA means that there is only one independent variable (often called *factor*). That independent variable has two or more levels. Gender would be a variable with two levels, whereas race, religion, and so forth may have varying numbers of levels depending on how the variable is defined. Two-way ANOVA indicates two independent variables, and *n*-way ANOVA indicates that the number of independent variables is defined by *n*. The dependent variable must be continuous and meet the assumptions described in the next section.

For example, Anderson and Helms (1998) used analysis of variance to compare hospitals grouped by size (small, medium, large, and very large) on their scores on the Referral Data Inventory (RDI), which "measures the amount and type of information an ECF (extended care facility) receives upon referral of an elderly patient from a hospital, as well as the organizational and medical factors associated with interorganizational communication" (p. 388). Table 6-1 contains an edited version of the results from the study. The RDI scores can range between 0 and 40, with higher scores indicating more data sent from hospital to extended care facility. Looking at the mean scores, we can see that large hospitals sent the least data (mean = 30.02), and very large hospitals sent the most (36.06). The *F* ratio is significant at $p < .0001$. The Scheffe posthoc test was

TABLE 6-1 *Summary of ANOVA for Size of Referring Hospital (N = 455)*

	Total Score
Hospital Size	*(0–40) M (SD)*
Small (Sm) (n = 50)	32.66 (2.37)
Medium (Med) (n = 126)	32.41 (2.46)
Large (Lg) (n = 257)	30.02 (3.06)
Very Large (Vlg) (n = 22)	36.06 (3.22)
SS (Between Grp.)	1171.85
SS (Within Grp.)	3661.43
df (Between [Within])	3 [451]
MS (Between Grp.)	390.61
MS (Within Grp.)	8.11
F ratio	48.11**
Scheffe	Lg < others Vlg > others

**$p < .0001$.
Adapted from Anderson, M. A., & Helms, L. B. (1998). Extended care referral after hospital discharge. *Research in Nursing & Health, 21*(5), 385–394.

used to determine which pairs of scores were significantly different. There are six pairwise comparisons that can be made, small with medium, small with large, small with very large, medium with large, medium with very large, and large with very large. You wouldn't expect the small- and medium-sized hospitals to be significantly different, given the closeness of their mean scores. The authors have reported the significant results at the bottom of the column. Large (with the lowest mean score) is reported as significantly less than the other three groups, and Very Large (with the highest mean score) is reported as significantly higher than the other three groups.

ASSUMPTIONS

ANOVA has been shown to be fairly *robust*. This means that even if the researchers do not rigidly adhere to the assumptions, the results may still be close to the truth. The assumptions for ANOVA are the same as those for the t test; that is, the dependent variable should be a continuous variable that is normally distributed, the groups should be mutually exclusive (independent of each other), and the groups should have equal variances (homogeneity of variance requirement).

TABLE 6-2 *Sample Size Determination*

	$u = 4$											
	f											
Power	*.05*	*.10*	*.15*	*.20*	*.25*	*.30*	*.35*	*.40*	*.50*	*.60*	*.70*	*.80*
.10	74	19	9	6	4	3	2	2	—	—	—	—
.50	514	129	58	33	21	15	11	9	6	5	4	3
.70	776	195	87	49	32	22	17	13	9	6	5	4
.80	956	240	107	61	39	27	20	16	10	8	6	5
.90	1231	309	138	78	50	35	26	20	13	10	7	6
.95	1486	372	166	94	60	42	31	24	16	11	9	7
.99	2021	506	225	127	82	57	42	33	21	15	11	9

Adapted from Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Assoc.

SAMPLE SIZE CONSIDERATIONS AND POWER

The principles relating to considerations of sample size and power are based on those outlined in Chapter 5, where two means were compared through use of the *t* test. Using means and standard deviations from previous work, we can calculate expected effect sizes, either by using the formulas provided in Cohen (1987) or by one of the software programs available. Given an expected effect size, desired power, and alpha level, we can determine sample size. For example, Table 6-2 contains an excerpt from Cohen's Table 8.44 (p. 384). It is used to determine appropriate sample size when alpha is .05 and the degrees of freedom (*df*; u) equal 4. Because the *df* is one less than the number of groups, this table is used when there are five groups. The effect sizes are indicated by *f* across the top of the table, and the power values are listed down the left side. Suppose we calculated an effect size of .30. (Cohen defines a moderate effect size as .25, which with two groups is still half of a standard deviation unit.) If we desired a power of .80, how many subjects would we need in each group, and how many would we need overall? We would need 27 subjects in each group, and because there are 5 groups, we would need a total of 135 subjects.

SOURCE OF VARIANCE

According to the null hypothesis, all groups are from the same population, and each of their scores comes from the same population of measures. Any variability of scores can be seen in two ways: First, the scores vary from each other in their own group; and second, the groups vary from each other. The first variation is called *within-group variation*; the second variation is called *between-group variation*. Together, the two types of variation add up to the total variation.

Students often are confused when we say that ANOVA tells us whether the means of groups differ significantly and then proceed to talk about analyzing variance. The *t* test was clearly a test of mean difference, because the difference between the two means was contained in the numerator of the *t* test formula. It is important to understand how analyzing the variability of groups on some measure can tell us whether their measures of central tendency (means) differ.

With ANOVA, the variance of each group is measured separately; all the subjects are then lumped together, and the variance of the total group is computed. If the variance of the total group (total variation) is about the same as the average of the variances of the separate groups (within-group variation), the means of the separate groups are not different. This is because if total variation is the sum of within-group variation and between-group variation, and if within-group variation and total variation are equal, there is no between-group variation. This should become more clear in the diagrams that follow. However, if the variance of the total group is much larger than the average variation within the separate groups, a significant mean difference exists between at least two of the subgroups. In that case, the within-group variation does not equal the total variation. The difference between them must equal the between-group variation.

To visualize the difference in the types of variation, consider three groups exposed to three different experimental conditions. Suppose that the three conditions yielded such widely different scores that there was no overlap among the three groups in terms of the outcome measure (Fig. 6-1). We could then represent our three

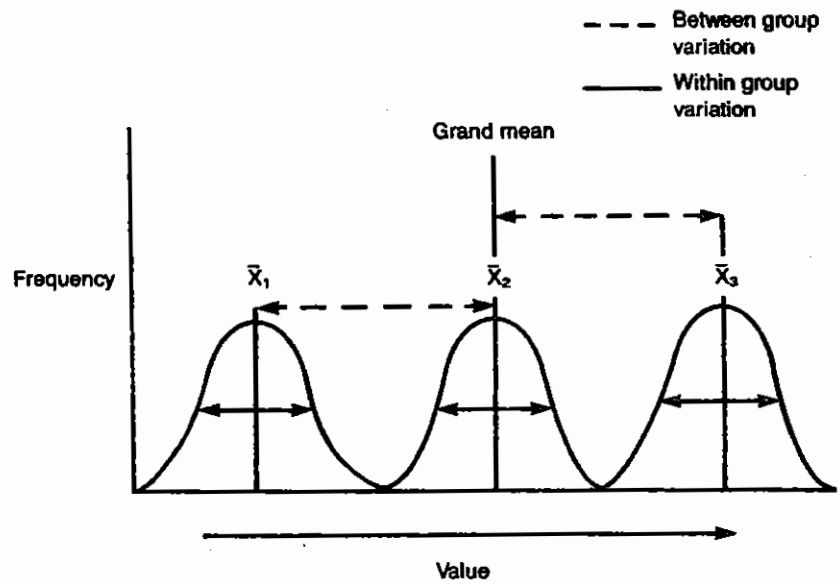


FIGURE 6-1. Between-group and within-group variation: The case of no overlap.

groups in terms of their relationship to each other and in terms of a total group. Each group would then have its own mean and its own distribution around its mean. At the same time, there would be a *grand mean*, which is a mean for all the groups combined. As shown in Figure 6-1, we can look at the variation within the groups and between the groups. The combination of the within-group and between-group variation equals the total variation.

The ANOVA test examines the variation and tests whether the between-group variation exceeds the within-group variation. When the between-group variance is greater (statistically greater) than the within-group variance, the means of the groups must be different. However, when the within-group variance is approximately the same as the between-group variance, the group's means are not importantly different. This relationship between the difference among groups and the different types of variance is shown in Figure 6-2.

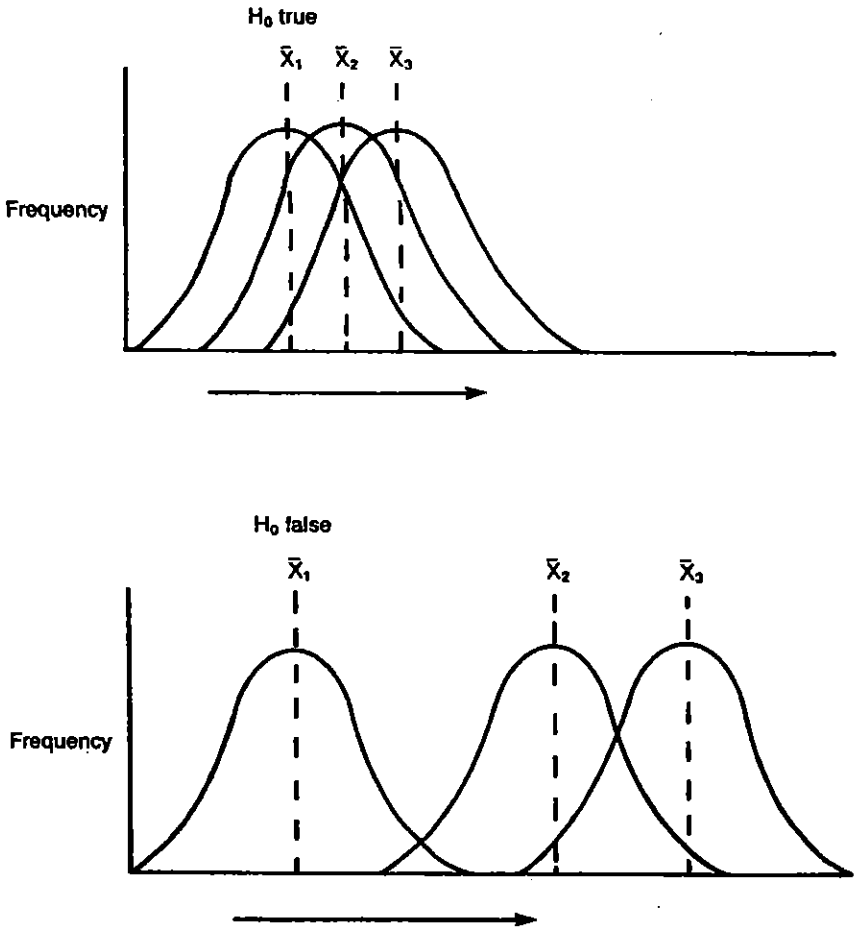


FIGURE 6-2. Relationship of variation to null hypothesis.

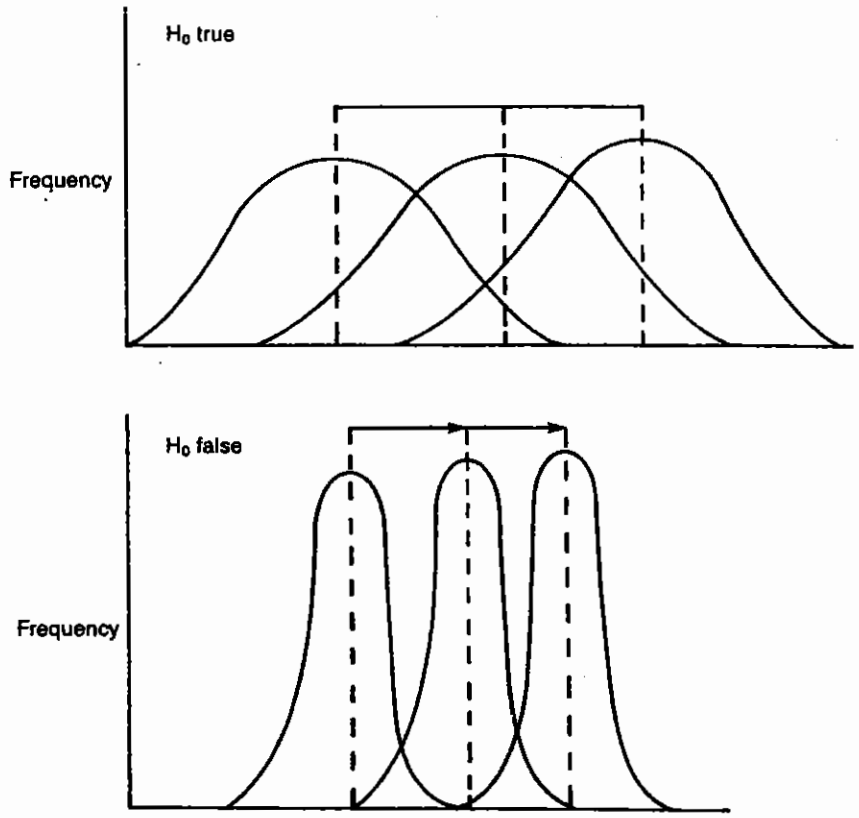


FIGURE 6-3. Effect of within-group variation on null hypothesis.

When the null hypothesis (H_0) is true, the groups overlap to a large extent, and the within-group variation exceeds the between-group variation. When the null hypothesis is false, the groups show little overlapping, and the distance between group is greater. In the lower portion of Fig. 6-2, we see that group 1 overlaps very little with group 2 and not at all with group 3. Groups 2 and 3 do overlap. In that case, it may be that group 1 scored significantly lower than groups 2 and 3, and that groups 2 and 3 do not differ significantly from each other. Thus, the group variation and the deviation between group means determine the likelihood that the null hypothesis is true.

Figure 6-3 illustrates the fact that when the variation within a group or groups is great, the difference between the groups must be greater than when the distribution within groups is narrow to reject the null hypothesis. In the same way, when the group distributions are narrow (low within-group variance), relatively small between-group differences will be significant.

MEASURE OF VARIANCE: SUMS OF SQUARES

The kinds of variation of scores within groups has an intuitive and a statistical meaning. We have discussed the intuitive meaning as the extent to which the scores

TABLE 6-3 *Calculation of Sum of Squares*

	Group 1	*Group 2*	*Group 3*
	1	4	6
	2	5	8
	3	3	5
	2	4	5
$\bar{X}_s$	2	4	6

Total Sum of Squares

Raw Scores	*Deviations from Grand Mean*	*Squared Deviation*
1	$1 - 4 = -3$	9
2	$2 - 4 = -2$	4
3	$3 - 4 = -1$	1
2	$2 - 4 = -2$	4
4	$4 - 4 = 0$	0
5	$5 - 4 = 1$	1
3	$3 - 4 = -1$	1
4	$4 - 4 = 0$	0
6	$6 - 4 = 2$	4
8	$8 - 4 = 4$	16
5	$5 - 4 = 1$	1
5	$5 - 4 = 1$	1
Grand Mean = 4	*Sum* = 0	42

Total sum of squares = 42

Within Sum of Squares

	Group 1	
Raw Scores	*Deviations from Group Mean*	*Squared Deviations*
1	$1 - 2 = -1$	1
2	$2 - 2 = 0$	0
3	$3 - 2 = 1$	1
2	$2 - 2 = 0$	0
$\bar{X} = 2$	*Sum* = 0	*Sum* = 2

(*continued*)

TABLE 6-3 *(Continued)*

Group 2

Raw Scores	*Deviations from Group Mean*	*Squared Deviations*
4	4 − 4 = 0	0
5	5 − 4 = 1	1
3	3 − 4 = −1	1
4	4 − 4 = 0	0
$\bar{X} = 4$	*Sum* = 0	*Sum* = 2

Group 3

6	6 − 6 = 0	0
8	8 − 6 = 2	4
5	5 − 6 = −1	1
5	5 − 6 = −1	1
$\bar{X} = 6$	*Sum* = 0	*Sum* = 6

Within sum of squares = 2 + 2 + 6 = 10

Between Sum of Squares

Deviations of Group Means from Grand Mean	*Squared Deviations*	*Number in Group*
Group 1 2 − 4 = −2	4	4
Group 2 4 − 4 = 0	0	4
Group 3 6 − 4 = 2	4	4

Between Sum of Squares = (4)(4) + (4)(0) + (4)(4) = 32

Summary Table

Source of Variance	*SS*	*df*	*MS*	*F*	*p*
Between group	32	2	16	14.41	<.01
Within group	10	9	1.11		
Total	42	11			

contains an example of the calculation of the sums of squares using the formulas based on the deviations of the scores from their respective means. The data consist of four scores in each of three groups. The means for the three groups are 2, 4, and 6, respectively.

Sum of Squares for Total Variation

The total sum of squares is equal to the sum of the squared deviations of each score in all groups from the grand mean. In our example, the grand mean (mean of the nine scores) is 4. The sum of the deviations around the mean equals zero, and the sum of the squared deviations equals 42. This total sum of squares represents the basis of the null hypothesis that all the subjects belong to one population, which is described by the grand mean.

Sum of Squares for within-Group Variation

The within-group variation is the total of the variation that occurs in each subgroup. It is calculated by finding the sum of squares for each group separately and then summing the results. The sums of the squared deviations for the three groups are 2, 2, and 6, and the sum across the three groups is 10.

Sum of Squares for between-Group Variation

The between-group variation examines how each of the groups varies from the grand mean. For this calculation, we use group means as representative of the individual groups. The between-group variation examines the variation of the group means from the grand mean. In Table 6-3, the mean for group 1 is two less than the grand mean. The sum of the deviations around the grand mean is (as always) zero, and the squared deviations are 4, 0, and 4, respectively. Because the weight of the difference of any mean from the grand mean is influenced by the number of the scores in the group, we weight the squared deviations by the number in the group. The weighted squared deviations are then summed to provide the between sum of squares (32).

In summary, these three sums of squares define the three different kinds of variation that exist when subjects are members of different groups and are measured on a single variable. They include the total variation of each of the scores around the grand mean, the variation of scores within their respective groups, and the deviation between groups measured by the deviation of group means from the grand mean.

DISPLAYING THE RESULTS: SUMMARY OF ANALYSIS OF VARIANCE

The results of the calculations leading to the F ratio are summarized in a table form that is standard for presenting ANOVA results. This presentation of the results is called the *summary of ANOVA table.* In Table 6-3, *SS* stands for sum of squares, *df* for degrees of freedom, *MS* for mean square, F for the statistic generated, and p for the probability level.

Degrees of Freedom

The *df* for the between-group variance is equal to the number of groups minus one. In our example, this is $3 - 1 = 2$. The *df* for the within-group variance is equal to the total number of subjects minus the number of groups, or $12 - 3 = 9$. The *df* for the total variance is equal to the number of subjects minus one ($12 - 1 = 11$). The mean square is the sum of squares divided by its *df*. Thus, the between-group sum of squares, 32, divided by 2 results in a mean square of 16.

Testing the Difference among Groups: The *F* Ratio

To determine whether the between-group difference is great enough to reject the null hypothesis, we compare it statistically to the within-group variance. The *F* represents the ratio of between to within variance and is calculated as the between mean square divided by the within mean square, or $16/1.11 = 14.41$. The *F* value is compared to the values obtained when the null hypothesis is true, and the scores are randomly selected from one population. To make the interpretation, we use the table that presents the *F* distributions (Appendix D). We locate the critical values for comparison by using the *df* for the between and within mean squares.

In the example, the between *df* was 2 and the within df was 9. We locate the between *df* on the row across the top of the table, and we locate the within *df* on the column on the left side of the table. With these points as coordinates, we locate two critical values for *F*. The top value (in light print) is 4.26. This is the value required to reject the null hypothesis at a probability level of 0.05 (given a one-tailed test). The value below (in bold print) is 8.02, the value required to reject the null hypothesis at the 0.01 level. The value of 14.41 is greater than the value required to reach an alpha of 0.01. Therefore, we can reject the null hypothesis at the 0.01 level. We say we have reached a probability level of "less than 0.01." In summary, we obtained an *F* value of 14.41. We therefore rejected the null hypothesis that there were no differences between the groups, and we concluded that the groups were different.

In other standard presentations of ANOVA summary tables, the within variance is sometimes called the *error variance* or *error term*. This terminology reflects the assumption of the ANOVA that the within difference is sampling error or random difference.

In addition to the summary table, often it is helpful to include a table in your results section that shows the means and standard deviations for the scores of each group. One can then see which group scored higher and by how much. Without further analysis, however, we do not know which pairs of means differ significantly. A posthoc analysis would allow us to compare group 1 with group 2, group 1 with group 3, and group 2 with group 3. Before discussing such contrasts in detail, however, we first present another example with a computer analysis of the data.

ONE-WAY ANALYSIS OF VARIANCE

We have one independent categorical variable with n levels and one continuous dependent variable that meets the assumptions, that is that it is normally distributed and that the variance is equal across the groups. To demonstrate, we used data from

One way

Descriptives

CONCERTS

	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Minimum	Maximum
					Lower bound	Upper bound		
Private home	195	5.4103	4.26879	.30569	4.8073	6.0132	.00	10.00
Apartment	86	5.6047	4.37447	.47171	4.6668	6.5425	.00	10.00
Elder housing	139	4.1079	4.56466	.38717	3.3424	4.8735	.00	10.00
Total	420	5.0190	4.42704	.21602	4.5944	5.4437	.00	10.00

Test of Homogeneity of Variances

CONCERTS

Levene statistic	df1	df2	Sig.
2.218	2	417	.110

AUTHOR COMMENTS
The assumption of homogeneity of variance has been met.

ANOVA

CONCERTS

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	174.729	2	87.364	4.533	.011
Within groups	8037.119	417	19.274		
Total	8211.848	419			

AUTHOR COMMENTS
The overall analysis is significant (p = .011).

FIGURE 6-4. Computer output of one-way analysis of variance with posthoc comparisons.

Posthoc Tests

Multiple Comparisons

Dependent Variable: CONCERTS
Scheffé

(I) Living recoded	(J) Living recoded	Mean difference (I-J)	Std. error	Sig.	95% confidence interval	
					Lower bound	Upper bound
Private home	Apartment	−.1944	.56829	.943	−1.5904	1.2016
	Elder housing	1.3023(*)	.48734	.029	.1052	2.4995
Apartment	Private home	.1944	.56829	.943	−1.2016	1.5904
	Elder housing	1.4967(*)	.60231	.047	.0171	2.9763
Elder housing	Private home	−1.3023(*)	.48734	.029	−2.4995	−.1052
	Apartment	−1.4967(*)	.60231	.047	−2.9763	−.0171

* The mean difference is significant at the .05 level.

FIGURE 6-4. (*Continued*)

Dr. Wood's study on promoting breast self-examination (1997). The housing of her subjects can be described by three categories: private home, apartment, and elder housing. Subjects were asked to rate the desirability of certain things that could be offered to them for participation in the study. One choice was concert tickets. Participants rated this choice from 0 = undesirable to 10 = very attractive. The research question is whether the three housing groups differ significantly in their rating of the desirability of receiving concert tickets.

COMPUTER ANALYSIS

To answer the research question, the data were submitted to analysis by the one-way program in SPSS for Windows. This program handles one-way ANOVA (one independent variable) and posthoc tests necessary to compare pairs of means. Figure 6-4 contains the computer output. Author comments have been added to ease interpretation and appear in a shaded box.

The dependent variable is the rating of the desirability of concert tickets, and the independent variable is housing group with three levels: private home, apartment, and elder housing. The descriptive statistics are given first. The groups are somewhat unequal with the smallest number living in apartments (86) and the largest number

living in private homes (195). Looking at the mean scores, we see that on a scale of 0 to 10, the groups are about in the middle. The elder housing group has the lowest mean rating (4.11), and the apartment group gave it the highest rating (5.60). The standard deviations and standard errors are listed. Based on the standard errors, 95% confidence intervals (CI) also are listed. For the private home group, the 95% CI is 4.8073 to 6.0132. This means that if 100 similar samples were drawn, in 95 out of 100 tests, the mean would fall between 4.8 and 6.0. The minimum and maximum scores for each group also are listed. For each group, the entire potential range of scores was covered, that is, all three groups had low scores of zero and high scores of 10.

The assumption of homogeneity of variance is met ($p = .110$). The ANOVA summary table is typical of what is reported in the literature. The variance is reported as between groups, within groups, and total. "Between groups" indicates the differences among the three groups, "within groups" is the error term, and "total" is the total variance in the dependent variable.

Sums of squares are reported first. Because there are three groups, $df = 2$ (number of groups minus one). Dividing the sum of squares by its associated df gives the mean square value. For example, for between groups, 174.729/2 = 87.364. The F is the ratio of between to within variance, or 87.364/19.274 = 4.533. This number is significant at the .011 level.

Because the overall F is significant, we want to know which pairs of means are significantly different. The Scheffé posthoc procedure, which will be described in the next section, was requested. All possible pairwise comparisons are tested. We see that there is a significant difference between the elder housing group and both of the other two groups. For the comparison with the private home group $p = .029$ and for the apartment group $p = .047$. The private home and apartment groups did not differ from each other ($p = .943$). Looking at the means to describe the results, we would say that the group that lived in elder housing rated the desirability of concert tickets (mean = 4.11) significantly lower than the private home group (mean = 5.41) and the apartment group (mean = 5.60).

MULTIPLE GROUP COMPARISONS

Two types of comparisons can be made among group means. The most commonly reported are posthoc (after the fact) comparisons and a priori (planned) comparisons, based on hypotheses stated before the analysis.

Posthoc Tests

When a significant F test is obtained, the null hypothesis that all the groups are from the same population or that all the populations are equal is rejected; that is, we can state that there is a difference among the groups. However, when more than two groups are being compared, we cannot determine from the F test alone which groups differ from each other. In other words, a significant F test does not mean that every group in the analysis is different from every other group. Many patterns of difference are possible. Some of the groups may be similar, forming a

cluster that is different from another select group; depending on the number of groups being compared, there may be wide deviation between each pair of the groups.

To determine where the significant differences lie, further analysis is required. Therefore, we must compare group means. However, if we decide to use the standard *t* test, we are confronted with the possibility of an increased rate of type I errors. To prevent this, secondary analyses following the computation of the *F* ratio are available to pinpoint the source of the difference.

Many techniques exist. A complete discussion of each is beyond the scope of this book, but the aim of all is to decrease the likelihood of making a type I error when making multiple comparisons. For more details on posthoc tests following ANOVA, we suggest Klockars and Sax (1991), and Toothaker (1993).

The *Scheffé test* is reported frequently. The formula is based on the usual formula for the calculation of a *t* test or *F* ratio. The critical value used for determining whether the resulting *F* statistic is significant is different. In other words, the *F* associated with comparing the two means is the same as if they had been compared in the usual ANOVA, but the critical value is changed based on the number of comparisons. The new critical value is simply the usual value multiplied by the number of groups being compared minus one. In our example in Fig. 6-4, the critical value at the 0.05 level with 2 and 417 *df* is 3.02 (see Appendix D). Multiplying that by 2 (the number of groups minus one) results in a critical value of 6.04. Thus, the critical value is twice as stringent when making all possible comparisons among three groups than it was for the overall analysis. The Scheffé test is stringent, but it can be used with groups of equal and unequal size.

The *Bonferroni correction* has been explained previously. The desired alpha is divided by the number of comparisons. For example, with an alpha of 0.05 and four comparisons, the significance level would have to be equal to or less than 0.0125 for the paired comparison to be significant.

The *Duncan test* is computed in the same way as the student Newman-Keuls, but the critical value is less stringent.

The *Least Significant Difference test* is equivalent to multiple *t* tests. The modification is that a pooled estimate of variance is used rather than variance common to groups being compared.

Student Newman-Keuls is similar to Tukey's honestly significant difference (HSD) but the critical values do not stay the same. They reflect the variables being compared.

Tukey's honestly significant difference (HSD) is the most conservative comparison test and as such is the least powerful. The critical values for Tukey remain the same for each comparison, regardless of the total number of means to be compared.

Tukey's wholly significant difference uses critical values that are the average of those used in Tukey's HSD and Student Newman-Keuls. It is therefore intermediate in conservatism between those two measures.

EXAMPLE FROM THE LITERATURE

Look again at Table 6-1 that contains the table from Anderson and Helms (1998). They used the Scheffe test for pairwise comparisons. Their *F* ratio (48.11) was significant at

$p < .0001$. It was, therefore, appropriate to use a posthoc test. It should be noted that even when the overall F is significant, it is possible that none of the pairwise comparisons will be significant. This is because the posthoc tests protect against a type I error by being more stringent. Given the four groups of hospitals, six posthoc comparisons could be made. The authors tell us that large hospitals differed significantly from the other three, as did the very large hospitals. That means that the only comparison that was not significant was between small and medium hospitals. Here is the breakdown of possible comparisons. An asterisk indicates significant differences. Take a moment to look at Table 6-1 and this outline to be sure you understand it.

Comparisons by hospital size:

Small with Medium
Small with Large*
Small with Very Large*
Medium with Large*
Medium with Very Large*
Large with Very Large*

Basically, since Large and Very Large are reported as significantly different from all other groups, any comparison that they are in is significant.

Planned Comparisons

Planned comparisons, or a priori contrasts, are based on hypotheses stated before data are collected. When you hypothesize ahead of time, you can use more powerful statistical tests. One way to do this is through the development of prespecified contrasts that are *orthogonal* to each other. Orthogonal means that the hypothesis tests are unrelated to each other; that is, knowing one result tells you nothing about the other. For an overview of planned comparisons versus omnibus tests, refer to Wu and Slakter (1990). Here we demonstrate how orthogonal contrasts can be developed and analyzed in SPSS for Windows. To have comparisons that are independent, only $n - 1$ comparisons can be made. In our three-group living arrangements example (Fig. 6-4), therefore, there could be only two orthogonal contrasts. In our example, we might want to test the hypothesis that the two independent living (private home and apartment) groups will score significantly higher on desire for concert tickets than the elder housing group and that there will be no difference between the apartment and private home groups. Table 6-4 contains the vectors necessary to code such a contrast. On vector 1 (V1), subjects in both independent living groups receive a -1, and the elder housing subjects receive a 2. This contrast tests the difference between the desirability of a concert ticket mean score for all the independent living subjects and the mean for the elder housing subjects. The second contrast is given in vector 2 (V2). The two independent living groups are compared. The elder housing group is not considered in the second contrast. (Note, in building the contrasts, you must list the groups in the order in which they are in the dataset. In our dataset, the values are 1 = Private home, 2 = Apartment, and 3 = Elder housing.)

TABLE 6-4 *Orthogonal Coding*

	Vectors	
Groups	*V1*	*V2*
Private home	−1	1
Apartment	−1	−1
Elder housing	+2	0

To ensure that hypothesized contrasts are orthogonal, three tests must be applied:

1. There must be only $n - 1$ contrasts.
2. The sum of each vector must equal zero. In the example, the sum of V1 is $(-1) + (-1) + 2 = 0$, and the sum of V2 is $1 + (-1) + 0 = 0$.
3. The sum of the cross-products must equal zero. In the example, $(-1 \times 1) + (-1 \times -1) + (2 \times 0) + = 0$.

Table 6-5 provides other examples of possible contrasts, given three groups. Are they all orthogonal? The vectors X1 and X2 reflect an orthogonal contrast, as do the vectors Y1 and Y2. Vectors Z1 and Z2 do not reflect an orthogonal contrast; group 1 is compared to group 2 and to group 3. The sum of the cross-products does not equal zero $(-1 \times 1) - (0 \times -1) + (1 \times 0) = -1$.

We now demonstrate the use of the contrasts specified in Table 6-4 in a computer analysis of these data. See Fig. 6-5 for the computer output of the a priori contrasts. In the first analysis (see Fig. 6-4), we requested a posthoc test and determined that the elder housing group scored significantly lower than the other two groups.

In the analysis in Fig. 6-5, we are testing a priori orthogonal contrasts. Since the Descriptives, Test of Homogeneity of Variance, and ANOVA table are the same as those in Fig. 6-4, they are not repeated here. Author's comments have been added

TABLE 6-5 *Contrasts*

	Pairs of Vectors					
Groups	*X1*	*X2*	*Y1*	*Y2*	*Z1*	*Z2*
1	2	0	−1	1	−1	1
2	−1	1	2	0	0	−1
3	−1	−1	−1	−1	1	0

Contrast Coefficients

Contrast	Type of living quarters		
	Private home	Apartment	Elder housing
1	−1	−1	2
2	1	−1	0

AUTHOR COMMENTS
The first contrast tests the difference between the two independent living groups (each given a coefficient of −1) and the elder housing group (coefficient = 2).
The second contrast tests the difference between the two independent living groups.

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Coupon for concerts	Assume equal variances	1	−2.7991	.93680	−2.988	417	.003
		2	−.1944	.56829	−.342	417	.732
	Does not assume equal variances	1	−2.7991	.95685	−2.925	259.302	.004
		2	−.1944	.56210	−.346	159.093	.730

AUTHOR COMMENTS
Because the homogeneity of variance assumption was met, the equal variance contrasts are appropriate. Only the first contrast is significant.

FIGURE 6-5. Computer output containing a priori contrasts.

to increase clarity and appear in a shaded box. This is a more powerful analysis; that is, it is more likely to find a significant difference among groups. This is because the contrasts are stated a priori and are restricted to orthogonal contrasts. In the case of posthoc tests, the overall F value must be significant before we can test pairwise comparisons. When using orthogonal contrasts, these contrasts can be examined even when the overall F is not significant. The first contrast (read across the row) compares the means of the two independent living groups with the elder housing group. The second contrast compares the two independent living groups.

The equal variance estimate is appropriate because the assumption of homogeneity of variance has been met (see Fig. 6-4, Levene test, $p = .110$). We hypothesized that the first contrast would be significant, but that the second would not. Our hypotheses have been supported. The first contrast is significant ($p = .003$), thus the independent living groups did differ significantly from the elder housing group.

Looking at the means, we see that as hypothesized the independent living groups rated the desirability of concert tickets significantly higher than did the elder housing group. The private home and apartment groups did not differ significantly ($p = .732$). In this example, the a priori and posthoc tests resulted in the same findings. This is not always the case.

A priori contrasts must be based on firm theoretical grounds.

EXAMPLE FROM THE LITERATURE

In a study of Australian nurses' experiences and attitudes in the "do not resuscitate" decision, Manias (1998) used a priori planned contrasts, although not orthogonal contrasts, to test specific contrasts. Nurses from four practice areas (intensive care, coronary care, acute medical, acute surgical) were compared on their experiences in decision making. The a priori contrasts showed that intensive care nurses considered themselves to be less effective in influencing a "do not resuscitate" order compared to the other three groups.

SUMMARY

One-way ANOVA is used to compare the means of two or more groups. When the overall F is significant and more than two groups are being compared, posthoc tests are necessary to determine which pairs of means differ from each other. Also, when directional hypotheses are appropriate, a priori contrasts may be specified and tested.

Application Exercises and Results

Exercises

Run the appropriate analyses to answer the question and test the hypotheses. Write a description of the results.

1. Do the three smoking groups differ significantly in their quality of life during the past month?
2. Test the following hypotheses:
   a. The smoking group will score significantly lower on quality of life during the past month than the other two groups.
   b. There will be no significant difference in quality of life between the group that quit smoking and the group that never smoked.

Results

1. To answer this question, a one-way ANOVA was run and the Scheffé posthoc test was requested. Exercise Fig. 6-1 contains the output.

   Looking at the descriptives, we see that the group that is still smoking had the lowest mean score (4.01) on the 6-point scale that ranged from a low of 1 (very dissatisfied, unhappy most of the time) to a high of 6 (extremely happy, could not be more pleased). The other two groups' scores were almost identical (4.31 and 4.30). The assumption of homogeneity of variance has been met ($p = .300$). The overall F is not significant ($p = .067$). Since the overall F is not significant, it is not appropriate to report the Scheffé test results.
2. One-way ANOVA with a priori contrasts was used to test the two hypotheses. Since the overall analysis is the same as the one in Exercise 1, Exercise Fig. 6-2 contains only the a priori contrasts. From Exercise 1, we know that the assumption of equality of variance has

One Way

Descriptives

Quality of life in past month

	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Minimum	Maximum
					Lower bound	Upper bound		
Never smoked	432	4.31	1.046	.050	4.21	4.40	1	6
Quit smoking	185	4.30	1.018	.075	4.15	4.44	1	6
Still smoking	78	4.01	1.026	.116	3.78	4.24	1	6
Total	695	4.27	1.039	.039	4.19	4.35	1	6

Test of Homogeneity of Variances

Quality of life in past month

Levene statistic	*df*1	*df*2	Sig.
1.205	2	692	.300

ANOVA

Quality of life in past month

	Sum of squares	*df*	Mean square	F	Sig.
Between groups	5.843	2	2.921	2.720	.067
Within groups	743.302	692	1.074		
Total	749.145	694			

EXERCISE FIGURE 6-1. One-way analysis of variance with posthoc test, Exercise 1.

been met (Levene's $p = .300$), and the overall F is not significant ($p = .067$). The first contrast tests whether the still-smoking group differs significantly from the other two groups. The second contrast tests whether the two nonsmoking groups differ from each other.

Because the homogeneity of variance assumption has been met, we use the equal variance contrasts. We would report that the first hypothesis was supported ($p = .000$). The group that is still smoking scored significantly lower on their quality of life score (mean = 4.01) than the other two groups combined. The second hypothesis was also supported ($p = .928$). There was no significant difference between the two nonsmoking groups on their reported quality of life. Thus, the a priori results indicate significant differences, whereas the posthoc did not.

Contrast Coefficients

Contrast	Smoking history		
	Never smoked	Quit smoking	Still smoking
1	1	−1	2
2	1	−1	0

Contrast Tests

		Contrast	Value of Contrast	Std. error	*t*	*df*	Sig. (2 tailed)
Quality of life in past month	Assume equal variances	1	8.03(a)	.252	31.913	692	.000
		2	.01	.091	.091	692	.928
	Does not assume equal variances	1	8.03(a)	.249	32.246	101.477	.000
		2	.01	.090	.092	356.918	.927

a The sum of the contrast coefficients is not zero.

EXERCISE FIGURE 6-2. One-way analysis of variance with a priori contrasts, Exercise 2.