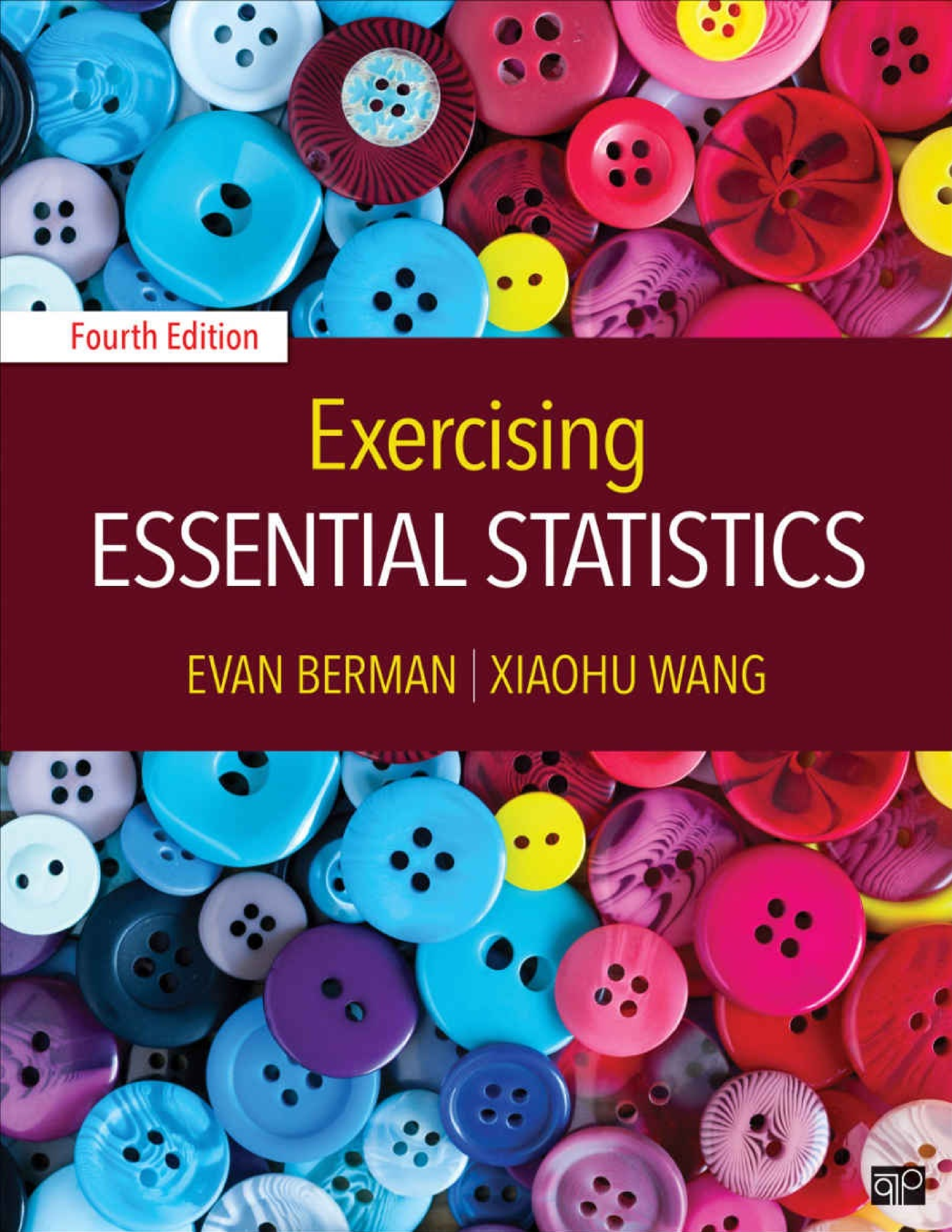


Fourth Edition

Exercising ESSENTIAL STATISTICS

EVAN BERMAN | XIAOHU WANG

Exercising Essential Statistics

Fourth Edition

Exercising Essential Statistics

Fourth Edition

Evan Berman
Victoria University of Wellington
XiaoHu Wang
City University of Hong Kong

SAGE | CQPRESS

FOR INFORMATION:

CQ Press

An Imprint of SAGE Publications, Inc.

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

SAGE Publications Ltd.

1 Oliver’s Yard

55 City Road

London, EC1Y 1SP

United Kingdom

SAGE Publications India Pvt. Ltd.

B 1/I 1 Mohan Cooperative Industrial Area

Mathura Road, New Delhi 110 044

India

SAGE Publications Asia-Pacific Pte. Ltd.

3 Church Street

#10-04 Samsung Hub

Singapore 049483

Printed in the United States of America

ISBN: 978-1-5063-4895-7

This book is printed on acid-free paper.

17 18 19 20 21 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Carrie Brandon

Editorial Assistant: Duncan Marchbank

eLearning Editor: John Scappini

Production Editor: Bennie Clark Allen

Copy Editor: Sarah J. Duffy

Typesetter: C&M Digitals (P) Ltd.

Proofreader: Jen Grubba

Cover Designer: Anupama Krishnan

Marketing Manager: Jennifer Jones

Contents

Introduction

This workbook is an integral part of the *Essential Statistics for Public Managers and Policy Analysts,* Fourth Edition, package. The textbook introduces theory and concepts and offers many examples that show relevance and application, and this workbook strengthens understanding by illustrating applications and encouraging analysts to think through statistical concepts in new and engaging ways. In short, practice makes perfect, and the applications found here are key to mastering statistics.

Chapters 1 through 17 correspond to those in the textbook. The "Chapter Objectives" at the beginning of each textbook chapter are a good guide to what you will be practicing in each workbook chapter. To help students understand new concepts, four sections are featured that target different aspects of learning. Specifically, each of these chapters is divided into the following sections:

- "Q & A" facilitates student learning of the statistical concepts through questions and answers. It is an ideal study guide for tests.
- "Critical Thinking" consists of short exercises to stimulate further conceptual insights into statistics and applications.
- "Application Exercises" in Chapters 1–5 and "Data-Based Exercises" in Chapters 6–17 help students develop hands-on skills using practical applications. The exercises in Chapters 6–17 draw on datasets provided on the companion website, http://study.sagepub.com/bermaness4e.
- "Further Reading" lists work toward self-study in areas of interest to readers, including articles that students may wish to access.

This workbook also includes chapters that extend the material covered in the textbook. Chapter 18 is a user's guide to essential statistics in Excel®. Chapter 19 describes how to use SPSS, a statistical analysis program. It provides a step-by-step approach for ease of learning. Chapter 20 provides documentation for the datasets provided on the companion website. The datasets are provided in SPSS®[*], SAS, SYSTAT, and Stata formats for easy use.

[*] SPSS is a registered trademark of International Business Machines Corporation.

The data sets are available at the SAGE Publishing/CQ Press website: **http://study.sagepub.com/bermaness4e**. If you have any difficulty accessing the datasets, just email Sage at orders@sagepub.com or call Sage at the number on this website. The datasets are the same as in the third edition, and a new one has been added, "Florida Counties." The website includes bonus data, a complete report based on an actual citizen survey, as well as a presentation in Microsoft PowerPoint, which will help users stretch their imaginations as they think about how to present their data to others in the public realm. The report and survey instrument are available electronically in Microsoft Word for your convenience. A file discussing the use of spreadsheets in public management and analysis, including examples in Microsoft Excel, is also available.

We hope this workbook will help users in their learning. Moreover, we hope they are able to readily apply some of these exercises to problems in their workplaces.

Have a question or feedback? Just send us an e-mail, and we’ll be happy to respond. Let us know what works for you and how we can further improve this workbook. We look forward to hearing from you.

Evan Berman

evanmberman@gmail.com

XiaoHu Wang

xwang1989@gmail.com

Chapter 1 Why Statistics for Public Managers and Analysts?

Q & A

1. ***Identify five ways in which analysis and data often are used.***

   The five ways are as follows: (1) to describe and analyze societal problems, (2) to describe policies and programs, (3) to monitor progress and prevent fraud, (4) to improve program operations, and (5) to evaluate policy and program outcomes.
2. ***How does quantitative analysis assist in decision making?***

   Quantitative analysis provides a factual underpinning of situations and responses by quantifying the extent of problems and situations and the actual or likely impact of proposed strategies. At the very least, a focus on facts and objective analysis can reduce judgment errors stemming from overly impressionistic or subjective perceptions that are factually incorrect.
3. ***Identify six competencies for analysis.***

   The six competencies are (1) being familiar with data sources in your line of work, (2) being able to collect your own data, (3) analyzing data, (4) communicating results from analysis, (5) bringing to quantitative analysis the theory and practice of management and policy analysis, and (6) having a strong sense of ethics relating to quantitative analysis.
4. ***What is scientific research?***

   Scientific research is the careful, systematic process of inquiry that leads to the discovery or interpretation of facts, behaviors, and theories. Scientific research is distinguished from personal and other forms of research or inquiry by rather strict standards for accepting new facts and theories as knowledge and by a process that includes other scientists in making such determinations.
5. ***What is statistics?***

   Statistics is the body of systematic knowledge and practice that provides standards and procedures for drawing conclusions from one’s data. Statistics includes specific tools for analyzing data, too.
6. ***Identify four stages of proficiency in quantitative analysis.***

   The four stages of proficiency are know-nothing, journeyman, technocrat, and sophisticated expert. Each stage is associated with distinct development objectives.
7. ***What three areas of ethical concern are identified in connection with analysis?***

   The three areas are as follows: (1) fully disclosing the purposes of analysis, (2) integrity in analysis and communication, and (3) concern for the impact of analysis and research on the welfare of human subjects.
8. ***What is scientific misconduct?***

   Scientific misconduct is generally understood as the violation of the standard norms of scholarly conduct and ethical behavior in scientific research. Scientific misconduct, when considered by others to be significant or severe, can diminish one's reputation and negatively affect one's career, including the possibility of dismissal from one's job and adverse legal action.

9. ***What is the specific problem of dual purposes?***

   Analysts must balance potentially conflicting purposes of (1) furthering programs and policies and (2) establishing objective truths about how well a program is performing.

10. ***Which practices are associated with furthering the integrity of analysis and communication?***

    Analysts should be honest, objective, accurate, and complete. Analysts should not hide facts, change data, falsify results, or consider only data that support a favored conclusion. Analysts should also fully report the sources of their data, data collection methodologies, and any possible gaps and shortfalls, and they should assess the impact of such shortcomings on their findings. Results should be presented in straightforward and nonmisleading ways. These norms provide essential guidance to analysts throughout the entire analytical process.

11. ***What concern should analysts have for the impact of research on the welfare of human subjects?***

    Researchers and analysts should recognize and minimize the potential harm that their research and analysis could have on research subjects. Most human subjects research is now subject to oversight by institutional review boards to ensure that risks to subjects are reasonable and that possible harm is identified and minimized.

Critical Thinking

Note to students: These questions further understanding of selected key points made in the textbook. Questions in the next section, “Application Exercises,” are designed to encourage application of the key points in practice.

1. **What is the difference between describing the extent of a social problem and describing the factors that give rise to it? Give an example. How can the latter be useful for developing programs and policies?**

2. **What is the role of statistics in connection with the six competencies mentioned in the text? What else might be needed to attain these competencies?**

3. **Many programs produce routine, administrative data that are used to monitor progress and prevent fraud. How useful are such data for the five common uses of analysis and data mentioned in the text? What other data might be needed, such as might be obtained from citizen or client surveys?**

____

____

____

4. **Identify a person or situation associated with each of the four stages of proficiency in quantitative analysis.**

____

____

____

____

____

5. **Explain how the following concerns of ethics can affect research and its utilization: (1) dual purposes, (2) full disclosure, (3) truthfulness, (4) alternative explanations, (5) communication, and (6) well-being of human subjects. Give examples of each.**

____

____

____

____

____

Application Exercises

Note to students: This section is called Data-Based Exercises in later chapters (starting in Chapter 6) and will provide you with hands-on exercises that involve real datasets.

1. **Identify five problems or challenges in your area of interest that would benefit from analysis or research.**

    ____

    ____

    ____

    ____

    ____

    ____

2. **Identify at least two examples, in your area of interest, of each of the five common uses of analysis and data.**

    ____

    ____

    ____

    ____

    ____

3. **What data exist in your area of interest? Are there any datasets with which managers and analysts are expected to be familiar?**

    ____

    ____

    ____

    ____

    ____

4. **At what stage of proficiency do you see yourself? What is necessary to get beyond this stage? Develop some learning objectives for yourself.**

5. **Explain how a customer or citizen survey might be useful in your area of interest. What topics might such a survey address? What challenges do you foresee?**

6. **(a) Consider the following proposition: “Almost every department needs people with analytical skills.” Verify this proposition by interviewing managers in your area of interest. Which analytical skills do they say that they are looking for? (b) Research salaries at the U.S. Bureau of Labor Statistics, National Industry-Specific Occupational Employment and Wage Estimates, at www.bls.gov/oes/current/oessrci.htm, and compare wages for occupations that vary in analytical content, for example, management positions in budgeting, information technology, human resource management, and parks and recreation.**

7. **Identify and consider some ethical situations that would challenge the integrity of your analysis and research, such as being asked to withhold relevant information. How might you deal with such situations?**

____

____

____

____

____

____

8. **Research the policies and practices that pertain to ethics in research in your agency or in an agency in your area of interest. If there are none, suggest two or three that would serve as a foundation for a more extensive set of policies.**

____

____

____

____

____

____

Further Reading

Various books offer additional information about the importance and uses of research and analysis in public service. A popular book is Eugene Bardach, *Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*, 5th ed. (Washington, D.C.: CQ Press, 2015). A rather different approach is William Dunn, *Public Policy Analysis: An Introduction* (New York: Prentice Hall, 2016). Scholarship about the use of policy analysis and research traces back to the development of policy analysis as a field in the 1970s and efforts to get public agencies to use it in the 1970s and 1980s. A classic text about the use of analysis is Aaron Wildavski, *Speaking Truth to Power: The Art and Craft of Policy Analysis* (Piscataway, N.J.: Transaction, 1987). Since then, research has focused varyingly on the capacity of government organizations to develop or have policy research and analysis capabilities, and the utilization of such knowledge in their decision-making processes.

Different events and contexts lead to different foci in research. In recent years, a focus has been the utilization of "evidenced-based" policy, resulting from increased capacity of performance measurement (see Chapter 4) and policy analysis. A representative article in this genre is Gary VanLandingham and Torey Silloway, "Bridging the Gap between Evidence and Policy Makers: A Case Study of the Pew-MacArthur Results First Initiative," *Public Administration Review* 76 (2016): 542–546, and, somewhat older, Michael Howlett, "Policy Analytical Capacity and Evidence-Based Policy-Making: Lessons from Canada," *Canadian Public Administration* 52 (2009): 153–175. While the use of evidence is ever more popular, concerns are growing as well; see Holger Strassheim and Pekka Kettunen, "When Does Evidence-Based Policy Turn into Policy-Based Evidence?" *Evidence & Policy: A Journal of Research, Debate and Practice* 10 (2014): 259–277. As these articles suggest, the problem of research utilization is found throughout the world. An older, award-winning article is Réjean Landry, Moktar Lamari, and Nabil Amara, "The Extent and Determinants of the Utilization of University Research in Government Agencies," *Public Administration Review* 63 (March/April 2003): 192–205. This article received the Louis Brownlow Award from the American Society for Public Administration for the best article published in *Public Administration Review* in 2003. Earlier, the focus was on the development of performance measurement. See, for example, Evan Berman and XiaoHu Wang, "Performance Measurement in U.S. Counties: Capacity for Reform," *Public Administration Review* 60 (September/October 2000): 409–420, which reflects the then-growing development of performance measures in local government. But as times change, so, too, does research.

A classic book about the ethics of analysis is Darrell Huff, *How to Lie with Statistics* (New York: Norton, 1993, 1954). Other books on this topic are Joel Best, *More Damned Lies and Statistics: How Numbers Confuse Public Issues* (Berkeley: University of California Press, 2012, 2001) and Matthew Robinson and Renee G. Scherlen, *Lies, Damned Lies, and Drug War Statistics: A Critical Analysis of Claims Made by the Office of National Drug Control Policy* (Albany: State University of New York Press, 2014). As the progress of these titles shows, the genre is getting ever more tailored around specific topics. The National Institutes of Health website provides educational materials that discuss protections for human subjects, which is applicable to all types of research, including studies that public policy managers and analysts might conduct that involve humans in some way, for example, by administering surveys. Training materials related to approval processes for such research (such as by institutional review boards that are found at many research centers and universities) may also be found on the web (see, for example, http://osp.od.nih.gov/office-clinical-research-and-bioethics-policy).

Chapter 2 Research Design

Note to students: This chapter includes questions and exercises relating to the textbook introduction to Section II (Research Methods), indicated by SI (Section Introduction).

Q & A

1. ***What is research methodology? (SI)***

   Research methodology is the science of methods for investigating phenomena. Research methods are used in almost every social science discipline and can be applied to many different kinds of problems, including those found in public and nonprofit management and analysis.
2. ***What is basic research? What is applied research? (SI)***

   Basic research is a research activity whose purpose is to develop new knowledge about phenomena such as problems, events, programs or policies, and their relationships. Applied research is a research activity whose purpose is to develop knowledge for addressing practical problems.
3. ***What are quantitative research methods? What are qualitative research methods? (SI)***

   Quantitative research methods involve the collection of data that can be analyzed using statistical methods. The purpose of quantitative research is to quantify the magnitude of phenomena, to provide statistical evidence about factors affecting these phenomena, and to quantify the impacts of programs and policies. Qualitative research methods involve the collection and analysis of words, symbols, or artifacts that are largely nonstatistical in nature. Typically, the purpose of qualitative research is to identify and describe new phenomena and their relationships.
4. ***What are variables?***

   Variables are empirically observable phenomena that vary.
5. ***What are attributes?***

   Attributes are the specific characteristics of a variable, that is, the specific ways in which a variable can vary. All variables have attributes. For example, the variable “gender” has two attributes: male and female.
6. ***How is descriptive analysis different from the study of relationships?***

   Descriptive analysis provides information about (the level of) individual variables, whereas the study of relationships provides information about the relationships among variables.
7. ***Define the terms* independent variable *and* dependent variable.**

   All causal relationships have independent and dependent variables. The dependent variable is the variable that is affected (caused) by one or more independent variables. Independent variables are variables that cause an effect on other variables but are not themselves shaped by other variables.
8. ***What is a causal relationship, and how is it different from an association?***

   Most studies describe relationships among variables. When relationships are causal, one variable is said to be the cause of another. When variables are only associated with each other, no effort is made to identify patterns of causation.

9. ***What is required for establishing a claim of causality?***

   To establish causality, there must be both empirical correlation and a plausible theory that explains how these variables are causally related.

10. ***What is a hypothesis?***

    A hypothesis is a relationship that has not yet been tested empirically.

11. ***What six steps are involved in program evaluation?***

    The six steps are as follows: (1) defining program goals and activities, (2) identifying which key relationships will be studied, (3) determining what type of research design will be used, (4) defining and measuring study concepts, (5) collecting and analyzing program data, and (6) presenting study findings.

12. ***What are rival hypotheses? What are control variables?***

    Rival hypotheses state threats to the credibility of study conclusions. Control variables are variables used in empirical research to evaluate rival hypotheses.

13. ***Explain the role of statistics in determining the impact of rival hypotheses (or control variables) on program outcomes.***

    The impact of rival hypotheses can seldom be ascertained through research design alone. Statistics, then, are used to examine these impacts. (Statistics are also used for other purposes, explained in subsequent chapters of *Essential Statistics for Public Managers and Policy Analysts*.)

14. ***How are quasi-experimental designs different from classic, randomized experiments?***

    In a classic experimental design, participants are randomly assigned to either a control or an experimental group. The *only* systematic difference between the groups is that study group participants receive an intervention (called a stimulus, such as a therapy, subsidy, or training). Any outcome differences between these two groups are then attributed to the systematic difference—the treatment. Such testing conditions are seldom possible in public management and policy. Quasi-experimental designs are imperfect research designs that may lack baselines, comparison groups, or randomization, which are present in classic, randomized experiments.

15. ***What are threats to external validity?***

    Threats to external validity are those that jeopardize the generalizability of study conclusions to other situations. These threats often concern unique features of the study population or research design.

16. ***What are threats to internal validity?***

    Threats to internal validity are those that jeopardize study conclusions about whether an intervention in fact caused a difference in the study population. These threats often question the logic of study conclusions.

Critical Thinking

1. **Give examples of basic and applied research questions that might be raised in the context of (1) a program to reduce adult illiteracy and (2) a program that fights international terrorism. (SI)**

2. **Why are both quantitative and qualitative methods indispensable in addressing questions of basic and applied research? (SI)**

3. **Give some examples of variables. Why are variables key to research?**

4. **A program aims to reduce adult illiteracy by providing reading sessions during evening**

hours. Identify the dependent and independent variables.

____

____

____

____

____

____

5. **A study examines the impact of gender and drug use on school performance and political orientations. Identify the dependent and independent variables.**

____

____

____

____

____

____

6. **It is said that in Sweden an empirical association exists between the presence of storks and the incidence of new babies. Explain what is necessary to establish a claim of causation. Do storks really bring babies?**

____

____

____

____

____

____

7. **A study examines the relationship between race and crime. Is this a causal relationship or an association? Explain.**

____

____

____

____

____

____

8. **Apply the following statement to program evaluation: "Research begins with asking questions." Think about a program that you know about as a basis for answering this question.**

____

____

____

____

____

____

9. **The developers of the adult literacy program mentioned in question 4 claim that the program is effective. By what measures might this effectiveness be demonstrated?**

____

____

____

____

____

____

10. **What might be some rival hypotheses regarding the effectiveness of this adult literacy program?**

____

____

____

____

____

____

11. **Discuss an experimental research design for testing the effectiveness of an anger management program. Then apply the three quasi-experimental designs mentioned in Box 2.2 in the text.**

____

____

____

____

____

____

Application Exercises

1. **Give examples of basic and applied research questions in your area of interest. (SI)**

   ____

   ____

   ____

   ____

   ____

   ____

2. **Give examples of quantitative and qualitative research methods in your area of interest. (SI)**

   ____

   ____

   ____

   ____

   ____

   ____

3. **Consider the following variables: the number of immigrants, attitudes toward abortion, and environmental pollution. What might be some attributes of each of the variables?**

   ____

   ____

   ____

   ____

   ____

   ____

4. **You have been asked to develop a neighborhood crime control program. Thinking ahead, you develop a strategy for evaluating the program in subsequent months and**

years. Define the program and identify dependent and independent variables that can be used to evaluate it.

____

____

____

____

____

____

5. **Identify a problem in your area of interest. Identify the dependent and one or more independent variables affecting this problem.**

____

____

____

____

____

____

6. **Consider a program or policy in your area of interest. How do the specific issues raised in the text regarding program evaluation apply to your program or policy? For instance, give some examples of how difficult it can be to document program outcomes.**

____

____

____

____

____

____

7. **Discuss how you can apply the six steps of program evaluation to a specific program in your area of interest.**

____

8. **Find an article that discusses a specific program evaluation and identify in it each of the six steps of program evaluation.**

9. **Identify some rival hypotheses (control variables) that might affect conclusions about the effectiveness of an adult literacy program. Then, discuss how an experimental research design and several quasi-experimental designs might be helpful for determining the effectiveness of the program.**

10. **Define the objectives of a job training program, and then identify some rival hypotheses regarding possible outcomes. Explain how baselines and comparison groups might be used.**

Further Reading

The all-time, best-selling, easy-to-read general textbook on research methods in social science (general) is Earl Babbie, *The Practice of Social Research,* 14th ed. or later (Belmont, Calif.: Wadsworth, 2016). A book with a focus on statistics in public administration is Maureen Berner, *Statistics for Public Administration: Practical Uses for Better Decision Making,* 2nd ed. (Washington, D.C.: ICMA, 2013). A variety of books on program evaluation may be found readily online, and we encourage readers to choose one. An example is David Royse, Bruce A. Thyer, and Deborah K. Padgett, *Program Evaluation: An Introduction to an Evidence-Based Approach,* 6th ed. (Boston, Mass.: Cengage, 2015). Another source with applications in public administration is Kathryn E. Newcomer, Harry P. Hatry, and Joseph S. Wholey (Eds.), *Handbook of Practical Program Evaluation,* 4th ed. (San Francisco: Jossey-Bass, 2015). A classic book on program evaluation is Peter Rossi, Mark W. Lipsey, and Howard E. Freeman, *Evaluation: A Systematic Approach,* 7th ed. or later (Thousand Oaks, Calif.: Sage, 2003), but that book is now getting dated.

Research methods are used widely in scholarly research, of course. The *Journal of Policy Analysis and Management* publishes many articles that evaluate specific programs and policies. Some of these articles, though not all, are grounded in economic thought. Policy journals are doing well these days, and many fields have their own policy journals, such as *Education Policy, Transport Policy, Space Policy*, *Environmental Policy and Governance*, *Research Policy,* and so on. Also, most empirical articles in the leading journals in public administration, political science, and nonprofit management use the terminology of independent, dependent, and control variables discussed in this chapter. You should have no problem picking up any leading scholarly journal in your field and finding these terms used. A few studies in public and nonprofit management and policy analysis use comparison groups and quasi-experimental designs, but most rely on statistical techniques to account for control variables. These techniques are discussed later in the textbook.

Chapter 3 Conceptualization and Measurement

Q & A

1. ***What is a scale?***

   A scale is the collection of attributes that is used to measure a specific variable. Scales are important because they define the nature of information about variables.
2. ***What is a nominal-level scale? What is an ordinal-level scale?***

   A nominal-level scale exhibits no ordering among the categories. The variable “region” is an example of a variable with a nominal scale. An ordinal-level scale exhibits order among categories, though without exact distances between successive categories. Likert scales are examples of ordinal scales. Variables with ordinal- or nominal-level scales are called categorical (or discrete) variables.
3. ***What is an interval-level scale? What is a ratio-level scale?***

   Interval- and ratio-level scales exhibit both order and distance among categories. The only difference between interval and ratio scales is that the latter have a true zero. Income is an example of a ratio-level variable when it is measured in actual dollars; someone who earns $75,000 per year makes exactly three times that of someone making $25,000, and it is possible to make $0 (no income). The distinction between ratio- and interval-level variables is typically of little relevance to public and nonprofit administration and policy analysis. Variables with interval- and ratio-level scales are also called continuous variables.
4. ***What are incomplete, ambiguous, and overlapping scales? Why must they be avoided?***

   An incomplete scale omits response categories, an ambiguous scale has ill-defined response categories, and an overlapping scale has at least one response that is covered by more than one category. Incomplete, ambiguous, and overlapping scales should be avoided because they have limited measurement validity.
5. ***Define and contrast the terms* variable *and* concept.**

   Variables are empirically observable phenomena that vary, whereas concepts are abstract ideas. Variables are observed directly; concepts are observed indirectly (through variables).
6. ***Describe the two steps involved in concept measurement.***

   Concept measurement involves two steps: first, the process of specifying all relevant dimensions of concepts (conceptualization) and, second, the process of specifying which variables will be used to measure (operationalization). Complex concepts and those that are key to the research design are usually conceptualized with greater rigor than are those that are simple or less key to the program or evaluation.
7. ***What three strategies of operationalization are mentioned in the text?***

   Three approaches to operationalization are (1) to develop separate measures for each dimension, (2) to develop a single set of measures that encompass the dimensions, and (3) to develop a single measure. These three strategies reflect a declining order of rigor.

8. ***What is the theorem of the interchangeability of indicators, and why is it important?***

   The theorem of the interchangeability of indicators states that if several measures are equally valid indicators of a concept, then any subset of these measures will be valid as well. In other words, there are many valid ways to measure a given concept. This theorem is important because it implies that the analyst's task is to choose one approach and then justify that that approach is a valid one.

9. ***What is an index variable?***

   An index variable is a variable that combines the values of other variables into a single indicator or score. Index variables are commonly used to empirically measure abstract concepts and multifaceted phenomena.

10. ***How are index variables constructed?***

    Index variables are constructed by summing the values of variables that measure distinct, though related, aspects of the concept. When a value of one or more measurement variable(s) is missing, the respective value of the index variable is also missing. (See Table 3.2 in the textbook for an example.)

11. ***Name four strategies for validating index variables.***

    First, analysts argue that their measurement variables are reasonable from a theoretical perspective (that is, they have content validity). Second, they show that, empirically, index measures have an appropriate range of values. Third, they show that the variables are correlated with each other (that is, they have high internal reliability). Fourth, index variables can be compared with other known measures, derived from external (other studies) or internal (the same study) sources. Respectively, these are referred to as criterion and construct validity.

12. ***What is Cronbach alpha? What values are acceptable?***

    Cronbach alpha, also called measure alpha, is a measure of internal reliability. This is the extent to which measurement variables are correlated with each other. Variables that measure the same concept are assumed to correlate with each other. Alpha values between 0.80 and 1.00 are considered good, values between 0.70 and 0.80 are acceptable, and values below 0.70 are poor and indicate a need for analysts to consider a different mix of measurement variables that measure their concept.

Critical Thinking

1. **Explain the following statement: “Scales should encompass all of the possible values that a variable can assume.”**

   ____

   ____

   ____

   ____

   ____

   ____

2. **Explain the following statement. “Continuous-level scales are preferred over ordinal-level scales, which in turn are preferred over nominal-level scales.”**

   ____

   ____

   ____

   ____

   ____

   ____

3. **Explain how measurement scales (for example, Likert scales) can affect the phrasing of survey questions.**

   ____

   ____

   ____

   ____

   ____

   ____

4. **The text states that “no correct number of dimensions or variables exist, only bad or**

lacking ones." Explain this statement.

____

____

____

____

____

____

5. **The text distinguishes three approaches to operationalization. When should the most rigorous approach be used? When should the least rigorous approach be used?**

____

____

____

____

____

____

6. **A study wishes to measure "citizen trust in government" through the number of lawsuits filed against the federal government. Evaluate the measurement validity of this approach.**

____

____

____

____

____

____

7. **Explain the theorem of the interchangeability of indicators.**

____

____

8. **Explain the following statement: “When one or more of the measurement variables are missing from an observation, the value of the index variables for that observation is missing, too.” What are the pros and cons of guessing the values of missing observations? Why should this not be done?**

9. **Explain the following concepts, and give an example of each: (1) face validity, (2) construct validity, (3) criterion validity, and (4) content validity.**

10. **Explain the following statement: “Analysts usually collect a few more variables than are minimally needed because they cannot know, prior to reliability analysis, which variable mix will have a sufficiently high alpha score to lend empirical support for the index measure.”**

Application Exercises

1. **Examine the citizen survey in the documentation for the Public Perceptions dataset, in Chapter 20 of this workbook. What level of measurement scale is used for the different items?**

2. **Give some examples of nominal-, ordinal-, interval-, and ratio-level variables in your area of interest.**

3. **An analyst wishes to measure public support for a new welfare program. (1) Develop some suitable measures using Likert items on a survey. (2) Show how incomplete, ambiguous, and overlapping scales create problems of measurement validity.**

____

4. **A survey of citizens assesses the extent to which they perceive that the federal government works democratically. A second study measures the extent to which the governments of different countries are democratic. Conceptualize democracy in each of these two study contexts.**

____

____

____

____

____

____

5. **Develop an index variable to measure “fear of statistics” among students in public and nonprofit management. Then develop an index variable to measure a concept in your area of interest.**

____

____

____

____

____

____

6. **Select a sample of six Government Accountability Office reports (see Box 2.1 in the text) or scholarly articles that use empirical data. Examine how these reports or articles address the matter of measurement validity.**

____

____

____

____

____

____

7. **Develop some measures that might be used in a study that assesses a neighborhood crime control program. Discuss some challenges of measurement validity as well as strategies for dealing with these challenges.**

____

____

____

____

____

____

Further Reading

Conceptualization and measurement are typically discussed in books on research methods and program evaluation, and readers are referred to those mentioned in Chapter 2 of this workbook.

It is instructive to consider articles that show different approaches to operationalizing study concepts. An example of the first approach to operationalization discussed in the text (measures of different dimensions combined in separate indices and subsequently aggregated into a “super” index) is XiaoHu Wang, Christopher V. Hawkins, Nick Lebredo, and Evan M. Berman, “Capacity to Sustain Sustainability: A Study of U.S. Cities,” *Public Administration Review* 72 (2012): 841–853. This study shows index measures of environmental sustainability practices, economic sustainability practices, and social sustainability practices. Each of these measures is composed of 10 or more survey items, and each of these index measures is subsequently aggregated into a super index of “sustainability.” A similar approach is taken on a very different topic in Evan M. Berman and Jonathan P. West, “Managing Emotional Intelligence in U.S. Cities: A Study of Public Managers,” *Public Administration Review* 68 (2008) 742–758. This article shows how “emotional intelligence” is conceptualized using four dimensions. The scientific literature is full of such examples on a broad range of topics, including environmental sustainability, organizational inclusion, neighborhood safety, and others. Many articles also validate their index measures such as through triangulation. You may want to research your library’s resources for articles that show indexes in your area of interest.

Chapter 4 Measuring and Managing Performance: Present and Future

Q & A

1. ***What is performance measurement?***

   Performance measurement provides a real-time assessment of what a program is doing, what resources it is using, and what it has accomplished recently. As an analytical process, it is designed to produce such information on an ongoing basis; it provides a snapshot that integrates important, frequently quantitative information about programs and policies. Performance measurement helps managers improve program monitoring and accountability and, by focusing on measurable results, improve program performance and stakeholder satisfaction, too.
2. ***How is performance measurement related to program evaluation?***

   Whereas program evaluation focuses on the past (what has a program or policy achieved?), performance measurement focuses on the present (what is a program or policy achieving?). Performance measurement developed from program evaluation. While thorough, program evaluation can be quite cumbersome and hence may produce information that is neither ongoing nor timely for management purposes. By contrast, performance measurement aims to be an up-to-date management information system.
3. ***What is the logic model?***

   The logic model is a way of conceptualizing program performance that shows relationships among inputs, activities, outputs, outcomes, and goals. (See Chapter 4 in the textbook for a schematic model.)
4. ***What is the difference between outputs and outcomes?***

   Outputs are the immediate, direct results of program activities. Outcomes are specific changes in behaviors or conditions that are measures of goal attainment.
5. ***What problem of measurement validity is mentioned in the text in connection with performance measurement, and how is it addressed?***

   Performance measures should avoid problems of inaccurate or incomplete measurement. In practice, performance measures do have these problems, and managers need to be clear about what their performance measures include and what they do not. Performance measures are best regarded as indicators only, to be used in conjunction with other, often qualitative information about programs and policies.
6. ***What is effectiveness?***

   Effectiveness is the level of results of a program or treatment. It is typically measured by one or more output or outcome measures.
7. ***What is efficiency?***

   Efficiency is the unit cost to produce a good or service. It is calculated as the output or outcomes over inputs, or O/I. Efficiency indicators can be calculated in different ways and should reflect program management concerns.

8. ***What are workload ratios?***

   Workload ratios are the ratios of activities over inputs, or A/I. For example, a workload ratio is the number of students in anger management courses per teacher providing such courses. Distinguishing between workload ratios and efficiency measures is important: a high caseload of clients does not mean that they are being served well.

9. ***What are benchmarks?***

   Benchmarks are standards against which performance is measured. Internal benchmarks are standards that organizations select based on what their own prior programs have achieved or on what they feel is appropriate. External benchmarks are standards that are set based on the performance of other organizations and programs.

10. ***What are equity measures?***

    Equity measures are used to compare performance across different groups. Often, output and outcome measures can be analyzed for different populations, types of organizations, and the like.

11. ***What is performance management?***

    Performance management is generally defined as activities to ensure that goals are consistently being met in an effective and efficient manner. Activities include using performance measures for improving accountability, service delivery, and managerial decision making, which is our focus here. Performance analysis is used in performance management to gain an understanding of program performance and the factors affecting it.

12. ***What is forecasting? How is it related to planning?***

    A forecast is a prediction about the future. This is sometimes also called a projection or prognosis. Forecasting is different from planning; whereas forecasting discusses what the future will look like, planning provides a normative model of what the future should look like. Planning often starts with forecasting to establish what the future is likely to look like in order to develop alternative futures or scenarios that might be preferred.

13. ***How are statistical methods used for forecasting?***

    Statistical methods typically describe and aim to extrapolate quantitative trends based on past and present data. Analysis can involve no more than the simple extrapolation of the past few data points, but it can also analyze complex cyclical patterns and model other variables affecting past and present levels.

14. ***What are judgment-based methods of forecasting?***

    Judgment-based methods of forecasting often use experts to assess the likelihood of futures occurring. Experts can be brought together in groups or as individuals. For example, the Delphi method is a forecasting method that asks experts to respond anonymously through several rounds of written surveys.

15. ***What are some key practices and standards for making forecasts?***

Forecasts are more reliable for shorter periods; forecasting should use multiple methods; data and experts should be as up-to-date and valid as possible; forecasts should use as much information as possible about the past, present, and future; assumptions and limitations should be stated clearly; the accuracy of forecasts should be determined wherever possible; forecasts of more complex methods are not always more accurate than simple ones; forecasting should begin by identifying a full range of possible future scenarios and events; forecasting should note unusual past events that affect past data and adjust forecasts or the data accordingly; and forecasters should expect their forecasts to be challenged.

Critical Thinking

1. **How is a system of key indicators such as performance measurement different from a system of comprehensive measurement?**

2. **Explain how performance measurement provides useful information about programs and policies, even if it is not free from measurement errors.**

3. **Is the number of arrests by police officers an activity, an output, or an outcome? Explain your answer.**

4. **Give an example of the distinction between efficiency and a workload ratio not**

mentioned in the text.

____

____

____

____

____

____

5. **Examine the measures in Table 4.1 in the textbook. Can you improve on these? Can you identify other measures? In what way might inaccurate or incomplete data affect these measures?**

____

____

____

____

____

____

6. **What practical problems can you foresee in using external benchmarks? Can these problems be overcome? If so, how?**

____

____

____

____

____

____

7. **Explain how equity measures are important, especially in the context of public management.**

____

8. **Explain how performance management uses performance measurement to improve program performance. Give a few examples from your area of interest.**

9. **Explain the following statement: “Experts can help identify future events that trend forecasting may overlook.” Give an example.**

10. **Discuss the following statement: “The accuracy of forecasts should be determined, such as by comparing predictions about the present against the observed reality of the present.”**

Application Exercises

1. **Identify outputs and outcomes of a program to increase high school student graduation rates through homework assistance for at-risk students.**

2. **Identify outputs and outcomes of a program to reduce traffic congestion by adding dedicated bus lanes (lanes that only buses can use).**

3. **Develop a complete performance measurement system for a program or policy in your area of interest. Identify inputs, activities, outputs, outcomes, and goals as well as measures of effectiveness, efficiency, and equity. See Table 4.1 in the textbook to get you started. You will want to add efficiency, effectiveness, and equity.**

4. **Research and compare examples of performance measurement in agencies in your area of interest. How similar are their measures? Can you explain the differences?**

5. **Look for examples of performance measurement and balanced scorecards on the Internet.**

6. **Identify some statistical forecasts in your area of interest. What assumptions do they make? How might they be improved?**

7. **Develop some scenario-based forecasts in your area of interest.**

Further Reading

Several books now exist on making logic models, such as Lisa Wyatt Knowlton and Cynthia C. Phillips, *The Logic Model Guidebook: Better Strategies for Great Results*, 2nd ed. (Thousand Oaks, Calif.: Sage 2013). For a practical discussion on performance measurement, see Harry P. Hatry, *Performance Measurement: Getting Results*, 2nd ed. (Washington, D.C.: Urban Institute, 2007). Many jurisdictions provide online training manuals, such as the one provided by the Office for Victims of Crime (www.ovcttac.gov/docs/resources/OVCTAGuides/PerformanceMeasurement/welcome.html). It is fair to say that performance measurement is now a basic competency, and applications also exist for nonprofit management, including Robert M. Penna, *The Nonprofit Outcomes Toolbox: A Complete Guide to Program Effectiveness, Performance Measurement, and Results* (New York: Wiley, 2011). The classic work is Harry P. Hatry et al., *How Effective Are Your Community Services? Procedures for Measuring Their Quality* (Washington, D.C.: Urban Institute, 1992). In recent years, performance measurement has grown into performance management, which makes use of measurement, of course, and there are many articles and a few books on the topic. Many more books can be found on performance management. A recent book that combines performance measurement and program evaluation is James C. McDavid, Irene Huse, and Laura R. L. Hawthorn, *Program Evaluation and Performance Measurement: An Introduction to Practice,* 2nd ed. (Thousand Oaks, Calif.: Sage, 2013).

Although many books discuss forecasting methods, most of these books are for business, for example, Rob J. Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice* (OTexts, 2013; www.otexts.org/fpp). Many policy analysis books include some chapters on forecasting, such as Dipak K. Gupta, *Analyzing Public Policy: Concepts, Tools, and Techniques*, 2nd ed. (Washington, D.C.: CQ Press, 2010). Another such book is William Dunn, *Public Policy Analysis: An Introduction*, 5th ed. (New York: Routledge, 2011). However, these chapters often provide little more than a very brief overview of the field.

Chapter 5 Data Collection

Q & A

1. ***What are administrative data, and for what purposes are they used?***

   Administrative data are generated in the course of managing programs and activities. Traditionally, administrative data are used to (1) ensure that resources are not misused, (2) monitor the status of activities, and (3) provide a record of what has been completed and accomplished. Today, administrative data are also collected to (4) meet the needs of performance measurement. These purposes may also be necessary for grant or contract compliance.
2. ***What are secondary data, and for what purposes are they used?***

   Secondary data are data that have been collected by other agencies for their own purposes but that are available to managers and may be relevant for their purposes. Secondary data can provide important information about communities and can be useful for needs assessment, benchmarking, and outcome measurement. Managers and analysts are expected to be familiar with the secondary data in their fields.
3. ***What types of surveys are mentioned in the text, and for what purposes are surveys used?***

   The four types of surveys are mail, Internet, phone, and in-person surveys. Surveys are commonly used in program evaluation research and, increasingly, performance measurement. Surveys are increasingly used when such knowledge needs to be quantitative, comprehensive, and systematic.
4. ***Discuss the pros and cons of different types of surveys. Why are phone surveys used increasingly?***

   Mail surveys allow for the most survey items, but the need for follow-up mailings increases the duration of data collection. Internet surveys have few survey items, and the lack of a sampling frame can pose problems. Phone surveys have important speed advantages but may have low response rates. In-person surveys offer the highest response rate but also carry the highest cost. Phone surveys are used increasingly because they can be completed in a short time and can be used to ask many questions. Cell phones are increasingly called in surveys, and U.S. regulations require that phone numbers be manually dialed by interviewers.
5. ***List some standards for writing survey questions.***

   Survey questions should be clear (unambiguous and specific) and easy to answer. They should avoid double-barreled and leading questions. And they should avoid negative statements.
6. ***Explain the limitations and uses of customer comment cards.***

   Customer comment cards generate samples that are typically not representative of all customers, and therefore they are not generalizable. They are useful, however, for obtaining feedback about problems that might need attention.

7. ***Which sources of data, other than administrative records and surveys, are mentioned in the text?***

   The chapter mentions trained observers, actors, experts, and focus groups.

8. ***What is a census?***

   A census is a survey or count (tally) of an entire group or population.

9. ***Why is obtaining a representative sample important? How is it different from a purposive sample?***

   Only representative samples allow for generalization to the population. Representative samples have a mix of characteristics similar to that of the population from which they are drawn, whereas purposive samples have an unrepresentative mix of characteristics (for example, “exemplary practices” surveys are often purposive samples). Some threats to validity for surveys are inadequate sampling frames and unrepresentative samples.

10. ***What is random sampling, and why is it important?***

    Random sampling is a sampling method whereby each member of the population has an equal chance of being selected for the sample. Random sampling is the most valid way of making representative samples.

11. ***Define sampling error. Do small or large samples have small sampling errors? Why?***

    The sampling error is the percentage by which sample findings vary in 95 of 100 repeated samples. Large samples better reflect population characteristics and thus have smaller sampling errors.

12. ***What is nonresponse bias?***

    Nonresponse bias occurs when the views of nonrespondents are different from those of respondents, thus affecting the generalizability of the sample.

13. ***Explain data coding, data input, and data cleaning.***

    Data coding is the process of preparing data (from pencil-and-paper surveys or electronic or other sources) for input into statistical software programs. Data input (also, data entry) is the activity of recording these data in statistical software programs. Data cleaning is the process of identifying and removing reporting and recording errors. Errors include mistyped values, errors that arise in the process of uploading, and other implausible values that have been recorded. It is common practice to assume that unexamined data usually contain various errors that must be identified and removed.

Critical Thinking

1. **Identify some examples of the use of administrative data.**

____

____

____

____

____

____

2. **Identify and discuss problems in the quality of administrative data.**

____

____

____

____

____

____

3. **Identify some examples of the use of secondary data.**

____

____

____

____

____

____

4. **Identify and discuss problems in the quality of secondary data.**

____

____

____

____

____

____

5. **Contrast the use of systematic surveys and focus groups. For which purposes is each best suited? Give some examples.**

____

____

____

____

____

____

6. **Consider the following statement: “Our city is larger than the city next door, so we need a larger sample size for doing our citizen survey.” Is this statement valid? Why or why not?**

____

____

____

____

____

____

7. **Discuss the following statement: “Over 50 items can be asked when questions (in phone surveys) are easy and asked in a similar format.” Find an example on the companion website for this workbook.**

____

____

____

8. **Explain the concept of sampling error and why it matters.**

9. **What practical problems do you see with drawing a random sample of all Americans as opposed to a random sample of those that live in your city?**

10. **Discuss why customer comment cards do not constitute a generalizable customer satisfaction survey.**

11. **Identify threats to validity arising from biased questions and sampling in surveys.**

____

____

____

____

____

____

12. **Explain the following statement: "It is common practice to assume that unexamined data usually contain various errors." Give some examples.**

____

____

____

____

____

____

13. **Identify four specific ways in which data that have not been thoroughly cleaned may be problematic.**

____

____

____

____

____

____

Application Exercises

1. **Identify administrative data in your field of interest. How can such data be used for managing programs?**

2. **Critique the validity of administrative data in your field of interest. How might the quality of administrative data be improved?**

3. **Identify secondary data in your field of interest. For what purposes might these be used?**

4. **Discuss how trained observers, actors, experts, and focus groups might be used in your field of interest.**

____

____

____

5. **Identify how client and citizen surveys might be used in your field of interest.**

____

____

____

____

____

6. **Use the Internet to find examples of client and citizen surveys in your field of interest. Then evaluate the (1) validity and (2) usefulness of these surveys.**

____

____

____

____

____

____

7. **Evaluate and improve the following survey question: "Please tell me whether you like living at your present location or would consider moving to another city."**

____

____

____

____

____

8. **Develop some other poor survey questions, and then improve them. Use a Likert scale (see Chapter 3 in the textbook).**

____

9. **Develop a survey of clients, citizens, or employees in your field of interest.**

10. **Examine the methods section in the survey report on the companion website for this workbook. Examine carefully how concerns about representative samples are dealt with.**

From Data Coding to Data Analysis

Below is a checklist for getting your data in order.

1. Does the dataset include all the observations (for example, respondents)?
2. Does the dataset include all the variables?
3. Has a sample of the dataset been compared against the actual observations for accuracy?
4. Do the variable names make sense?
5. Does each variable have a label (if needed)?
6. Does each variable value have a label (if needed)?
7. Have variable values that indicate a missing value been coded as missing?
8. Has each variable been checked for implausible values (including outliers)?
9. Have implausible values been corrected or omitted?
10. Do the ranges of variables make sense? Which variables have limited ranges, and what implications follow for subsequent analysis?

Further Reading

The leading text on survey research is Don Dillman, Jolene D. Smith, and Leah Melani Christian, *Internet, Mail and Mixed-Mode Surveys: The Total Design Method,* 4th ed. (New York: Wiley, 2014). For conducting citizen surveys for administrative purposes, see Thomas Miller and Michele Miller, *Citizen Surveys for Local Government: A Comprehensive Guide,* 3rd ed. (Washington, D.C.: International City/County Management Association, 2008). A book on public opinion polling from a political perspective is Herbert Asher, *Polling and the Public: What Every Citizen Should Know,* 8th ed. (Washington, D.C.: CQ Press, 2010). A classic, older book on issues with secondary data sources is Mark Maier, *The Data Game,* 3rd ed. (Armonk, N.Y.: M. E. Sharpe, 1999). Despite the publication date, this book contains authoritative discussions about the quality of secondary data in different fields of interest, such as demography, housing, health, education, crime, the economy, labor, and business. These chapters are essential reading for anyone dealing with secondary data in these specific fields.

Few articles or books deal with data cleaning. Several sources are available on the Internet, such as from BetterEvaluation, http://betterevaluation.org/evaluation-options/data_cleaning. Some books on data mining, which deals with getting information from large databases (big data), discuss the importance of data cleaning. Some scholarly articles draw attention to problems with data, though these books often are quite sophisticated. In the end, the message is clear: we need to get our data in shape by trying to detect errors.

Chapter 6 Central Tendency

Note to students: This chapter includes questions and exercises relating to the textbook introduction to Section III (Descriptive Statistics), indicated by SI (Section Introduction).

Q & A

1. ***What are descriptive statistics? Give some examples. (SI)***

   Descriptive statistics provide summary information about variables, such as their average and frequency distribution.
2. ***What is the distinction between univariate and bivariate analysis? (SI)***

   Univariate analysis describes single variables, whereas bivariate analysis examines the relationship between two variables.
3. ***What are some important tasks of analysts engaged in statistics?***

   Some important tasks of analysts engaged in statistics are as follows: (1) understanding the definition and purpose of a statistic, (2) ensuring that a statistic is appropriate to the data and problem at hand, (3) understanding the test assumptions of a statistic, (4) applying a statistic to the problem at hand in ways that are mindful of the preceding points, (5) drawing correct conclusions, and (6) communicating results in ways that are appropriate for both professional and general audiences.
4. ***What are variables? What are scales?***

   Variables are succinctly defined as empirically observable phenomena that vary (see Chapter 2 in the text). Scales are the collection of specific attributes (or values) used to measure a specific variable (see Chapter 3 in the text). There are four levels of measurement scales: nominal, ordinal, interval, and ratio. You need to be familiar with these concepts, as they are key to choosing the correct statistic.
5. ***What role does the measurement level play in univariate analysis?***

   The type of univariate statistics that should be used depends on the level of measurement.
6. ***Name the three measures of central tendency. How is each defined?***

   The three measures of central tendency are mean, median, and mode. The mean is the sum of a series of observations, divided by the number of observations in the series. The median is the middle value in a series (or array) of values that have been ordered from low to high. The mode is the most frequent (typical) value(s) of a variable.
7. ***How should analysts deal with the problem of missing data in calculating statistics?***

   The most common approach is to exclude such observations from calculations.
8. ***What is a weighted mean? For what purposes is it sometimes used?***

   The weighted mean is defined as a mean for which the observations have been given variable weights. Weighted means are commonly used to adjust for over- and undersampling in surveys, for example.
9. ***What is the formula for determining the location of the median?***

The location of the median is determined by the formula (n + 1)/2. For example, if there are 97 observations, the median is the value of the 49th observation, when observations have been ordered. When there are 98 observations, the median is the mean of the 49th and 50th observations.

10. ***When should both the mean and median be used? When should the mode be used?***

The median should be reported along with the mean when a few very large or very small observations affect the value of the mean. The mode is used infrequently, but an advantage of the mode is that it can be used with nominal-level data, which is not possible for calculating the mean or median.

Critical Thinking

1. **Give some examples of univariate and bivariate analyses. (SI)**

____

____

____

____

____

____

2. **Why is the mean frequently used?**

____

____

____

____

____

____

3. **Consider the following statement: “Calculating the mean is straightforward, but managers and analysts may encounter some practical issues that, for the most part, concern the data rather than the formula itself.” Give examples of these practical issues.**

____

____

____

____

____

____

4. **Why are observations with missing values typically removed before calculating specific**

statistics?

5. **Discuss the challenge of using the mean for calculating the central tendency of an ordinal-level variable (for example, a survey question that uses a Likert scale; see Box 3.1 in the text).**

6. **Explain the following statement: “The median should always be used when a few very large or very small values affect estimates of the mean.” Give some examples of variables for which the median is typically used.**

7. **Explain the following statement: “An advantage of the mode is that it can be used with nominal-level data, which is not possible for calculating the mean or median.”**

Data-Based Exercises

1. **Although computer software is used to calculate statistics, some students find that hand calculation furthers their conceptual understanding. Consider the following values: 4, 5, 7, 9, 11, 13, 13, 16, 18. Calculate the mean, median, and mode. After completing the exercise below, you might return to your results and verify them using the computer. (To practice, you can make up your own data, too.)**

   Mean: $\sum x_i/n=$ ____

   Median: Value of observation at location (n + 1)/2:

   ____

   Mode: Most frequent value: ____

2. **This is an exercise in data input. Data input is a skill that is specific to the software package you are using. Chapter 19 provides data input examples for SPSS. Input the data shown in the workbook Screen W19.4, and use SPSS to calculate the mean, median, and mode. Verify your results with those shown in Screen W19.8.**
3. **This is an exercise in data coding and input. Input the data from Table W6.1 into your statistical software program. For simplicity, the table shows only three respondent records of the three survey items. Note that respondent 3 did not answer the first question. This needs to be coded as missing. Code this value as a 9 and instruct your statistical software to treat this value as missing. Also, create variable and value labels, as shown in Screens W19.3 and W19.4.**

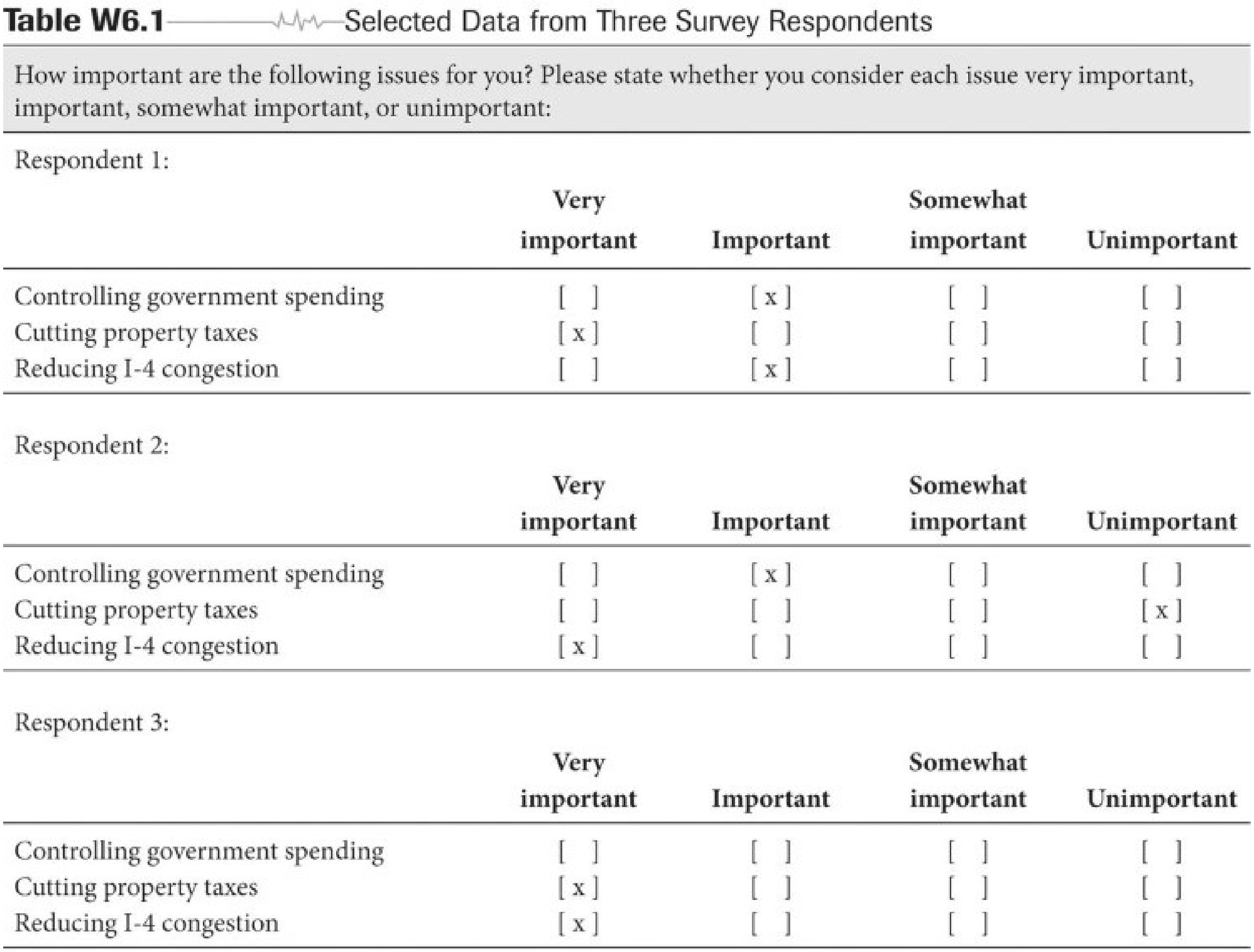


Table W6.1 Selected Data from Three Survey Respondents

How important are the following issues for you? Please state whether you consider each issue very important, important, somewhat important, or unimportant:

Respondent 1:

	Very important	Important	Somewhat important	Unimportant
Controlling government spending	[ ]	[x]	[ ]	[ ]
Cutting property taxes	[x]	[ ]	[ ]	[ ]
Reducing I-4 congestion	[ ]	[x]	[ ]	[ ]

Respondent 2:

	Very important	Important	Somewhat important	Unimportant
Controlling government spending	[ ]	[x]	[ ]	[ ]
Cutting property taxes	[ ]	[ ]	[ ]	[x]
Reducing I-4 congestion	[x]	[ ]	[ ]	[ ]

Respondent 3:

	Very important	Important	Somewhat important	Unimportant
Controlling government spending	[ ]	[ ]	[ ]	[ ]
Cutting property taxes	[x]	[ ]	[ ]	[ ]
Reducing I-4 congestion	[x]	[ ]	[ ]	[ ]

4. **Read about the Community Indicators dataset in Chapter 20 of this workbook. Then open this dataset using your statistical software package (for example, SPSS). Compare the mean and median values of the number of murders (Murder) in these different cities. Are the values of the mean and median similar or dissimilar? Calculate the mean and median of other variables of your choice, too. In Chapter 7, we will take this analysis further by making frequency distributions.**
5. **A local government operates a small park near its city hall. The department uses the attendance to measure its workload in determining the budget for the park. The city manager has long believed that the park has too few visitors and that the department should reach out to more customers or its budget will need to be reduced. You are an analyst for the city. The city manager has asked you to prepare an analysis for the park to determine whether the budget request for the park is justified in the city's budget proposal. You randomly selected 25 days in the past year and calculated the park attendance data: 5, 3, 10, 1, 2, 3, 4, 3, 5, 100, 4, 3, 2, 4, 25, 150, 3, 3, 5, 4, 8, 7, 10, 15, and 30.**
    1. Use computer software (SPSS or Excel) to calculate the mean, the median, and the mode for the data. Write a paragraph to explain the meaning of these statistics. Do you recommend the use of the mean in your presentation? Why or why not? Do you recommend the use of the median or the mode? Why or why not?

2. (*optional*) Prepare a report to discuss recommendations to the city manager.

6. **Using the Data_FL_County dataset, what is the mean and median of conservation spending in 2008? Why is the mean so much greater than the mode? (In the dataset, the variable name is Cons08 and the variable label is 2008 Conservation Spending.)**

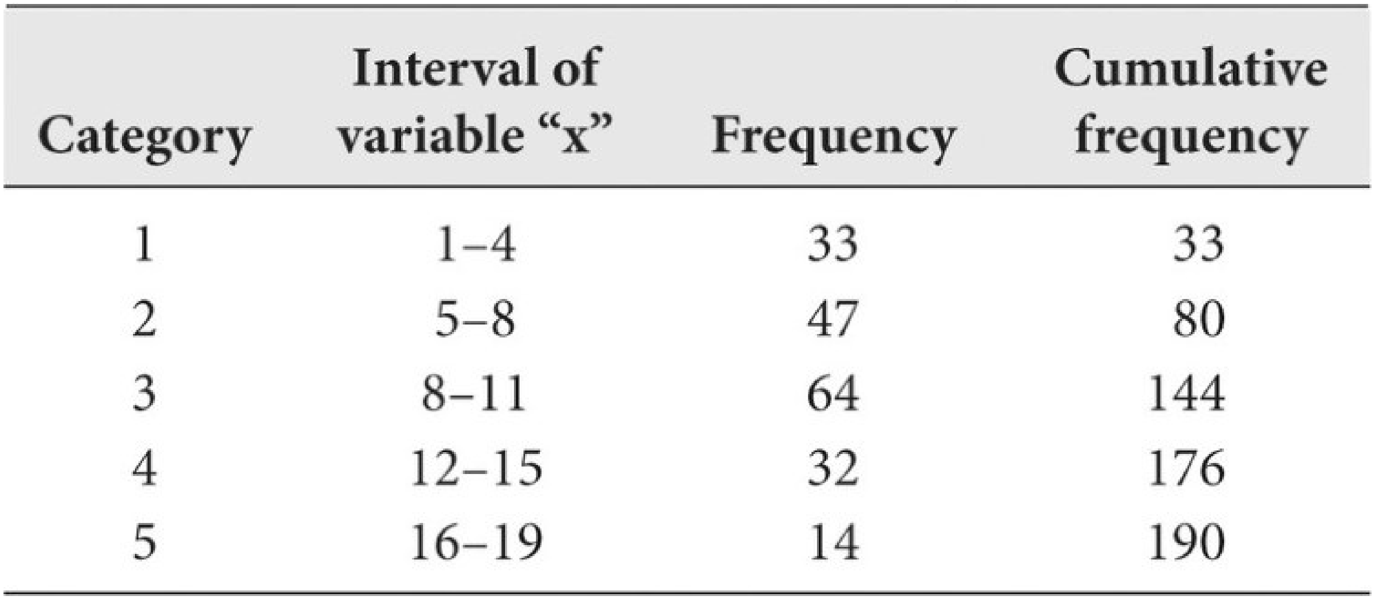
7. **Using the Community Indicators dataset, calculate the mean burglary rate per capita. To do this, you first need to create a new variable, which is the burglary rate per capita for each city in the dataset. Let's call this variable "burglaryrate" and define it as burglary/pop, whereby "burglary" and "pop" correspond to the variable names in the Community Indicators dataset. Use your statistical software program to calculate the mean and median burglary rates per capita. Which three cities have the highest burglary rates per capita? Are these the same cities as those with the largest number of burglaries?**

Note for students using SPSS: See Screen W19.2. Type “burglaryrate” (without quotes) in the target variable field, and type “burglary/pop” (without quotes) in the numeric expression field. Select OK and see the new variable created in your Data Editor screen. It will have been added at the end of the existing variables, all the way to the right in the Data View screen or at the end of the list in the Variable View screen. You may want to change the number of decimals shown to, say, five (5); see Screen W19.3 for this purpose.

8. **Open the Watershed dataset and read the description of this survey in Chapter 20. Calculate the mean and the median of the number of samples that exceed pollution standards (Conpolut) and fish and wildlife advisories (Advisory).**

9. **(*optional*) Consider the table of grouped data (Table W6.2). Calculate the grouped mean and grouped median (see Chapter 6 appendix in the text).**

Table W6.2 Frequency Table



Category	Interval of variable “x”	Frequency	Cumulative frequency
1	1–4	33	33
2	5–8	47	80
3	8–11	64	144
4	12–15	32	176
5	16–19	14	190

Grouped Mean

Step 1: Calculate the weighted mean of categories:

____

____

____

Step 2: Calculate the estimated value of the variable mean:

____

____

____

Grouped Median

Step 1: Determine the location of the median:

____

____

____

Step 2: Calculate the estimated value of the variable median:

____

____

____

Note: Chapter 7 contains more introductory data-based exercises, such as involving frequency charts and frequency distributions. Many analyses, including exercises in data cleaning, involve the techniques discussed in that chapter.

Further Reading

The statistics described in this chapter are quite basic. A fun, introductory book on statistics is Larry Gonick and Woollcott Smith, *The Cartoon Guide to Statistics* (New York: Harper Perennial, 2015). It does a nice job of explaining statistical concepts and includes many of the statistics described in subsequent chapters. Somewhat less useful is Murray Spiegel and Larry Stephens, *Schaum's Outlines: Statistics*, 4th ed. (New York: McGraw-Hill, 2014). It has chapters on descriptive statistics and other statistics discussed in this course, but the treatment is not very applied. Of course, you can consult many basic books on statistics, but few focus on public affairs. For example, H. Tokunaga, *Fundamental Statistics for the Social and Behavioral Sciences* (Thousand Oaks, Calif.: Sage, 2015) is written in an accessible style, and it emphasizes learning through understanding and application of mathematical calculations. Another option is Chava Frankfort-Nachmias and Anna Leon-Guerrero, *Social Statistics for a Diverse Society*, 7th ed. (Thousand Oaks, Calif.: Sage, 2014). The suggested readings in Chapter 2 related to the ethics of data analysis and presentation also provide useful advice on basic statistics, for example, Darrell Huff, *How to Lie with Statistics* (New York: Norton, 1993). An online source is "Getting Started with Statistics Concepts," www.statsoft.com/textbook/elementary-concepts-in-statistics.

A good exercise on the road toward mastering concepts and applications in statistics is to view their use in scholarly research and professional practice. For example, read some issues of leading academic journals and note their use of descriptive statistics. Some of the leading journals in public affairs and political science are *American Political Science Review, Public Administration Review*, and *Journal of Public Policy Analysis and Management*. Of course, many good scholarly journals are available, and you might ask your professor for a reading list or even for specific articles highlighting the uses of statistics discussed in this and subsequent chapters.

If you are using SPSS and you want more information than is offered in Chapter 19 of this workbook, try the most recent version of *SPSS Base User's Guide* (Chicago: SPSS) for whatever version of SPSS you are using. Or search online for "IBM SPSS Statistics 23 Brief Guide" or whatever version you are using (or call IBM at 800-543-2185). You can also try Earl Babbie and W. Wagner, *Adventures in Social Research: Data Analysis Using SPSS*, 9th ed. (Thousand Oaks, Calif.: Sage, 2015). There are also some excellent instructional videos on YouTube (e.g., search "introduction to SPSS").

Chapter 7 Measures of Dispersion

Q & A

1. ***What are measures of dispersion?***

   Measures of dispersion provide information about how the values of a variable are distributed.
2. ***What are frequency distributions? What are they used for?***

   Frequency distributions describe the range and frequency of values of a variable. They are used for nominal- and ordinal-level data. Frequency distributions often are a prelude for generating data tables and attractive graphics and are also used for data cleaning.
3. ***What is a histogram? How is it different from a bar chart?***

   A histogram shows the number of observations in different categories (or values) of the variable. Analysts can define the number (or widths) of categories that are used to group the different values of the variable. Bar charts are similar, but they show the number of observations for each different value of the variable. By convention, histograms are used for continuous variables, and bar charts are used for categorical variables.
4. ***Discuss the use of bar charts, pie charts, and line graphs.***

   Bar charts show the frequency of occurrences through stacks, which can be used to highlight the importance of categories (values). Bar charts are used with ordinal- and nominal-level variables. Pie charts typically are used to focus on equality: Who gets most (or least) of what? Pie charts are used with nominal-level variables. Line graphs are usually used for continuous variables, partly to avoid displaying a large number of bars.
5. ***What are outliers? How are they dealt with?***

   Outliers are analyst-defined observations with unusual values relative to other values in the data. Outliers are defined as observations whose values are either less than the inner fence or greater than the outer fence. Outliers may be the result of data-coding errors or reflect actual but unusual values in the sample. The textbook suggests that observations that are flagged as outliers generally should be retained when they are not coding errors, when they are plausible values of the variable in question, and when they do not greatly affect the value of the mean (of continuous variables).
6. ***What is the standard deviation?***

   The standard deviation is a measure of dispersion that is calculated based on the values of the data.
7. ***What statistical property makes the standard deviation a desirable statistic?***

   When data are normally distributed, 68.3 percent of the observations lie within ±1 standard deviation from the mean, 95.4 percent lie ±2 standard deviations from the mean, and 99.7 percent lie ±3 standard deviations from the mean.

8. ***How do analysts determine whether a variable is normally distributed?***

   Some analysts rely on a visual inspection, aided by a computer-generated curve that is superimposed over the histogram. Analysts also use measures of skewness and kurtosis to determine whether the shape of the observed curve is consistent with a normal distribution. Sample data are not expected to match a theoretical bell-shaped curve perfectly because of deviations due to chance selection.
9. ***What are standardized variables?***

   Standardized variables (also called z-scores) are variables that have been transformed such that their means are exactly 0 and their standard deviations are exactly 1 (or unity).
10. ***(optional) What is a boxplot? For what purpose is it used? (Chapter 7 appendix)***

    A boxplot is a graphical device that shows various measures of dispersion. Boxplots are useful for obtaining a quick, visual, preliminary understanding of data; they are also useful tools for data cleaning. Statistics associated with boxplots are calculated based on the location of data.

Critical Thinking

1. **Explain why frequency distributions are widely used.**

   ____

   ____

   ____

   ____

   ____

   ____

2. **Explain how frequency distributions can assist with data cleaning.**

   ____

   ____

   ____

   ____

   ____

   ____

3. **Explain the following statement: "When recoding variables or creating histograms, a practical question is how wide each category should be. To avoid perceptions of lying with statistics, a rule of thumb is that categories should be based on category ranges of equal length."**

   ____

   ____

   ____

   ____

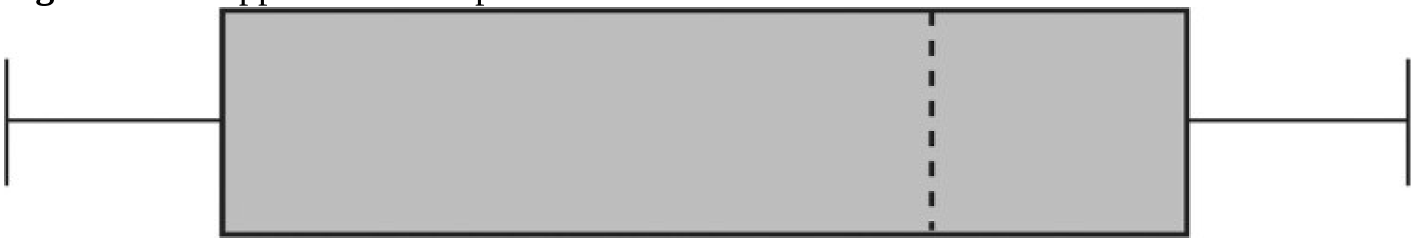
   ____

   ____

4. **Describe the difference between a histogram, and a bar chart.**

5. **Research some innovative ways of using graphs to get your points across.**
6. **What percentage of observations has a higher z-score value than 0.86? What percentage of observations has a lower z-score value than −1.15? What useful questions can a z-score statistic help answer? Hint: In the normal distribution table (see Appendix A in the textbook), look up values of the absolute difference between the above z-score values and 0.50 (|0.5 − z-score|).**
7. ***(optional)*** **What is the purpose of a boxplot? Is a boxplot fence an actual observation or a calculated number? What about a whisker? (Chapter 7 appendix)**
8. ***(optional)*** **Using Figure W7.1, draw a boxplot of a variable with these values: 4, 5, 7, 9, 11, 13, 13, 16, 18, 24. Are there any outliers? Why or why not? (Chapter 7 appendix)**

Figure W7.1 Application: Boxplot



Data-Based Exercises

1. **Although the computer calculates statistics, some students find that hand calculation furthers their conceptual understanding. Consider the following values: 4, 5, 7, 9, 11, 13, 13, 16, 18. (These were also used in workbook Chapter 6, Data-Based Exercise 1.) Now, calculate the standard deviation. You can use the computer to verify your results.**

Standard deviation: Σ (x i – x ¯) 2 n – 1 $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

x ¯ $\bar{x}$ ___________

x_1 – x ¯ $\bar{x}$: ___________ (x_1 – x ¯ $\bar{x}$)^2 : ___________

x_2 – x ¯ $\bar{x}$: ___________ (x_2 – x ¯ $\bar{x}$)^2 : ___________

x_3 – x ¯ $\bar{x}$: ___________ (x_3 – x ¯ $\bar{x}$)^2 : ___________

x_4 – x ¯ $\bar{x}$: ___________ (x_4 – x ¯ $\bar{x}$)^2 : ___________

x_5 – x ¯ $\bar{x}$: ___________ (x_5 – x ¯ $\bar{x}$)^2 : ___________

x_6 – x ¯ $\bar{x}$: ___________ (x_6 – x ¯ $\bar{x}$)^2 : ___________

x_7 – x ¯ $\bar{x}$: ___________ (x_7 – x ¯ $\bar{x}$)^2 : ___________

x_8 – x ¯ $\bar{x}$: ___________ (x_8 – x ¯ $\bar{x}$)^2 : ___________

x_9 – x ¯ $\bar{x}$: ___________ (x_9 – x ¯ $\bar{x}$)^2 : ___________

∑ (x_i – x ¯ $\bar{x}$)^2: ____________________________

Σ (x i – x ¯) 2 n – 1 $\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$: __________________________

2. **Standard deviation is a useful concept in performance management. It can be useful in performance comparison, performance monitoring, and performance evaluation. Let's say that a director in a local fire department wants to know any variation between the**

performance of this year and that of last year. He draws a sample of 10 response times from this year (in minutes): 3.0, 12.0, 7.0, 4.0, 4.0, 6.0, 3.0, 9.0, 11.0, 15.0, comparing them with a sample of 10 response times from last year (in minutes): 8.0, 7.0, 8.0, 6.0, 6.0, 9.0, 7.0, 9.0, 8.0, 6.0.

1. Does he see a performance variation by the mean?

   ____

2. Does he see a performance variation by the standard deviation? If he does, is it a performance improvement or a performance deterioration from the last year? Why?

   ____

   ____

3. Now, imagine that you are a citizen receiving fire protection services from this local fire department. How do you evaluate the response times of this fire department, by the mean, by the standard deviation, or by both?

   ____

   ____

4. If the average response time improves (shortens) but the standard deviation of the response times deteriorates (increases), what is your conclusion of the performance as a service recipient? In other words, do you want a quicker average response at the expense of a more unpredictable response? What is your recommendation to the director on which statistical measures he should use and how he should use them?

   ____

   ____

   ____

   ____

3. **This is an exercise in data cleaning. Prior to performing any type of data evaluation, you must first become familiar with the dataset and the nature of the variables you wish to evaluate. The first dataset is from a citizen survey performed in Orange County, Florida. Open the Public Perceptions dataset.**
   1. Read the description of this dataset, the methodology, the survey instrument, and notes regarding the variables and values.
   2. How many observations are contained in the dataset?

      ____

   3. How many variables are contained in the dataset?

      ____

   4. What are the measurement levels of the variables Gender, Age, and Yearsorc? Yearsorc measures the number of years that the respondent has lived in Orange County. Determining measurement levels is relevant to the selection of statistics later. The

measurement level can be determined in many ways, for example, by making a frequency chart and examining the category values.

Gender: ____

Age: ____

Yearsorc: ____

5. We now examine whether the present variable values are plausible (that is, whether they make sense). Make a frequency distribution for the variable Yearsorc. Are these values plausible?

   ____

6. (*optional*) To further validate your suspicions, make a boxplot of the variable Yearsorc. What do you conclude?

   ____

7. Let's remove this observation. It is quite unlikely that someone has actually lived 301 years in Orange County. It might be a coding error (or a practical joke), but in any event we are unable to obtain the correct value at this point. Note also that the revised boxplot still indicates some outliers. However, these are quite plausible values. We use our judgment and decide to retain them.
8. (*optional*) Run some boxplots on other variables to see if there are any outliers. Do you need to make any other changes?

   ____

4. **According to the Public Perceptions survey, do residents feel that the county has done a good job of balancing growth with environmental concerns? Make a frequency distribution. The variable is named Balance. Also, construct two bar charts: one that includes all three categories and one that omits the category "don't know."**

   ____

   ____

   ____

   ____

   ____

5. **According to the Public Perceptions survey, how do residents feel about their service experience? Specifically, what percentage of respondents agree or strongly agree that employees were helpful? What percentage of respondents agree or strongly agree that service was provided in a timely manner?**

   Note: Analyze data for only those respondents who stated that they have had contact with county employees during the last 12 months.

____

____

____

____

____

6. **Among the items listed in Section I of the Public Perceptions survey, which five items are the most important? Which items are the least important for residents? Based on what statistic do you decide this? What type of graph might you use to present your findings?**

____

____

____

____

____

7. **Open the Community Indicators dataset. Examine the frequency distributions of the variables murder and nonnegligent manslaughter (Murder), burglary (Burglary), and forcible rape (Rape). Then use boxplots to determine whether any cities might be considered outliers among these measures. Based on your findings, should you include or exclude these cities from any further analysis? If you include them, should you note and study the impact that these cities have on your results?**

____

____

____

____

____

8. **The Employee Attitudes survey contains several items that reasonably might affect workplace performance. Read the description of this survey. How do employees feel about the following items: the morale among county employees, the extent that their organization welcomes change, receiving timely feedback about performance, and cooperation among departments? The variables are, respectively, Himorale, Welchang, Feedback, and Coopdept.**

____

____

____

____

____

9. **Open the Watershed dataset and read the description of the survey. Calculate the mean and median of the number of samples that exceed pollution standards (Conpolut) and fish and wildlife advisories (Advisory). Make a bar chart of the quality of watersheds (Wshedch).**

____

____

____

____

____

Further Reading

The readings discussed in Chapter 6 should help with the material here, too. Basic statistics books will discuss standard deviations, bar charts, and the like. Another useful book is Mark L. Berenson, David M. Levine, and Kathryn A. Szabat, *Basic Business Statistics*, 13th ed. (Upper Saddle River, N.J.: Pearson Education, 2014).

Few separate articles or books used to deal with data cleaning, but increasingly more are found, such as Jason W. Osborne, *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data* (Thousand Oaks, Calif.: Sage, 2012). Some books that address data mining, which deals with getting information from large databases (big data), also discuss the importance of data cleaning. Some scholarly articles draw attention to problems with data, though these often are quite sophisticated. In the end, the message is clear: We need to get our data in shape by trying to detect errors.

Graphical displays are increasingly used in presentations, of course. This may be a good time to familiarize yourself with standards for making presentations (for example, with PowerPoint) and the use of graphics. See, for example, Cliff Atkinson, *Beyond Bullet Points*, 3rd ed. (Redmond, Wash.: Microsoft Press, 2011). In addition, you might want to visit a website like www.smartdraw.com to become familiar with graphics and other visual representations.

Chapter 8 Contingency Tables

Q & A

1. ***What are bivariate statistics? How do bivariate analyses help managers and policymakers?***

   Bivariate statistics is the study of relationships among two variables. Providing empirical information about relationships is a key step in determining program effectiveness, and bivariate statistics can help managers and policymakers by determining how variables affect key outcomes.
2. ***What is a contingency table?***

   A contingency table expresses the relationship between two categorical variables. One variable is shown in rows and the other in columns. The cells show the number (and, often, percentages) of observations associated with specific values of the two categorical variables.
3. ***How does causality affect the design of contingency tables?***

   When bivariate relationships between two categorical variables are causal, the independent variable should be placed in the columns and the dependent variable in the rows. When relationships are associations, there is no preference concerning the placement of variables. Although this convention often is violated, following it makes the analysis of tables easier when column percentages are used, as is commonly the case.
4. ***How are column percentages calculated? How are they used for examining relationships?***

   Column percentages are calculated by dividing each frequency by the column total. Relationships in contingency tables are usually examined by comparing column percentages across (groups of) rows.
5. ***What are marginal totals?***

   Marginal totals are the row and column totals.
6. ***What is a statistical definition of a relationship?***

   A statistical relationship means that as one variable changes, so, too, does another.
7. ***What is a positive relationship? What is a negative relationship?***

   Relationships involving ordinal or continuous variables are characterized as positive or negative. A positive relationship means that large values of one variable are associated with large values of the other variable and that small values of one variable are associated with small values of the other variable. A negative relationship implies the opposite: large values of one variable are associated with small values of the other variable, and vice versa.
8. ***What is a pivot table? How is it different from a contingency table?***

   Pivot tables show statistics of one or more continuous variables for one or more categorical variables in the data cells. By contrast, the cells of contingency tables show the number (and, often, percentages) of observations associated with specific values of the two categorical variables.

9. ***What is a layer variable?***

   A layer variable is one that defines the subset of data used for subsequent data tables.

10. ***What does the term* transposing *mean?***

    Transposing means interchanging the column and row variables; column variables become row variables, and vice versa.

Critical Thinking

1. **Identify all the features of contingency tables: title, clear column and row headings, data cell frequencies, column percentages, marginal totals, and grand total.**

2. **Explain the following statement: “Examining relationships in contingency tables is usually based on comparing column percentages for each or groups of rows.” Give an example.**

3. **Calculate column percentages for Table W8.1:**

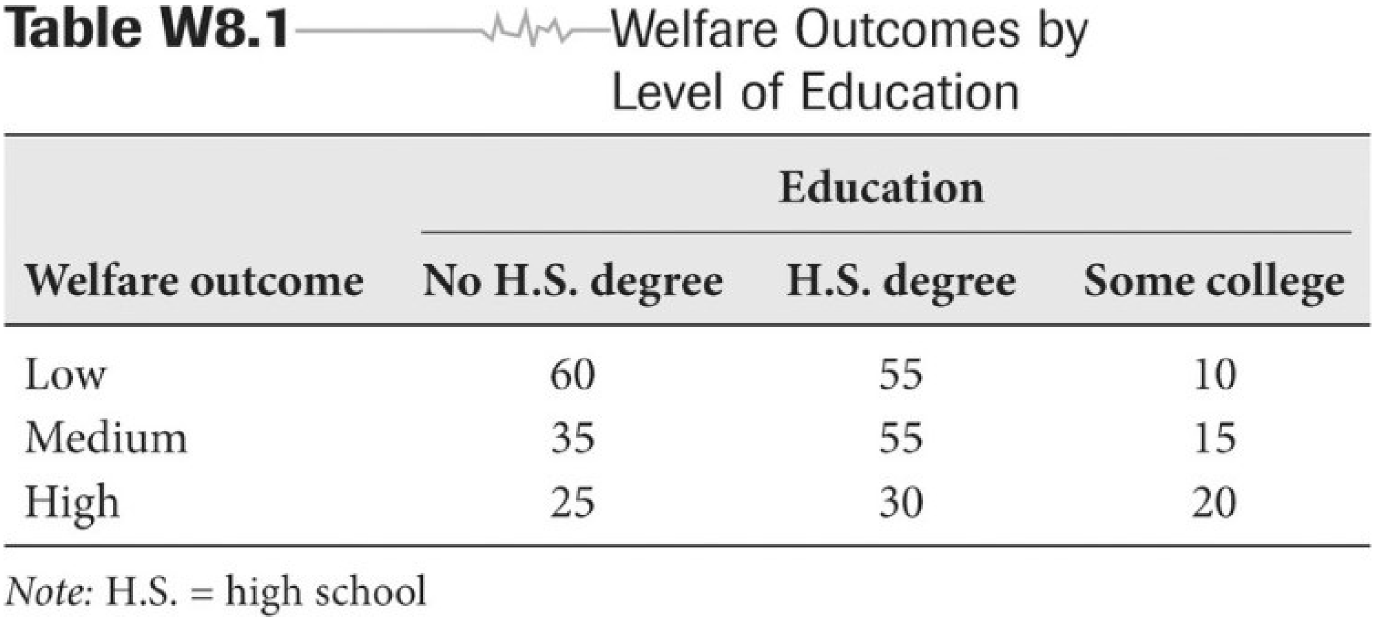


Table W8.1 Welfare Outcomes by Level of Education

	Education		
Welfare outcome	No H.S. degree	H.S. degree	Some college
Low	60	55	10
Medium	35	55	15
High	25	30	20

Note: H.S. = high school

Note: H.S. = high school

4. **Practice writing up the results for the preceding question. Identify the number of recipients and how they differ by education and welfare outcome. Then discuss the relationships between these two variables. Use statements in the form of "Whereas xx percent of recipients with no high school degree have high welfare outcomes, xx percent of recipients with some college have high welfare outcomes," and so on.**

5. **Explain how pivot tables and contingency tables are relevant to analysis of problems in your area of interest. Provide some examples.**

____

6. **Study the latest issues of research journals or professional reports in your field of interest; examine the use of tables and how the reports are written up. Record your findings on a separate sheet.**

Data-Based Exercises

Note to students: The following exercises are designed to give you practice making contingency tables and pivot tables. However, because the interpretation of contingency tables is often difficult without the use of the statistics presented in Chapter 11, some of the write-up and interpretation regarding these tables is postponed until that chapter.

1. **Use the Employee Attitudes dataset (see Chapter 20). Examine the relationship between gender and the morale of county employees (Himorale). Then consider the relationship between gender and stress (Stressed). Use column percentages, and write up the results.**

   ____

   ____

   ____

   ____

   ____

2. **Use the Public Perceptions dataset. Consider the relationship between gender (Gender) and trusting the county government to do what is right most of the time (Trust). Is the difference between men and women large or small? Is it meaningful to describe the relationship as being positive or negative?**

   ____

   ____

   ____

   ____

   ____

3. **Use the Public Perceptions dataset. The three most important issues (see Part I of the survey in Chapter 20) are as follows: helping public schools (Pubschl), fighting illegal drug use (Figtdrug), and dealing with the problems of gangs (Gangs). Do whites and nonwhites agree on the importance of these priorities? On which issues is there a difference?**

   ____

   ____

____

____

____

4. **Use the Employee Attitudes dataset. A manager wishes to examine the relationship between race (Race) and perceptions that the people who get promoted are among the best qualified (Bestqual) in the Public Works Department. However, very few employees are minorities, and a manager is concerned that separate analysis for each minority group might reveal their identity. Therefore, the manager wants to compare Caucasian employees against non-Caucasian employees. Recode the variable Race in this manner, and report on the above relationship.**

____

____

____

____

____

5. **Use the Employee Attitudes dataset. Examine the relationship between the morale of county employees (Himorale) and being satisfied with one's job (Satisjob). Is this table easy or difficult to interpret?**

   Note: Chapter 11 discusses Kendall's tau-c, which can also be used to interpret whether a relationship exists between these two variables in a contingency table and, if so, whether it is positive or negative.

____

____

____

____

____

6. **Open the Watershed dataset, and make a pivot table from variables in that dataset. For example, consider region and drinking water impairment as the two categorical variables, and show the mean level of conventional pollutants in each data cell. You may choose other continuous variables as well.**

   Note: See footnotes in Chapter 8 of the textbook for information on how to create pivot tables in SPSS and Excel.

7. **Does class attendance affect a student’s academic performance? To answer this question, you can conduct a study based on your personal experience. Consider all the classes you have taken in college. Code as ordinal variables your attendance in a course (for example, good or not so good) and your academic performance (for example, A or not A). Conduct a contingency table analysis to examine the relationship between attendance (the independent variable) and academic performance (the dependent variable). Write a paragraph to report your finding.**

Further Reading

This chapter deals with contingency tables and pivot tables. Although no specialized texts cover these matters, user's manuals and published articles provide further reading. You might search online for "SPSS contingency tables," which will bring up some useful PDFs and YouTube videos on how to use contingency tables in SPSS. For a light conceptual introduction, check out www.khanacademy.org.

Many articles and reports show the use of contingency tables, and students can examine how these authors interpret and write up the results—almost every research journal has articles with contingency tables, of course. You can also visit the Government Accountability Office website for examples of reports with tables (see www.gao.gov/browse/date/week).

Chapter 9 Getting Results

Q & A

1. ***What are analyses of outputs and outcomes? Can you give an example?***

   These are efforts that define, calculate, and display output and outcome measures of a program or an effort. The text discusses, as an example, outputs and outcomes related to a job training program. Outputs include the number of training sessions taught, whereas outcomes refer to trainees who find jobs.
2. ***What are analyses of efficiency and effectiveness? Can you give an example?***

   These are efforts that define, calculate, and display efficiency and effectiveness measures of a program or an effort. As an example, we can compare the costs to graduate each student in two schools (that is, cost/number of students graduated) to identify which school is more efficient by this measure. Effectiveness refers to outcomes such as the number of students graduated.
3. ***What are analyses of equity? Can you give an example?***

   These are efforts that define, calculate, and display how program resources, efforts, and outcomes vary across groups. These analyses may also concern matters of equal treatment across groups. As an example, we might assess the gender participation rates in a local government workforce over time to identify the trend in equal access to work opportunity.
4. ***What are quality-of-life analyses? Can you give an example?***

   A quality-of-life measurement is a composite index that assesses a variety of aspects in service or life quality of a community. A quality-of-life analysis is the effort to define, calculate, and display measures of the quality-of-life index of a community. As an example, we can create a quality-of-life index measure for an urban area that includes outcome measures in service areas such as public safety, health care, transportation, environmental protection, education, and other essential public services and specify the trend of the index by examining the values of the index over time.
5. ***What are forecasts?***

   A forecast is a calculation or estimate of a future event. (For a general discussion on forecasting, see Chapter 4 in the text, especially, the six principles and practices of forecasting.)
6. ***Why and how should forecasts be validated?***

   Forecasts vary according to what methods are used. Therefore, multiple methods should be used and forecasts should assess how well the model can accurately predict current values; we cannot place much credence in a forecasting model that does a very poor job of predicting today’s known values. Regardless of this, forecasts are thought to be more certain for shorter periods.
7. ***Compare forecasting based on prior moving averages, prior moving changes, and***

known ratios.

Forecasting based on prior moving averages predicts future values based on the mean of preceding values. Forecasting based on prior moving changes predicts future values based on the immediately preceding value and the mean of preceding increases. Forecasting based on known ratios uses the value of one variable to predict another variable, assuming that both the ratio of variables and an accurate forecast of the other variable are known. Forecasts based on prior moving averages tend to be conservative.

Critical Thinking

1. **Explain the difference between performance measurement and performance management.**

2. **Give some examples of how statistical concepts and tools can enhance the application of performance measurement and performance management.**

3. **Consider the following expression: “What gets measured gets done.” Discuss the pros and cons of this approach to performance management.**

4. **What limitations would a manager face if he or she understood the concepts involved in**

performance measurement and performance management but did not have the ability to produce statistical analyses?

____

____

____

____

____

____

5. **How is the application of statistics in performance measurement and performance management enhanced through the use of charts (that is, graphs and tables)? Conduct a simple experiment in which the results of a statistical analysis are presented to an audience with and without the use of charts. What differences does the presentation of charts make? Besides the visual enhancement provided, what other benefits do you see from presenting information using graphs and tables?**

____

____

____

____

____

____

6. **How is your understanding of statistical concepts enhanced by studying the analyses of outputs, outcomes, efficiency, and effectiveness in this chapter?**

____

____

____

____

____

____

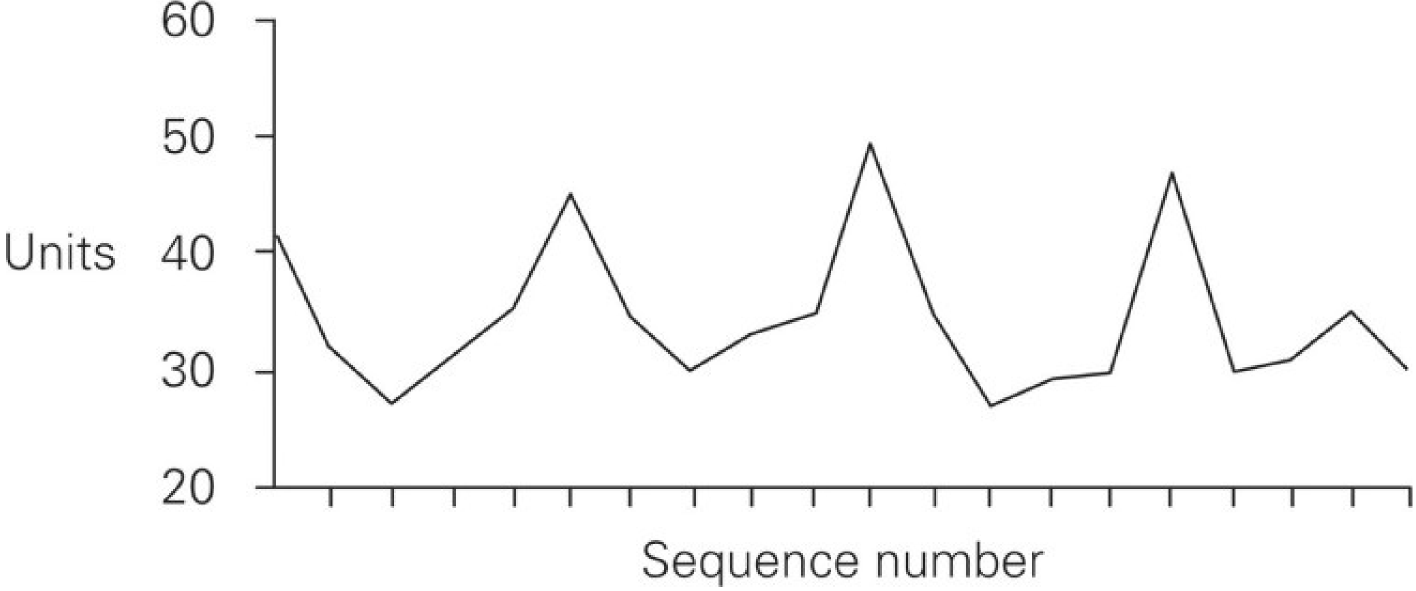
7. **Is your understanding of statistical concepts enhanced by studying the quality-of-life analyses in this chapter? Why or why not?**

8. **Explain the difference between prediction and forecasting.**

9. **Discuss why prior moving averages forecasting may be more conservative than forecasting based on prior average changes.**

10. ***(optional)*** **Consider the trend shown in Figure W9.1 Should forecasting involve periodicity, and, if so, what period of time should be considered? What is the consequence of not using periodicity? (Chapter 9 appendix)**

Figure W9.1 Trend



Data-Based Exercises

1. **The department of transportation in a state government has four vehicle registration offices in a metropolitan area. The department recently conducted a series of studies to determine the number of mistakes made in issuing or renewing driver's licenses (an outcome measure). Twelve tests were conducted in each of these four offices, and the results are shown in Table W9.1.**

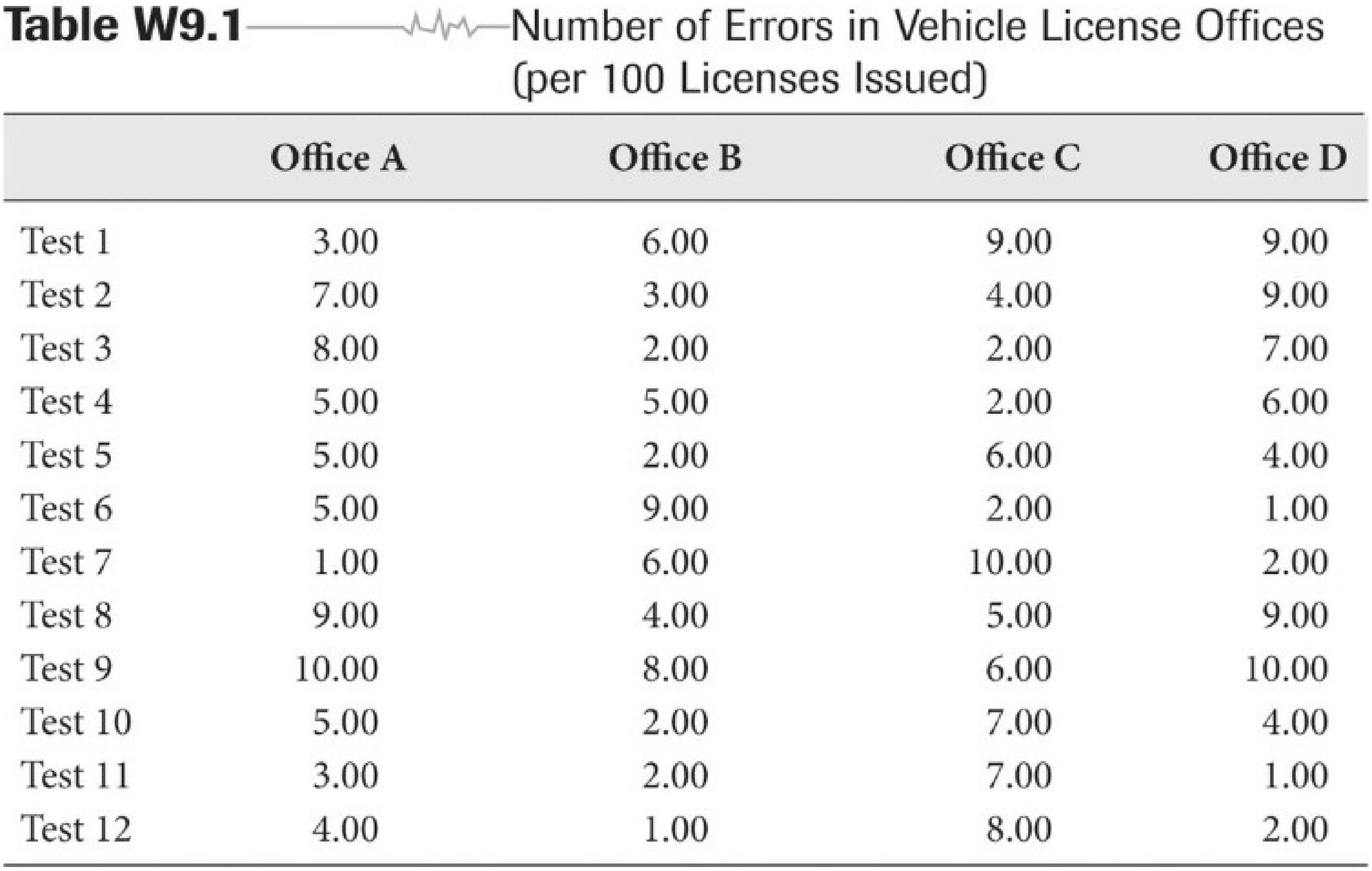


Table W9.1 Number of Errors in Vehicle License Offices (per 100 Licenses Issued)

	Office A	Office B	Office C	Office D
Test 1	3.00	6.00	9.00	9.00
Test 2	7.00	3.00	4.00	9.00
Test 3	8.00	2.00	2.00	7.00
Test 4	5.00	5.00	2.00	6.00
Test 5	5.00	2.00	6.00	4.00
Test 6	5.00	9.00	2.00	1.00
Test 7	1.00	6.00	10.00	2.00
Test 8	9.00	4.00	5.00	9.00
Test 9	10.00	8.00	6.00	10.00
Test 10	5.00	2.00	7.00	4.00
Test 11	3.00	2.00	7.00	1.00
Test 12	4.00	1.00	8.00	2.00

Calculate the mean, the sample variance, the sample standard deviation, the maximum, the minimum, and the range for each office. Discuss the meaning of each of these performance measures for the offices.

____

____

____

____

____

2. **Which office has the best performance, based on these measures? Which has the worst? What makes you draw these conclusions? If the state decides to retrain offices**

one at a time, what would be your priority list for retraining? Explain your rationale for creating this list.

____

____

____

____

____

3. **Calculate the frequency and cumulative distributions for the number of errors for each office. Let's say that the department's goal is for at least 50 percent of each office's tests to find no more than five errors. Which office or offices meet this performance benchmark, and which don't? Explain your findings.**

____

____

____

____

____

4. **Use five or fewer errors as a benchmark to categorize the test scores for each office (that is, with categories of "five or fewer errors" and "more than five errors"). Perform a contingency table analysis for all four offices. Discuss your findings.**

____

____

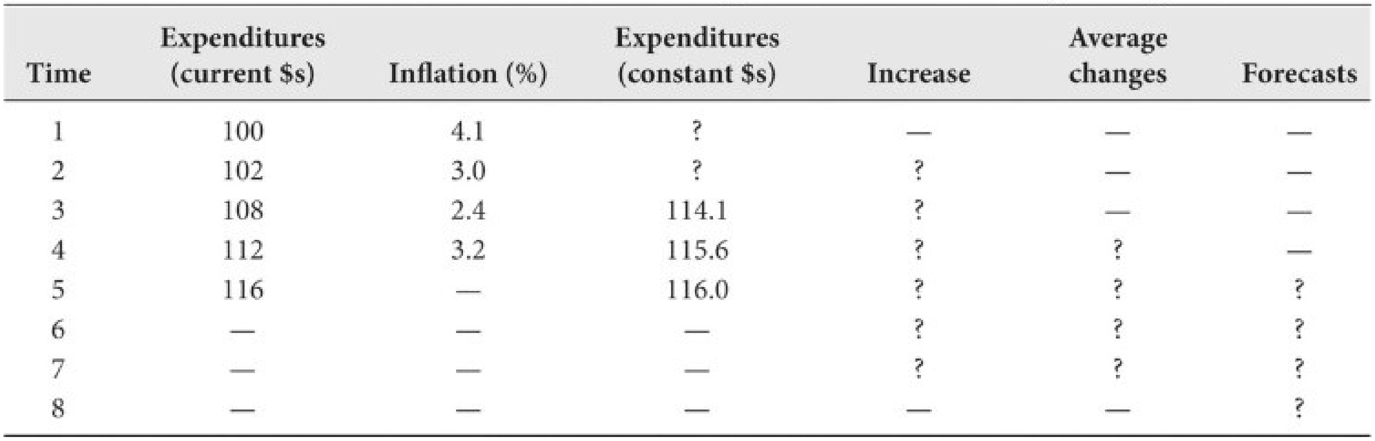
____

____

____

5. **Table W9.2 shows expenditures for the past 5 years. Forecast the next three periods using (1) prior moving averages (3-year spans) and (2) prior average moving changes (with last period as base). Add your forecast to the table.**

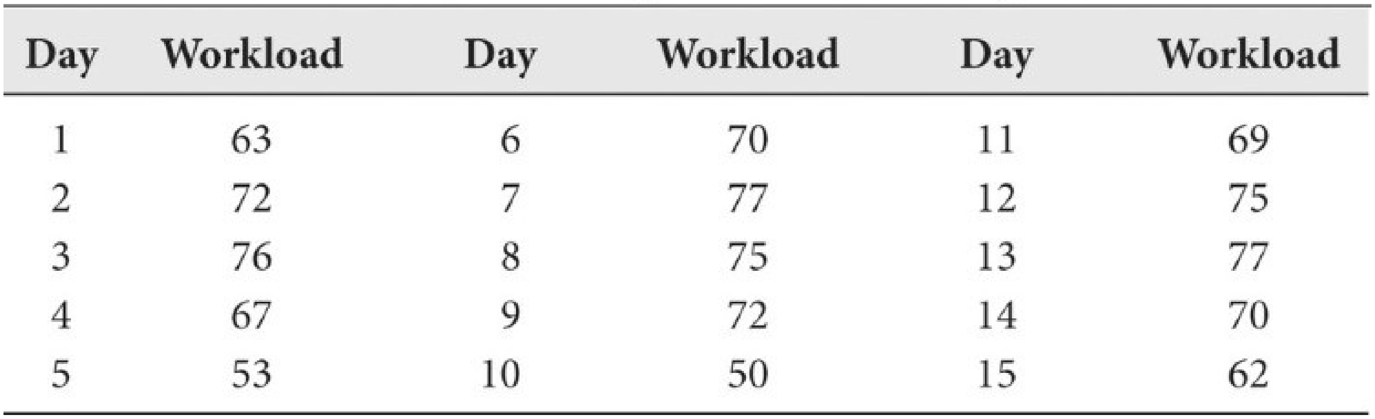
Table W9.2 Worksheet for Forecasting Based on Prior Moving Changes



Time	Expenditures (current $s)	Inflation (%)	Expenditures (constant $s)	Increase	Average changes	Forecasts
1	100	4.1	?	—	—	—
2	102	3.0	?	?	—	—
3	108	2.4	114.1	?	—	—
4	112	3.2	115.6	?	?	—
5	116	—	116.0	?	?	?
6	—	—	—	?	?	?
7	—	—	—	?	?	?
8	—	—	—	—	—	?

6. **Daily workload activities are shown in Table W9.3. Forecast workloads for Week 6.**

Table W9.3 Daily Workload Activity



Day	Workload	Day	Workload	Day	Workload
1	63	6	70	11	69
2	72	7	77	12	75
3	76	8	75	13	77
4	67	9	72	14	70
5	53	10	50	15	62

7. **Practice making a few professional-looking charts using Microsoft Excel. Excel allows analysts to create different types of charts, including column charts, bar charts, pie charts, and line charts. To create a chart in Excel, click on the Insert tab and then select the chart type from the Charts group. You can modify charts, apply predefined styles and layouts, and add various formatting features. Chapter 18 of this workbook provides an example of how to create a chart using Excel.**
8. **Table W9.4 shows attendance data of a public museum. Make a worksheet for calculating forecasts based on (1) prior moving averages and (2) prior moving changes for the next five periods. Also, validate your forecasts by comparing actual and predicted values.**

Table W9.4 Public Museum Attendance

Time	Attendance	Time	Attendance
1	1,205	9	1,468
2	1,309	10	1,602
3	1,325	11	1,625
4	1,226	12	1,698
5	1,450	13	1,550
6	1,529	14	1,623
7	1,679	15	1,708
8	1,543	16	1,767

Further Reading

See Chapter 4 of this workbook for many good sources regarding the performance measurement and management literature. For more on the application of performance measurement, see David N. Ammons, *Municipal Benchmarks: Assessing Local Performance and Establishing Community Standards*, 3rd ed. (New York: Routledge, 2012). For the application of statistical tools in performance management, see XiaoHu Wang, *Performance Analysis for Public and Nonprofit Organizations* (Burlington, Mass.: Jones and Bartlett, 2010) or, on financial management, see XiaoHu Wang, *Financial Management in the Public Sector: Tools, Applications and Cases,* 3rd ed. (New York: Routledge, 2014).

Chapter 10 Introducing Inference: Estimation from Samples

Q & A

Note to students: This chapter includes questions and exercises relating to the textbook introduction to Section IV (Inferential Statistics), indicated SI (Section Introduction).

1. ***What are inferential statistics? (SI)***

   Inferential statistics are statistics used to make inferences about characteristics in the population from which sample data were drawn. For example, inferential statistics are used to determine, on the basis of a sample, whether a relationship exists in the population from which the sample data were drawn.
2. ***Name three types of bivariate tests. Explain when each should be used. (SI)***

   There are many statistical tests for bivariate data. The type of data (level of measurement) determines which statistical test should be used. The section introduction discusses the following:
   - When two variables are categorical, tests based on contingency tables should be used.
   - When one variable is dichotomous (for example, gender) and the other is continuous, the t-test should be used.
   - When both variables are continuous, simple regression analysis should be used.
3. ***What are populations and parameters?***

   The population is the entire set of subjects in a study. Parameters are characteristics of the population, for example, the mean age of all the residents living in a country. Because population data can be difficult to obtain, samples are often used to estimate population parameters.
4. ***What are samples and sample statistics?***

   A sample is a subset of the population. Sample statistics are characteristics of a sample, such as the mean age of a sample of residents living in a city. Sample statistics are often used to estimate population parameters.
5. ***What is a confidence interval?***

   A confidence interval is the range within which the unknown but true population parameter is estimated to lie. (A confidence interval estimates the population parameters using sample statistics and is sometimes also expressed as the range within which a statistic is expected to fall on repeated sampling.)
6. ***What is a 95 confidence interval? Why is a 99 percent confidence interval always wider than a 95 percent confidence interval? Discuss the trade-off between the confidence and accuracy in estimation.***

   A 95 percent confidence interval is the range within which sample statistics such as the mean will fall in 95 of 100 samples, whereas a 99 percent confidence interval is the range within which sample statistics such as the mean will fall in 99 of 100 samples. A 99 percent confidence interval gives *more confidence* that the population parameter will fall within the calculated range, but that range is wider and hence *less accurate* than the range of a 95 percent confidence interval. (By analogy, the bigger a target, the more likely you will hit it. However, hitting a larger target does not mean that you are as accurate as a shooter hitting a smaller target!)

7. ***What is estimation error? Could it be zero?***

   Estimation error is a term that indicates the difference between the sample statistic and the population parameter. It is very rare to have an estimate error of zero, but it can happen.

8. ***What is a probability distribution?***

   A probability distribution is a statistical function which describes all possible values that a variable can take. Some examples are the normal distribution and the t-distribution.

9. ***What is the difference between the* t*-distribution and the normal distribution?***

   The *t*-distribution is used when the sample size is smaller than 30. Because *t*-values are very close to the *z*-values for large samples, *t*-values can be used for both small and large sample tests. (This is why *t*-distributions are more popular in statistical testing than the normal distribution.)

10. ***What does the Central Limit Theorem state?***

    The Central Limit Theorem is a theory that states that an infinite number of relatively large samples will be normally distributed, regardless of the distribution of the population from which they are drawn.

Critical Thinking

1. **Would you need to use inferential statistics when population parameters are known? Why or why not?**

2. **Can a population in one context be a sample in another context? For example, consider a class of 50 students—can this be both a sample and a population?**

3. **How can you improve the accuracy of a confidence interval estimation? (i.e., obtain a smaller interval. Tip: look at the formula!)**

4. **Can population parameters ever change?**

____

____

5. **Give a few real-life examples of when a confidence interval estimate might be used.**

____

____

____

____

____

6. **Give your own numerical example of the Central Limit Theorem.**

____

____

____

____

____

Data-Based Exercises

1. **The authority in a school district is very concerned about a recent newspaper report of the high absence rate among students in the district. The report says that the absence rate of the district may be well above the state average of 11.30 days a year per student. The authority wants to closely monitor the absence rate trend. To start, a sample of 250 students in the district was drawn to develop a baseline performance target for the performance tracking. The average absence rate of the sample is 11.90 days per student with a standard deviation of 3.90. Create a 95 percent and a 99 percent confidence interval to estimate the absence rate of the district.**

   ____

   ____

   ____

   ____

   ____

2. **A citizen satisfaction survey on police services was conducted in a city of about 100,000 residents. Of 1,934 residents in the sample who returned the survey, 895 said they were either "satisfied" or "very satisfied" with the services and 1,039 said that they were "dissatisfied" or chose "don't know." Create a 95 percent and a 99 percent confidence interval to predict the percentage of residents who are "satisfied" or "very satisfied" with the police services.**

   ____

   ____

   ____

   ____

   ____

3. **The number of vehicle accidents is sampled from a city's 20 major traffic intersections and their adjacent areas several times a day to assess traffic congestion. The police department uses the information to allocate resources of traffic control. A performance standard of an average of 2.0 accidents per hours per area is established. A traffic emergency control system is mobilized and resources are directed to traffic control once the traffic condition is equal to or worse than the standard. Table W10.1 shows the accident information taken from 7:00 am to 8:00 am today. Should the police department mobilize the traffic emergency control system? Why or why not?**

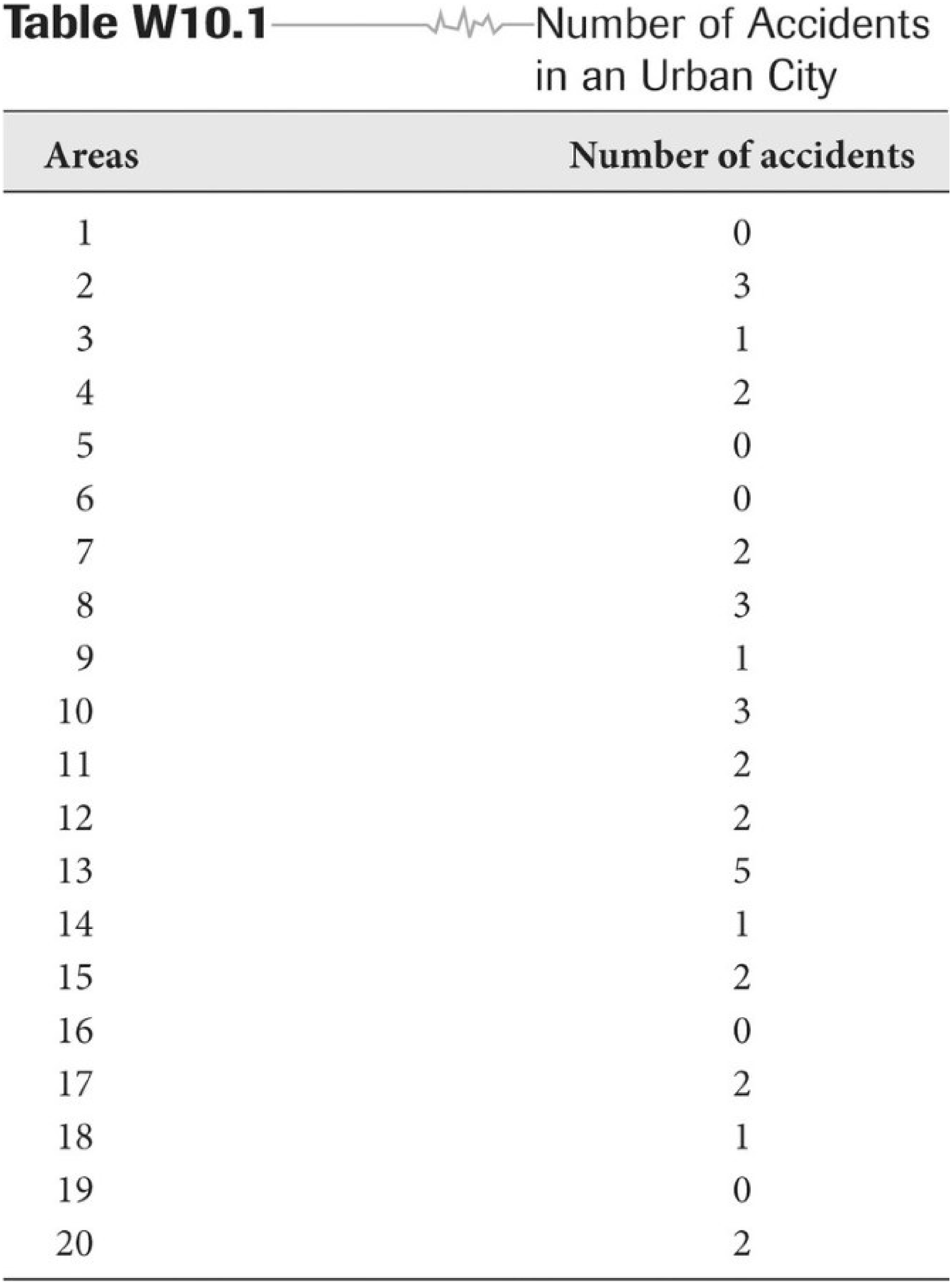


Table W10.1 Number of Accidents in an Urban City

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

4. **The dataset Data_FL_County (SPSS) includes Florida county governments' total spending, administrative spending, environmental spending, conservation spending, public safety spending, and economic development spending from 1999 to 2008. The data are from all Florida counties (except one which does not provide data). The**

dataset also includes demographic and sociopolitical variables—population, income, education, presidential votes, and many other variables. Use the county conservation spending in total spending in 2008 (ConsTotal08) to create a confidence interval. Explain the meaning of this estimate.

____

____

____

____

____

Further Reading

Additional exercises in public management and policy can be found in XiaoHu Wang, *Performance Analysis for Public and Nonprofit Organizations* (Sudbury, Mass.: Jones and Bartlett, 2010), Chapter 7, “Statistical Performance Monitoring.” In-depth discussion of probability and various probability distributions can be seen in Alan Agresti and Barbara Finlay, *Statistical Methods for the Social Sciences* (Harlow, UK: Pearson, 2008), Chapter 4, “Probability Distributions,” and Chapter 5, “Statistical Inference: Estimation.”

Chapter 11 Hypothesis Testing with Chi-Square

Q & A

1. ***Which five steps are followed in deciding the statistical significance of relationships?***
    1. State the null hypothesis (in Greek letters).
    2. Choose a statistical test.
    3. Calculate the test statistic (t.s.), and evaluate test assumptions.
    4. Look up the critical value (c.v.) of the test.
    5. Draw conclusion: If |t.s.| < c.v., do not reject null hypothesis. If |t.s.| ≥ c.v., reject the null hypothesis.
2. ***Explain the purpose of stating the null hypothesis.***

    The purpose of stating the null hypothesis is to establish a reasonable ground that a relationship exists. By assuming that a relationship doesn't exist, we need only to find a reasonable ground that it does exist, which is that it should be very unlikely to find a test statistic (in the sample) of a given (large) magnitude when in fact no relationship exists in the population. If we assume that a relationship does exist, we might be guilty of not trying hard enough to prove that it doesn't exist.
3. ***Discuss why notations in hypotheses are usually stated in Greek letters instead of Roman letters.***

    Greek letters are used because hypotheses refer to relationships in the general population, not in the specific sample.
4. ***What is the value of chi-square when two variables are unrelated to each other? What happens to the chi-square value as variables are more closely related to each other?***

    When two variables are (perfectly) unrelated to each other, the value of chi-square is (exactly) zero. The value of chi-square increases as two variables are more related to each other.
5. ***Define chi-square mathematically. Explain the concepts of "observed" and "expected" frequencies.***

    Chi-square is defined as ∑ i (O i – E i) 2 E i $$\sum_i \frac{(O_i - E_i)^2}{E_i}$$ where O_i is the observed frequency in a cell and E_i is the expected frequency in a cell. Observed frequencies are the actual counts of observations in cells. Expected frequencies are cell counts that are expected when no relationship exists between the variables.
6. ***What are the three test assumptions that must be satisfied in order to ensure that the chi-square test is valid?***

    First, the variables must be categorical. Second, the observations must be independent. Third, cells must have a minimum of five expected observations. (When the latter is not the case, it is usually because the contingency table contains a large number of rows and columns relative to the number of observations. That is, the data are spread too thin across too many cells. To correct this problem, adjacent rows or columns are combined to create a smaller number of cells.)

7. ***Explain the concept of statistical significance.***

   The concept of statistical significance relates to the following question: On repeated sampling, *how often* would I be wrong to conclude that a relationship exists when in fact it doesn't exist? The statistical standard in the social sciences is 5 percent; that is, we are willing to tolerate a 1-in-20 chance of concluding that the null hypothesis should be rejected when in fact it shouldn't. Standards of 1-in-100 (1 percent) are also used.

8. ***What is a critical value? Which two parameters determine it?***

   The critical value is the minimum value of a test statistic that is used to reject the null hypothesis. The critical value of any test statistic is determined by (1) the desired level of statistical significance and (2) the number of degrees of freedom (df).

9. ***What are the rules for determining whether a relationship is statistically significant?***

   We compare the absolute value of the test statistic against the critical value of the test statistic at a given level of significance. When the absolute value of the test statistic is greater than or equal to the critical value, we conclude that a relationship exists at the given level of significance. By convention, we report that relationships are not significant, are significant at the 5 percent level, or are significant at the 1 percent level. *Note:* Statistical software programs calculate the test statistic and report the level of statistical significance at which the test statistic is significant.

10. ***Explain how the sample size affects the level of statistical significance.***

    Relationships are more likely to be statistically significant when we are working with large datasets than when we are dealing with small ones. Having more information increases our confidence in our conclusions. The sample size affects the magnitude of many widely used test statistics, including chi-square.

11. ***What is the practical relevance of relationships? How is this different from statistical significance?***

    Hypothesis testing merely establishes whether a relationship is present. If a relationship is present, managers and analysts will want to describe it further in order to determine its importance for management action and public policy. This is called practical relevance (or practical significance). After the statistical significance of a relationship has been determined, analysts should ask the following questions: (1) What is the *direction* of the relationship; specifically, is it a positive or negative relationship? (2) By *how much* does one variable increase or decrease as a result of changes in the other? (3) What is the *strength* of the relationship?

12. ***What is the goodness-of-fit test?***

    The goodness-of-fit test is used to determine whether an ordinal distribution or test result is consistent with a predetermined norm.

13. ***13. What are nonparametric statistics?***

    Nonparametric statistics are a family of statistics which derive their name from the fact that they have very few test assumptions. They are a bit less powerful, but quite useful.

14. ***What is Kendall's tau-c?***

    Kendall's tau-c is a nonparametric alternative to chi-square. Chi-square provides no information about the direction and strength of a relationship, and it has some test assumptions, too. By contrast, Kendall's tau-c provides information about the level of significance as well as the strength and direction of relationships, and it does not have the assumption that cells must have a minimum of five expected observations. Hence, Kendall's tau-c is a useful alternative to chi-square.

15. ***(optional) Discuss the strategy for evaluating rival hypotheses in contingency table analysis. Also, define the terms* explanation, replication, specification, *and* suppression. *(Appendix 11.1)***

    The strategy for evaluating rival hypotheses in contingency table analysis involves taking control variables into account, discussed in the Appendix 11.1 in the text. Explanation is what occurs when statistically significant bivariate relationships are explained away (cease to exist) after adding the control variable. Replication is what occurs when statistically significant bivariate relationships remain significant after adding the control variable. Specification is what occurs when some statistically significant bivariate relationships are explained away after adding the control variable. Suppression is what occurs when a statistically insignificant bivariate relationship becomes significant after adding the control variable. Suppression is rare.

16. ***16. (optional) What is proportional reduction in error? (Appendix 11.2)***

    Proportional reduction in error (PRE) is the improvement, expressed as a fraction, in predicting a dependent variable from knowledge of the independent variable. PRE ratios range from 0.00 (no association or improvement in prediction) to 1.00 (perfect association or prediction). Although there are no absolute standards for PRE scores, many analysts regard scores of less than 0.25 as indicating a weak association, scores between 0.25 and 0.50, a moderate association, and scores above 0.50, a strong association.

Critical Thinking

1. **Why does it make sense to test for relationships by stating the null hypothesis?**

2. **Why are standards of 1 percent and 5 percent often used? What objections might be raised against using standards of, say, .01 percent or 10 percent?**

3. **Explain why the null hypothesis is rejected when the test value exceeds or is equal to the critical value.**

4. **Distinguish between statistical significance and practical relevance. Can you find some examples in your area of interest in which a statistically significant result might be**

practically irrelevant?

5. **Why must all cells in a chi-square test have a minimum of five expected observations? (See footnote 8 in Chapter 11 of the textbook.)**

6. **What is the critical value for a chi-square test with 13 degrees of freedom at the 1 percent level of significance? If the chi-square test statistic were 16.98, what would you conclude regarding the null hypothesis? What would you conclude if the chi-square value were 68.03?**

   Critical value:

   Conclusion if chi-square were 16.98:

   Conclusion if chi-square were 68.03:

____

7. **A statistical software program reports a test to be significant at p = .035. At what level of significance should this result be reported by the analyst (that is, 1 percent or 5 percent)? And what about p = .056, p = .0000, and p = .9989?**

   p = .035:

   ____

   ____

   p = .056:

   ____

   ____

   p = .0000:

   ____

   ____

   p = .9989:

   ____

   ____

8. **Formulate three null hypotheses in a study that examines the impact of community-based policing on neighborhood crime.**

   Null hypothesis 1:

   ____

   ____

   ____

   Null hypothesis 2:

   ____

   ____

   ____

Null hypothesis 3:

____

____

____

9. **Use Table W10.1 to design a contingency table of overall health and the number of days worked. Make up your own data. Identify the dependent variable, and be careful where you place it in the table. Show both percentages and frequencies.**

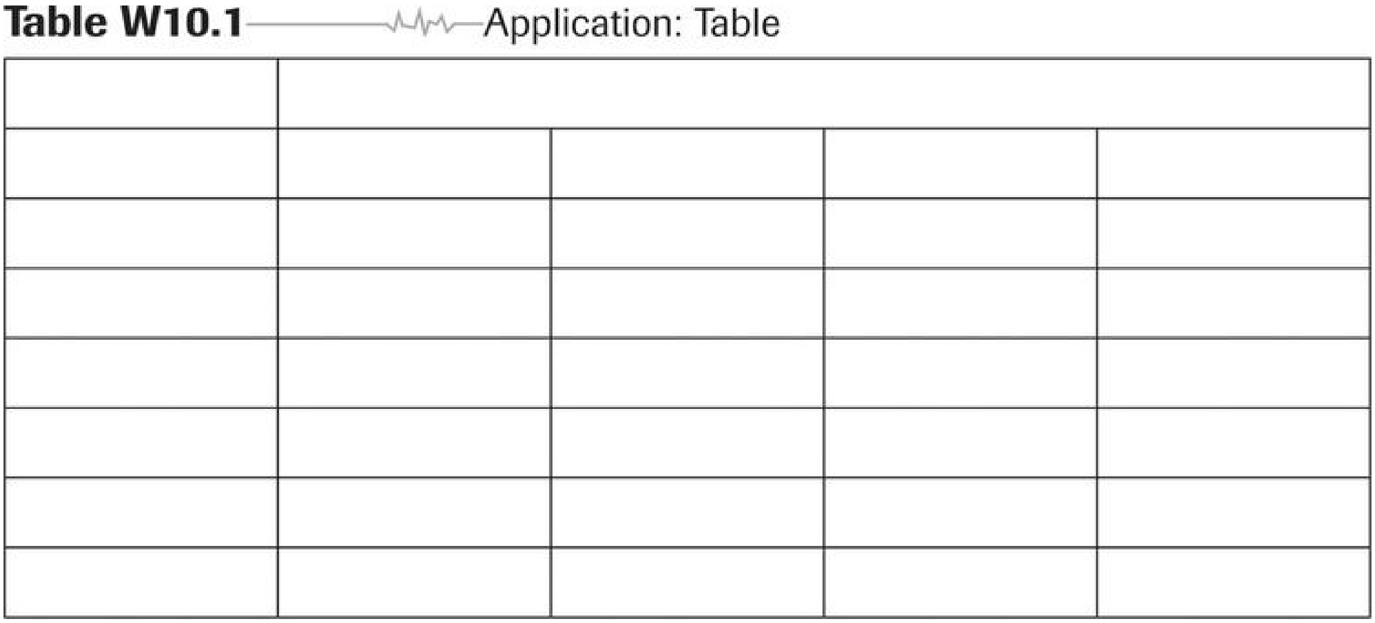


Table W10.1 Application: Table

10. **Calculate chi-square for the hypothetical data shown in Table W10.2, and determine whether the relationship is statistically significant. Welfare outcomes (for example, time to find a job) are examined as a function of the education levels of welfare recipients. Use Table W10.3 to calculate your expected frequencies. Use Table W10.4 to calculate the chi-square values for each cell.**

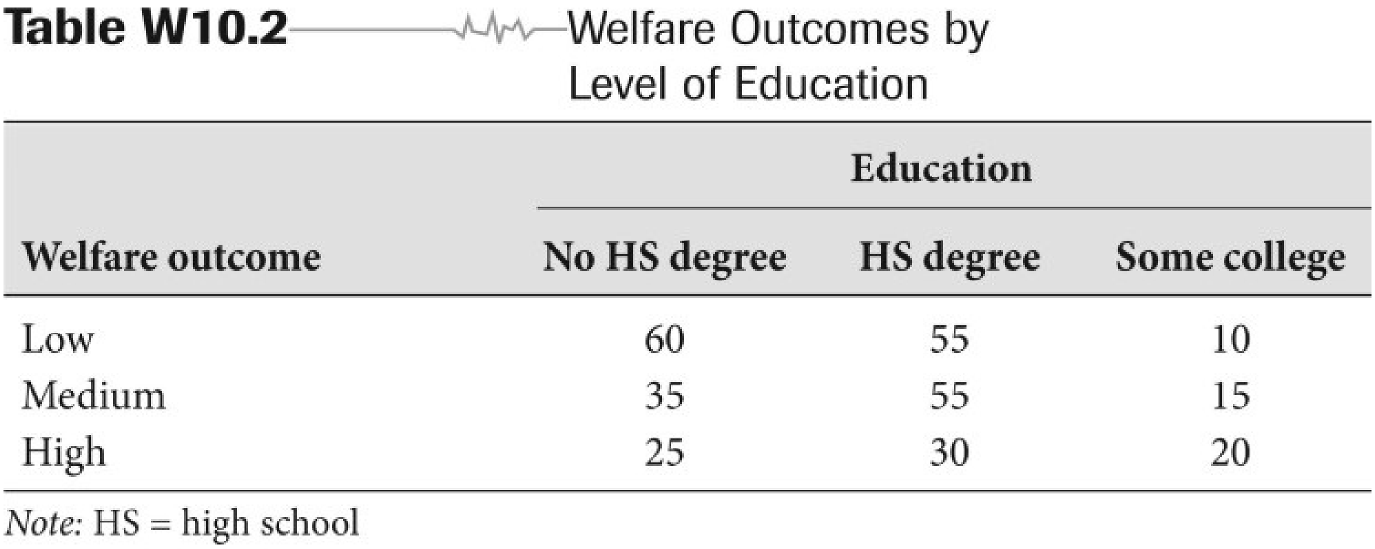


Table W10.2 Welfare Outcomes by Level of Education

	Education		
Welfare outcome	**No HS degree**	**HS degree**	**Some college**
Low	60	55	10
Medium	35	55	15
High	25	30	20

Note: HS = high school

Note: HS = high school

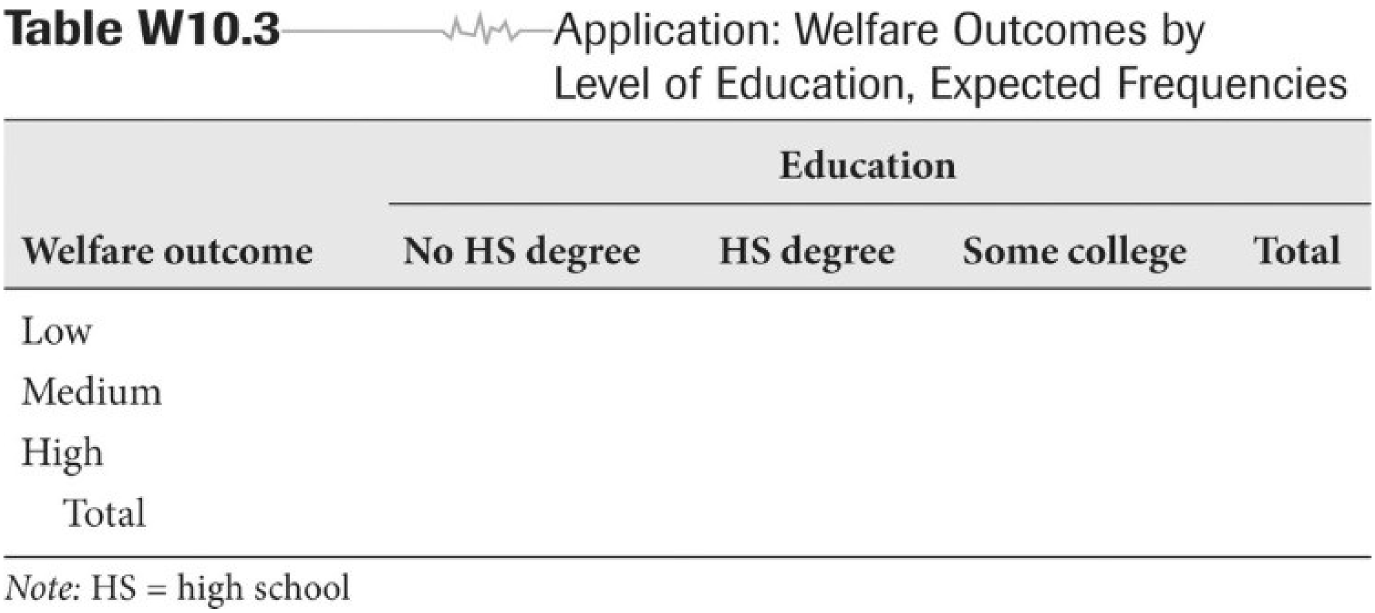


Table W10.3 Application: Welfare Outcomes by Level of Education, Expected Frequencies

	Education			
Welfare outcome	No HS degree	HS degree	Some college	Total
Low				
Medium				
High				
Total				

Note: HS = high school

Note: HS = high school

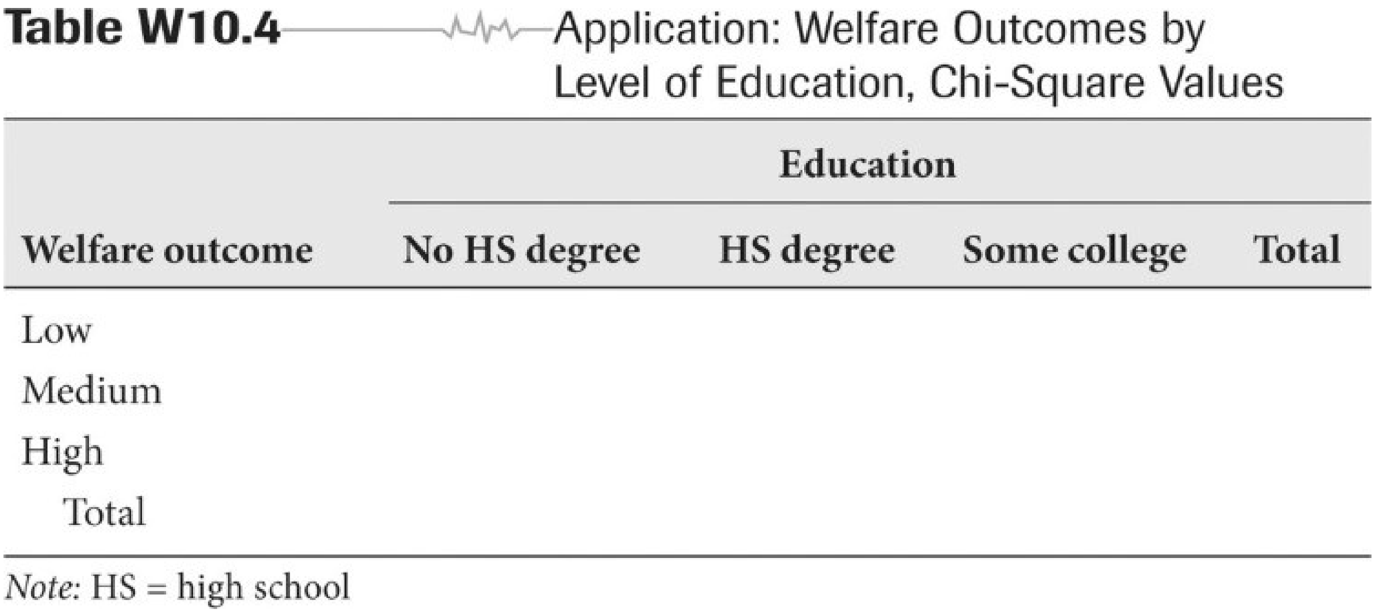


Table W10.4 Application: Welfare Outcomes by Level of Education, Chi-Square Values

	Education			
Welfare outcome	No HS degree	HS degree	Some college	Total
Low				
Medium				
High				
Total				

Note: HS = high school

Note: HS = high school

11. **Examine Table W10.5. Does this contingency table satisfy the test assumptions for chi-square? Why or why not? If not, what might be done to correct the problem?**

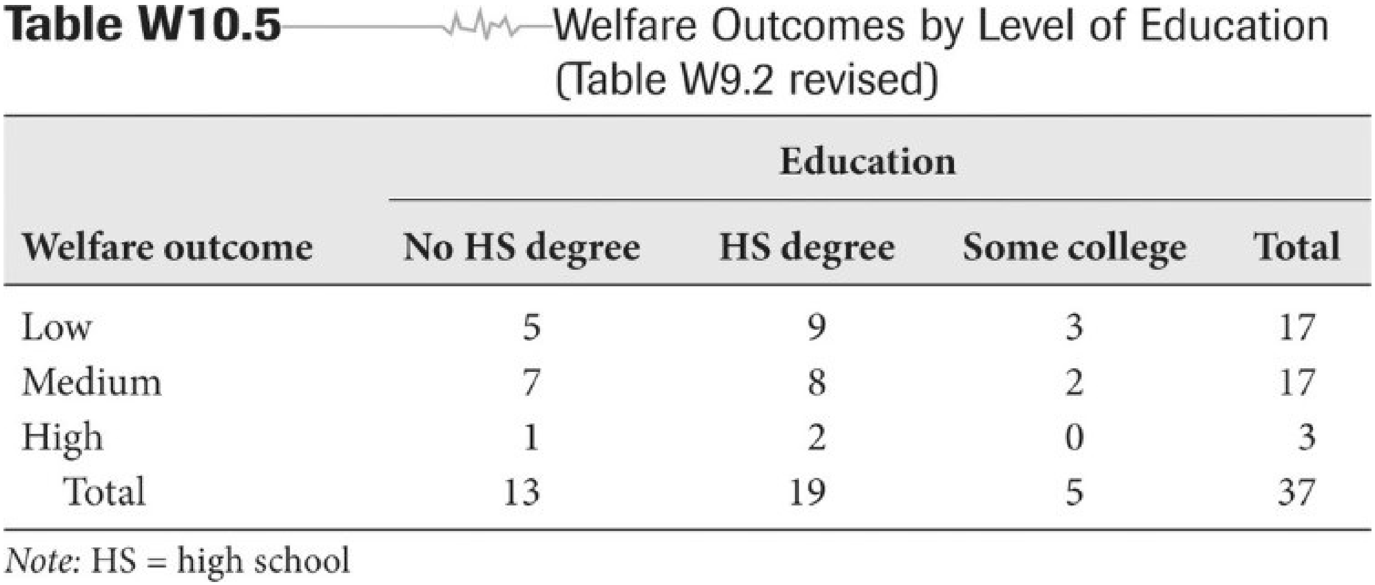


Table W10.5 Welfare Outcomes by Level of Education (Table W9.2 revised)

	Education			
Welfare outcome	No HS degree	HS degree	Some college	Total
Low	5	9	3	17
Medium	7	8	2	17
High	1	2	0	3
Total	13	19	5	37

Note: HS = high school

Note: HS = high school

12. **Good writing starts with knowing what to write—an outline or list of topics. Assume that you are an analyst with the National Institutes of Health, and your data show a statistically significant relationship between exercise and the risk of heart disease. Identify five things that you will want to report on in the Results section of your report. Consider statistical as well as practical significance. List any charts or visual aids you will use.**
    1. ______________________________
    2. ______________________________
    3. ______________________________
    4. ______________________________
    5. ______________________________
13. **In a sample of 57 students, 35 pass a test. Is this result consistent with a norm that states that at least 67 percent of students should pass the test? Is it consistent with a norm of 80 percent? What test statistics should you use?**

    ____

    ____

    ____

    ____

    ____

    ____
14. **Explain why Kendall's tau-c is a useful alternative to chi-square.**

    ____

Data-Based Exercises

Note: Some of these exercises draw on those first presented in Chapter 8. Now you can practice chi-square as discussed in the text.

1. **Use the Employee Attitudes dataset. Examine the relationship between stress (Stressed) and the morale of county employees (Himorale). Note the measurement scale of the variables. What do you conclude? Do you consider this a causal relationship or an association? Does the analysis satisfy the assumptions of the chi-square test? If not, what categories might you combine to overcome this problem?**

   ____

   ____

   ____

   ____

   ____

2. **Use the Public Perceptions dataset. Is the relationship between watching Orange TV (Watch), the county's cable television station, and trusting the county government to do what is right most of the time (Trust) statistically significant? Do you consider this a causal relationship or an association? Does the analysis satisfy the assumptions of the chi-square test? If not, how might you address this problem?**

   ____

   ____

   ____

   ____

   ____

3. **Use the Public Perceptions dataset. Examine the relationship between residents who trust the county government to do what is right most of the time (Trust) and their belief that county government works efficiently (Works). What is the practical significance of this relationship?**

   ____

   ____

____

____

____

4. **Use the Public Perceptions dataset. In workbook Chapter 8, Data-Based Exercise 3, you evaluated the importance of selected issues. The three most important issues were helping public schools (Pubschl), fighting illegal drug use (Figtdrug), and dealing with the problems of gangs (Gangs). Do whites and nonwhites agree on the importance of these priorities? On which issues is there a difference? Discuss the practical importance of any significant differences.**

____

____

____

____

____

5. **Examine the variables in the Employee Attitudes dataset. Identify five relationships that you hypothesize to be statistically significant. Test these hypotheses. What do you find?**

____

____

____

____

____

6. **Compare some of the above results using chi-square and Kendall’s tau-c. What do you conclude?**

____

____

____

____

____

7. **In Chapter 8, Data-Based Exercise 7, you examined the relationship between attendance and academic performance based on your college experience. Now, treat those data as a sample so that you can conduct a hypothesis test. Do you find a relationship between these two variables based on the sample? Do you draw a different conclusion from the one you drew in Chapter 8? If yes, why is it different?**

____

____

____

____

____

8. **(*optional*). Use the Employee Attitudes dataset. This is an exercise in examining control variables (see Appendix 11.1 in the textbook). First, examine the relationship between gender (Gender) and the perceived morale of county employees (Himorale). What do you conclude? Next, test the rival hypothesis that this relationship is spuriously caused by stress: maybe one gender experiences a higher level of stress. (To this end, recode the variable Stressed into two groups: high stress and low stress.) What do you conclude?**

____

____

____

____

____

Further Reading

There exists no shortage of "general purpose" introductory statistics texts, and most of them discuss hypothesis testing. If you seek additional assistance on this topic, we suggest that you peruse your local library shelves until you find a book that you like. Statistics books explain this topic in similar yet subtly different ways. The textbooks suggested here explain hypothesis testing and contingency table analysis in a clear manner. We hope that you like our approach, but feel free to choose whatever works for you. Sometimes, it is just a matter of seeing the same thing from different angles.

Some statistics texts have been mentioned in Chapters 6 and 7, and these will likely help here, too. Some "oldies but goodies" are Sam Kash Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed. (New York: Radius Press, 1991). It is an inexpensive book that does an exemplary job of explaining advanced statistics concepts in nonmathematical terms, but the coverage of topics is advanced, perhaps aimed at doctoral students. Another option is Tari Renner, *Statistics Unraveled: A Practical Guide to Using Data in Decision Making* (Washington, D.C.: International City/County Management Association, 1988). Of course, you can also compare our treatment with that of another book in public affairs, Kenneth Meyer, Jeffrey Brudney, and John Bohte, *Applied Statistics for Public Administration*, 9th ed. (Florence, Ky.: Wadsworth, 2014). Many basic statistics books have separate chapters on the use of statistics for categorical variables. Comprehensive treatment of this subject is found in Lawrence L. Giventer, *Statistical Analysis for Public Administration*, 2nd ed. (Burlington, Mass.: Jones & Bartlett, 2007). This book covers the material presented in this single chapter over multiple chapters, and numerous applications and examples of hand calculations of test formulas are provided.

Chapter 12 The T-Test

Q & A

1. ***When should a t-test be used?***

   T-tests are used for testing whether two groups have different means. One variable is dichotomous, whereas the other is continuous.
2. ***What is the null hypothesis of a t-test?***

   The null hypothesis of a t-test is that the means of a variable do not differ between two groups in the population; that is, the two means are equal. Rejecting the null hypothesis implies that the two group means are different in the population.
3. ***Name four t-test assumptions.***
   1. One variable is continuous, and the other variable is dichotomous.
   2. The two distributions have equal variances.
   3. Observations are independent.
   4. The two distributions are normally distributed.
4. ***How do researchers test for the normality of variables?***

   Researchers typically use a combination of visual inspection and statistical tests, such as the Kolmogorov-Smirnov test, to determine the normality of variables. It is acceptable to consider variables as being normally distributed when they visually appear to be so, even when the null hypothesis of normality is rejected by normality tests. Of course, variables are preferred that are supported by both visual inspection and normality tests.
5. ***What is the purpose of the Levene's test of the equality of variances?***

   The purpose of the Levene's test is to test whether groups have equal variances. This test is used to test one of the four t-test assumptions and is always used prior to testing the equality of means.
6. ***What is a paired t-test?***

   The paired t-test often is used when using before-and-after measurements, such as when assessing student scores before and after tests. Paired t-tests are used when analysts have a dependent rather than an independent sample.
7. ***What does the term* dependent samples *refer to?***

   Dependent samples (also called related samples) are those in which the selection of one subject in a sample affects the selection of subjects in another group. Examples of dependent samples are those that involve before-and-after test scores, subjects that have been matched (or paired) in some way, and evaluators' ratings. Separate statistical tests exist for dependent samples.
8. ***What is a one-sample t-test?***

   A one-sample t-test is used to test whether the mean of a variable is significantly different from a user-specified value.

9. ***What is the difference between parametric and nonparametric tests? Which type of test is the t-test?***

   Parametric tests make assumptions about the distribution of data and also are used to make inferences about population parameters. Formally, the term *parametric* means that a test makes assumptions about the distribution of the underlying population. Parametric tests have more test assumptions than nonparametric tests, and most typically assume that the variable is continuous and normally distributed.

10. ***What are the advantages and disadvantages of using nonparametric alternatives to t-tests?***

    The chief advantage of nonparametric alternatives is that they do not require that continuous variables be normally distributed. The chief disadvantage of nonparametric alternatives is that they are less likely to reject the null hypothesis. A further, minor disadvantage is that nonparametric alternatives do not provide descriptive information about variable means; separate analysis is required for that.

11. ***What is a nonparametric alternative to the independent-samples t-test?***

    Nonparametric alternatives to the independent-samples test are the Mann-Whitney (U) and Wilcoxon (W) tests. The Mann-Whitney and Wilcoxon tests are equivalent, and both are simplifications of the more general Kruskal-Wallis H test, discussed in Chapter 11 of the text.

12. ***What is a nonparametric alternative to the paired-samples t-test?***

    A nonparametric alternative to the paired-samples t-test is the Wilcoxon signed rank test.

13. ***What is a nonparametric alternative to the one-sample t-test?***

    The Wilcoxon signed rank test can also be adapted as a nonparametric alternative to the one-sample t-test. Then, analysts create a second variable that, for each observation, is the test value.

Critical Thinking

1. **Explain the importance of the four test assumptions of the independent-samples t-test.**

   ____

   ____

   ____

   ____

   ____

   ____

2. **Why is the assumption of equal variances irrelevant for the paired-samples t-test?**

   ____

   ____

   ____

   ____

   ____

   ____

3. **Table W12.1 is the printout of a t-test (independent samples). The continuous variable is an index variable of environmental concern. The dichotomous variable is a measure of education (college versus no college). Interpret and write up the results. What other information would you like to have about this relationship?**

   ____

   ____

   ____

   ____

   ____

   ____

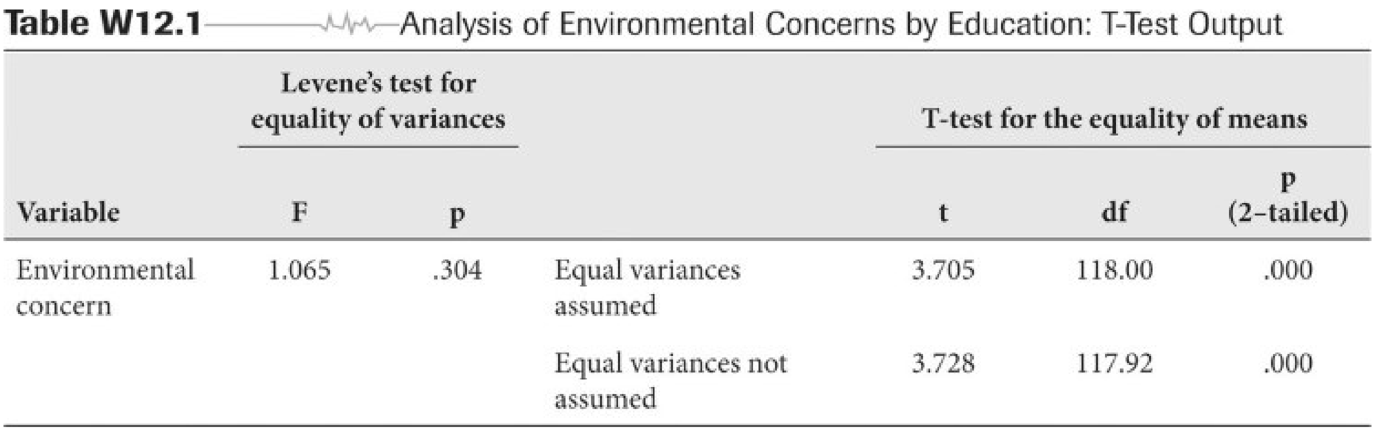


Table W12.1 Analysis of Environmental Concerns by Education: T-Test Output

	Levene's test for equality of variances			T-test for the equality of means		
Variable	F	p		t	df	p (2-tailed)
Environmental concern	1.065	.304	Equal variances assumed	3.705	118.00	.000
			Equal variances not assumed	3.728	117.92	.000

4. **Table W12.2 is the printout of a paired-samples t-test. The data are before-and-after measurements of a public safety program. Interpret and write up the results. What other information would you like to have about this relationship?**

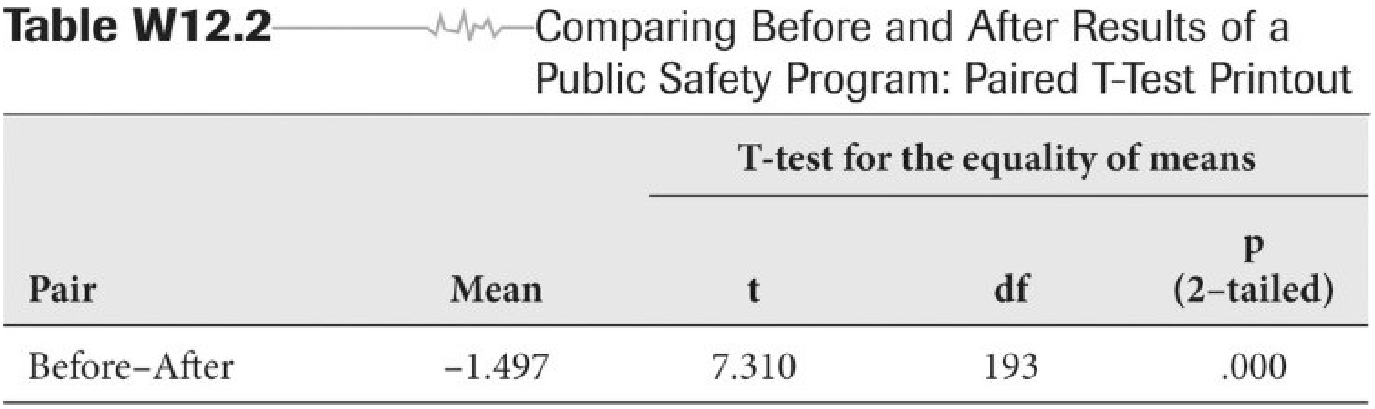


Table W12.2 Comparing Before and After Results of a Public Safety Program: Paired T-Test Printout

		T-test for the equality of means		
Pair	Mean	t	df	p (2-tailed)
Before–After	–1.497	7.310	193	.000

5. **Explain the following statement regarding tests for normality: "Whereas failure to reject the null hypothesis indicates normal distribution of a variable, rejecting the null hypothesis does not indicate that the variable is not normally distributed."**

6. **Table W12.3 shows the result of the Kolmogorov-Smirnov test for a variable and three of its transformations. Which variable(s) should the analysts consider for subsequent use?**

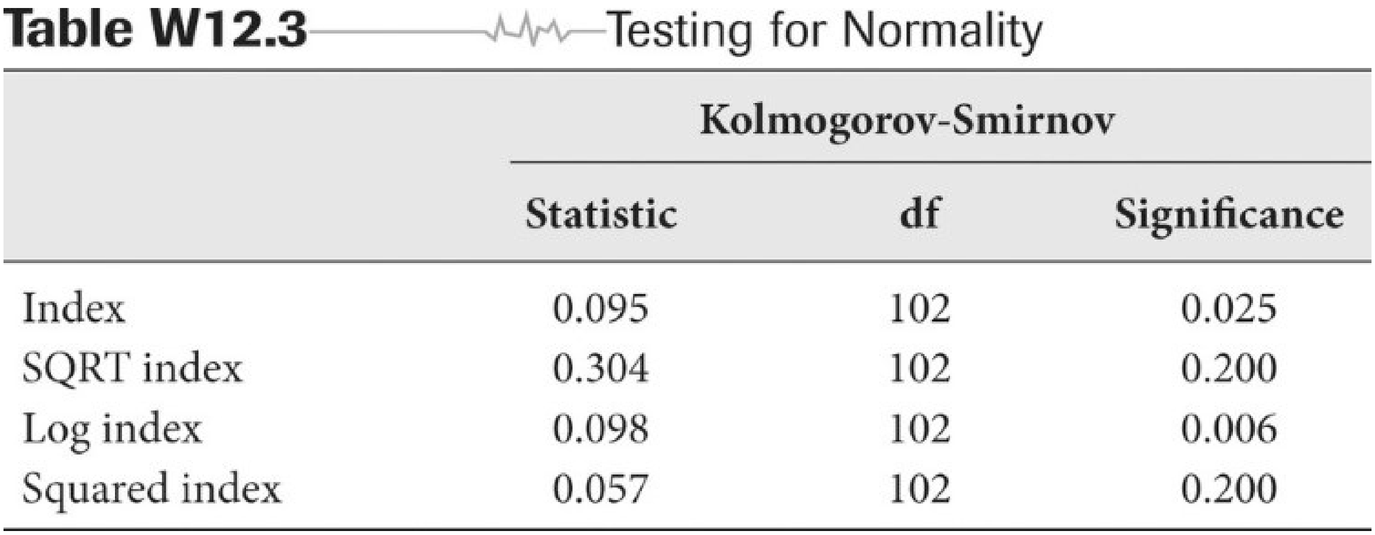


Table W12.3 Testing for Normality

	Kolmogorov-Smirnov		
	Statistic	df	Significance
Index	0.095	102	0.025
SQRT index	0.304	102	0.200
Log index	0.098	102	0.006
Squared index	0.057	102	0.200

7. **You are analyzing your data and, using a histogram or boxplot, you see some outliers. Should your initial approach be to remove them or to keep them? Why? What are the pros and cons of each approach?**

Data-Based Exercises

1. **Use the Watershed dataset. Replicate the analysis in the text. First, analyze the normality of the variable Conpolut and consider various transformations, as shown in Figures 12.4 and 12.5 in the textbook. Then, replicate the t-test for comparing the East (defined as Northeast and Southeast combined) with all other regions.**

    ____

    ____

    ____

    ____

    ____

2. **Use the Watershed dataset. Do states in the East vary in the number of fish and wildlife advisories that have been made? (Use the variable Advisory.) Is the variable Advisory normally distributed? If not, what transformation do you suggest?**

    ____

    ____

    ____

    ____

    ____

3. **Use the Public Perceptions dataset. An analyst wants to know whether men and women have different perceptions of customer service. To this end, we will use an index variable of customer satisfaction (see Chapter 3); the index variable is provided on the dataset as the last variable, Satisfac, but you can also practice making this index variable in the following way:**

    From among those respondents who have had contact with a county employee (that is, if contact = 1), create an index variable of "customer service" that is composed of the six survey items: "employees were helpful," "employees treated me with courtesy and respect," "employees were friendly," "service was provided without mistakes," "the service experience exceeded my expectations," and "service was provided in a timely manner." If you are using SPSS, see Chapter 19 of this workbook for instructions on creating index variables in SPSS. After you have either created this variable or identified this variable in the dataset, address the question of whether men and women vary in their customer satisfaction experience.

4. **Use the Public Perceptions dataset. An analyst wants to know whether the index variable of customer service varies across race. Recode the Ethnic variable to distinguish among whites, blacks, other races, and Hispanics. If you are using SPSS, see Chapter 19 of this workbook for instructions on recoding variables in SPSS.**

5. **Do coastal counties spend more on conservation? Conduct a t-test to compare conservation spending in total spending using 2008 data (ConsTotal08) from the Data_FL_County dataset. The dataset contains the dichotomous variable Coastal.**

6. **A group of 12 welfare recipients participate in training. Before-and-after abilities are measured through a standardized test. See Table W12.4. Is there evidence of improvement? Compare the results of both the paired-samples t-test and the nonparametric alternative.**

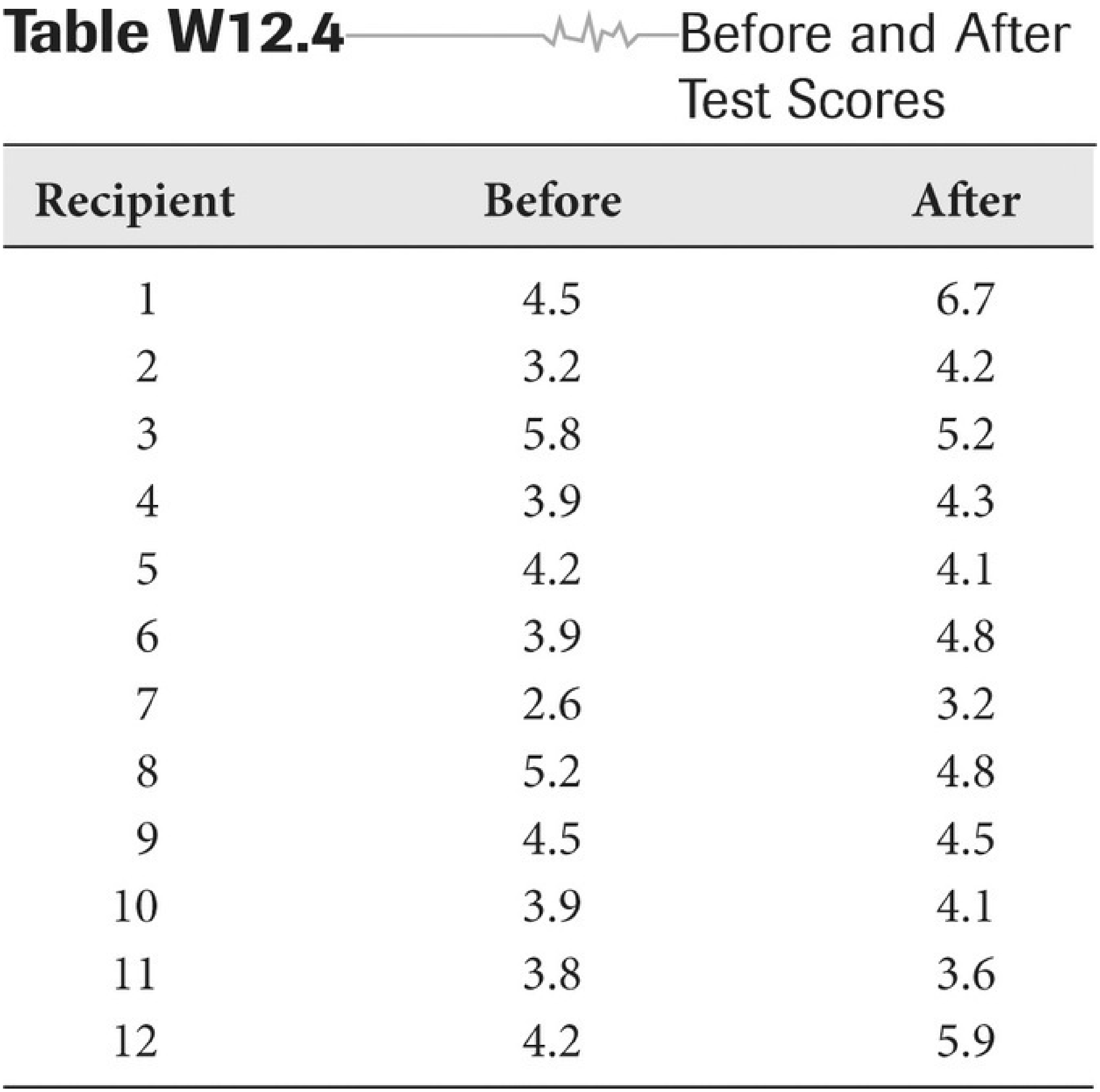


Table W12.4 Before and After Test Scores

Recipient	Before	After
1	4.5	6.7
2	3.2	4.2
3	5.8	5.2
4	3.9	4.3
5	4.2	4.1
6	3.9	4.8
7	2.6	3.2
8	5.2	4.8
9	4.5	4.5
10	3.9	4.1
11	3.8	3.6
12	4.2	5.9

7. **Consider Table W12.4 again. Is the after test score consistent with a norm of 5.0? With a norm of 6.0? Use both parametric and nonparametric tests. Compare the results.**

8. **T-tests can be used in performance analysis and evaluation. The psychosocial functional score (PFS) is used to assess school-age children’s psychosocial behaviors. A score of 25 points or above is considered normal. A school counseling program has used the PFS to assess participating students’ progress. A sample of 15 participating students is tested, and their PFS scores are 29, 32, 18, 23, 27, 19, 34, 32, 27, 27, 23, 26, 32, 29, and 34. Evaluate whether participants’ average PFS score is greater than 25. Write a summary to report your findings to the program manager, who knows nothing about statistics.**

9. **(*optional*) Men and women are asked to rank six issues in the order of priority. See Table W12.5. Do men and women differ in their order of priorities? Which test statistic would you use? Can you think of some other situations that match this scenario?**

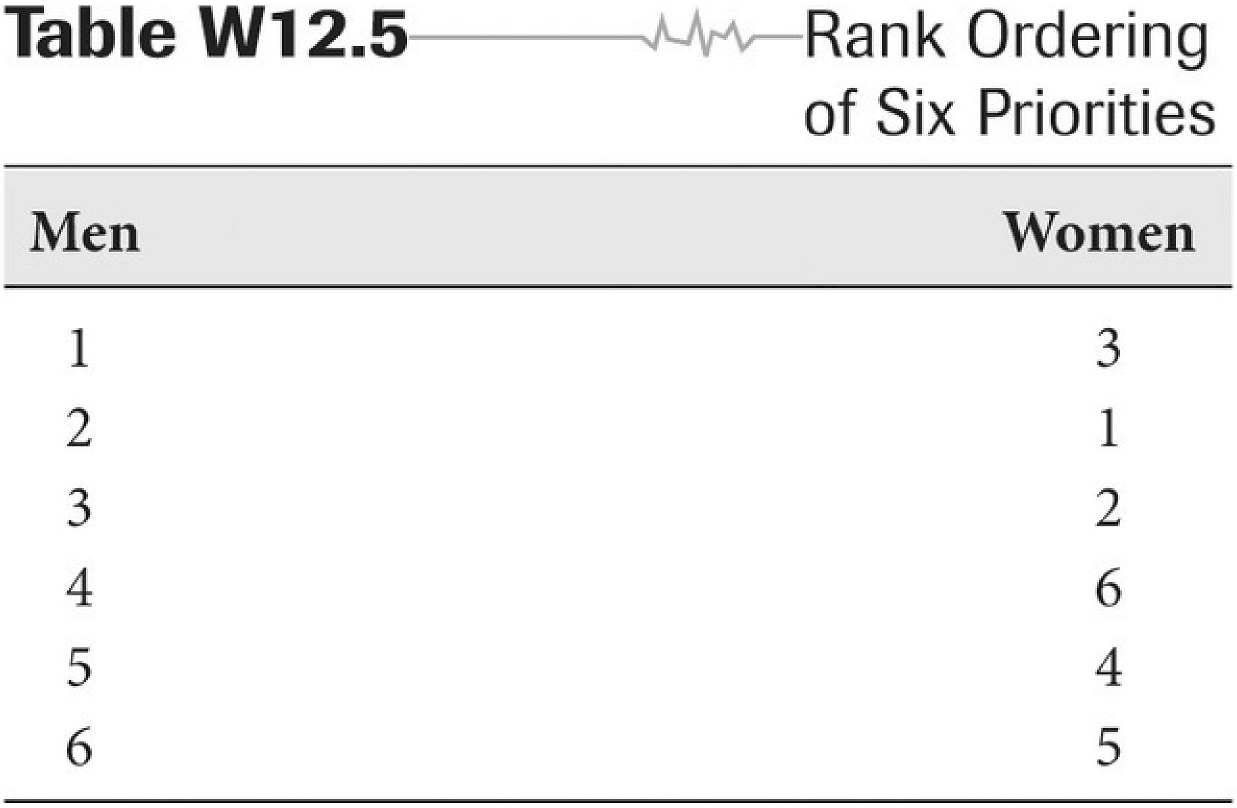


Table W12.5 Rank Ordering of Six Priorities

Men	Women
1	3
2	1
3	2
4	6
5	4
6	5

Further Reading

Most statistics books discuss t-tests, although they vary in their coverage of formulas and test assumptions. An in-depth discussion of t-tests is found in David Howell, *Statistical Methods for Psychology*, 8th ed. (Belmont, Calif.: Wadsworth, 2012), Chapter 7. The *SPSS User’s Guide* (Chicago: SPSS) has excellent examples of the use of t-tests in practice and, in a separate chapter, gives some further examples of variable transformations. Another solid treatment is found in Ronet Bachman and Raymond Paternoster, *Statistical Methods for Criminology and Criminal Justice*, 3rd ed. (Thousand Oaks, Calif.: Sage, 2016).

chapter

Chapter 13 Analysis of Variance (ANOVA)

Q & A

1. ***For what purpose is ANOVA used?***

   ANOVA is used for testing whether three or more groups have different means and, if so, which groups have different means. By comparison, t-tests are used for testing whether two groups have different means.
2. ***What is the purpose of the global F-test in ANOVA?***

   The ANOVA global F-test tests for differences among any of the means.
3. ***What are the four test assumptions for ANOVA?***
   1. One variable is continuous, and the other variable is ordinal or nominal.
   2. The group distributions have equal variances.
   3. Observations are independent.
   4. The variable is normally distributed in each of the groups.
4. ***What is a post-hoc test? Name three post-hoc tests.***

   Post-hoc tests are tests that test all possible group differences and do so in a manner that maintains a true (5 percent or 1 percent) level of significance. Three popular post-hoc tests are Tukey, Bonferroni, and Scheffe. The Scheffe test is the most conservative, the Tukey test is best when many comparisons are made (that is, there are many groups), and the Bonferroni test is preferred when few comparisons are made.
5. ***What are homogenous subsets?***

   Homogeneous subsets are groups with insignificant differences of their means.
6. ***How can ANOVA test the linearity of relationships that involve a continuous variable and an ordinal variable?***

   ANOVA can also be used to test the linearity of interval-ordinal relationships, that is, whether the change in means follows a linear, an increasing, or a decreasing pattern according to the ordering of the ordinal variable. The appropriate F-test is the statistic reported as the "linear term for unweighted sum of squares" in ANOVA tables.
7. ***What is MANOVA? How is it different from two-way ANOVA?***

   MANOVA is a technique for analyzing the effects of multiple independent variables on multiple dependent variables. Two-way ANOVA is a technique for analyzing the effects of two independent variables on one dependent variable.
8. ***What is the Kruskal-Wallis H test?***

   The Kruskal-Wallis H test is a nonparametric alternative to one-way ANOVA. As a nonparametric method, the Kruskal-Wallis H test does not assume normal populations, but the test does assume similarly shaped distributions for each group. A limitation of this nonparametric test is that it does not provide group means, post-hoc tests, or analysis of which groups are homogenous.

Critical Thinking

1. **Your data include a measure of air pollution that is measured on a continuous scale, as well as another variable that measures the location of that air pollution in five counties. Which test statistic should you use to examine whether mean pollution levels vary across counties?**

2. **Table W13.1 is a sample output ANOVA table, comparing educational outcomes across five different groups of students. Interpret the output.**

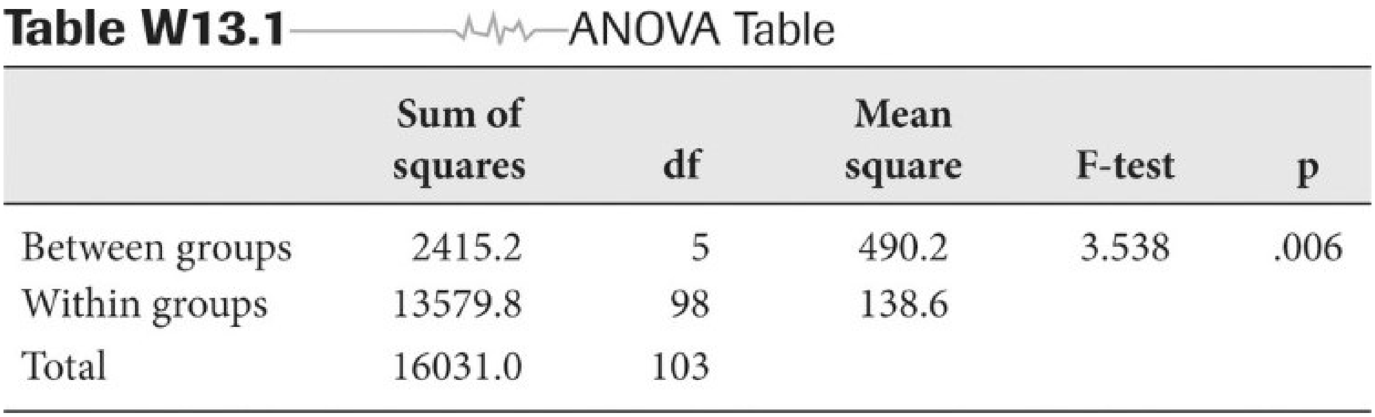


Table W13.1 ANOVA Table

	Sum of squares	df	Mean square	F-test	p
Between groups	2415.2	5	490.2	3.538	.006
Within groups	13579.8	98	138.6		
Total	16031.0	103			

3. **ANOVA can be useful, even outside your field of work. Conduct the following**

experiment. Choose one of your favorite locations away from your home. It could be a supermarket, a shopping mall, or a playground. Identify at least three routes to get there. Now, choose the route you use most often. Record your travel time to that location (in minutes) up to 10 times. Then, choose an alternative route and record your travel time (in minutes) up to 10 times. Choose a third route and record your travel time up to 10 times. In this experiment, make sure you use the same mode of transportation (for example, car, bike, or public transit) each time and obtain your measurements at approximately the same time of day to rule out the impact of these confounding variables on travel time. Compare travel times for each route to determine whether the mean travel times differ between routes. Is the route you use most often the fastest?

____

____

____

____

____

____

4. **In the preceding exercise, the dependent variable is travel time. You tested the impact of various routes (the independent variable) on travel time. Because these routes are unrelated to each other, you tested the mean differences of three independent samples. You can follow the same logic to test the impact of times of day on travel time. For example, you could record your travel times at 10:00 am, 2:00 pm, and 7:00 pm up to 10 times each to determine whether there is any mean difference in travel time at different times of a day. You could use the same method to test your travel times to work. If you have a flexible schedule that allows you to be at work at the time of your choice, you could use this method to choose the least time-consuming and most energy-saving way to get to work. This would truly be a performance improvement for you and for your organization. Can you think of any other ways to use ANOVA to improve your work performance or your life?**

____

____

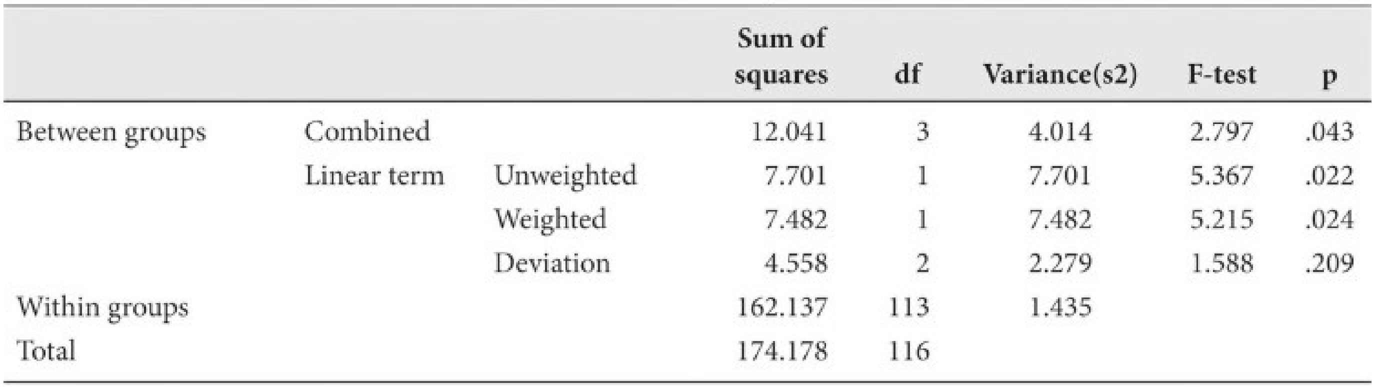
____

____

____

5. **Table W13.2 is a sample output ANOVA table with linearity test. Interpret the output.**

Table W13.2 ANOVA Table with Linearity Test



			Sum of squares	df	Variance(s2)	F-test	p
Between groups	Combined		12.041	3	4.014	2.797	.043
	Linear term	Unweighted	7.701	1	7.701	5.367	.022
		Weighted	7.482	1	7.482	5.215	.024
		Deviation	4.558	2	2.279	1.588	.209
Within groups			162.137	113	1.435		
Total			174.178	116			

6. **An analyst conducts a test of homogeneity of variances in ANOVA and finds that the Levene's statistic = .911, p = .461 > .05. What does the analyst conclude?**

7. **Table W13.3 is a sample output two-way ANOVA table. Interpret the output.**

Table W13.3 Two-Way ANOVA Table

Source	Sum of squares	df	Mean square	F-test	p
Model	1642.8	24	68.44	4.507	.000
Var A	229.8	4	57.45	3.783	.007
Var B	85.5	5	17.11	1.127	.351
Var A * Var B	409.5	14	29.25	1.926	.032
Error	1488.2	98	15.19		
Total	3130.9	122			

Data-Based Exercises

1. **ANOVA is a useful tool in performance comparisons. A nonprofit foundation funds a program to reduce accidents among school-aged children. The employees in the program conduct regular education workshops for students and teachers. The daily number of accidents per 1,000 students is used to assess the performance of the program in reducing accidents among students. The program is operated in school A, which collected the following data on 10 randomly selected days: 3.82, 1.01, 5.96, 8.99, 8.85, 5.54, 0.14, 4.07, 7.65, and 1.39. You are a performance analyst whose job it is to assess the performance of this program in school A. To conduct your evaluation, you collect data from two schools (school B and school C) that do not have these programs (that is, the control groups). The following data are for school B: 6.58, 0.59, 7.00, 7.17, 7.83, 8.68, 1.17, 5.50, 6.84, and 4.09. For school C the data are as follows: 9.47, 4.86, 8.65, 5.54, 6.45, 2.71, 8.63, 9.30, 7.37, and 7.09. Assume normal distributions of the data in each group and determine whether the program in school A has a lower number of accidents.**

2. **Use the Public Perceptions dataset. An analyst wants to know whether the index variable of customer service varies across race. Recode the variable Ethnic to distinguish among whites, blacks, other races, and Hispanics, and then calculate the means for each of these groups. Then use ANOVA to determine whether any of these differences are statistically significant.**

3. **Use the Public Perceptions dataset. An analyst wants to know whether incomes vary by age groups. Treat the income variable as a continuous variable, and treat the age variable as an ordinal variable. Calculate the means for each of these groups, and then use ANOVA to determine whether any of these differences are statistically significant. For which age groups is the relationship linear?**

4. **Use the Watershed dataset. An analyst wants to know whether the level of conventional pollutants (conpolut) is associated with the number of aquatic/wetland species at risk (speatrsk). What are your conclusions?**

Further Reading

Numerous books discuss ANOVA. See Sam Kash Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed. (New York: Radius Press, 1991); Chapter 5 provides a most lucent discussion of ANOVA. The *SPSS Base Applications Guide* has an excellent, practical discussion and carries this topic further. Analysts in health care settings will want a more detailed discussion, such as in Barbara Tabachnick and Linda Fidell, *Using Multivariate Statistics,* 6th ed. (New York: HarperCollins, 2012), or Michael Kutner, *Applied Linear Statistical Models* (New York: McGraw-Hill, 2013), 5th ed.

Chapter 14 Simple Regression

Q & A

1. ***For what purpose is simple regression used?***

   Simple regression is used for testing the relationship between two continuous variables, whereby one variable is the dependent variable and the other is the independent variable.
2. ***How is simple regression used for testing hypotheses? What test statistic is used, and how is it defined?***

   Simple regression involves estimating the relationship between variables through a straight line (regression line), $y = a + bx$, where a is the intercept (or constant) and b is the slope. The slope is also called the regression coefficient. If the slope is statistically different from zero, then a relationship is said to exist between the variables in the population. To determine whether the slope equals zero, a t-test is performed. The test statistic is defined as the slope (b) divided by the standard error of the slope (se[b]).
3. ***What is the interpretation of the coefficient of determination, R-square* (R^2)*, and what values can it assume?***

   R-square can assume values between 0 and +1. The value of R-square is interpreted as the percentage of variation in the dependent variable that is explained by the independent variable.
4. ***Why is linearity important in regression analysis?***

   Regression analysis estimates the relationship between variables as a straight line over the *entire range* of observations. If the pattern is curved (for example, parabolic) or broken (for example, upward sloping for one part of the observations and downward sloping for the rest), the assumption of linearity is violated, and the statistical significance of regression coefficients will be underestimated.
5. ***What is the difference between the observed and predicted values of the dependent variable? What is the error term?***

   The predicted value of y (defined, based on the regression model, as $y = a + bx$) is typically different from the actual observed value of y. The predicted value of the dependent variable y is sometimes indicated as (pronounced “y-hat”). The difference between y and is called the regression error or error term (e). Hence, $y = + e$.
6. ***What is Pearson’s correlation coefficient,* r*? What values can it assume, and how is it interpreted?***

   Pearson’s correlation coefficient, r, measures the association between two continuous variables. It does not distinguish between a dependent and an independent variable, as does simple regression. Pearson’s correlation coefficient ranges from −1 to +1. The sign indicates the direction of the relationship, which is the same sign as the slope coefficient. Values of r^2 between 0.0 and 0.20 indicate weak associations, values between 0.20 and 0.40 indicate moderate associations, values between 0.40 and 0.65 indicate strong associations, and values above 0.65 are considered to indicate very strong associations. For the two-variable, simple regression model, $r^2 = R^2$.

7. ***How is the Pearson's correlation coefficient,* r*, different from the slope (regression coefficient),* b?**

   Comparison of the measures *r* and *b* (the slope) sometimes causes confusion. Pearson's correlation coefficient, *r*, indicates not the regression slope but rather the extent to which observations lie close to it. A steep regression line (large *b*) can have observations that lie either loosely or closely scattered around it, as can a shallow (more horizontal) regression line.

8. ***What purpose does Spearman's rank order correlation coefficient (ρ) serve?***

   Spearman's rank order correlation coefficient, *ρ*, is a nonparametric alternative to simple regression and Pearson's correlation coefficient, *r.* Spearman's rank order correlation coefficient looks at correlation among the ranks of the data rather than among the values. Because Spearman's rank correlation coefficient examines correlation among the ranks of variables, it can also be used with ordinal-level data. Spearman's rank correlation coefficient has a "percent variation explained" interpretation, similar to the other measures described here.

Critical Thinking

1. **Explain why variables are assumed to be linearly related in simple regression.**

2. **Explain why the slope can be used as a test of the relationship of two variables in simple regression.**

3. **Explain why, if *b* is statistically significant, then *r* should be statistically significant, too.**

4. **Draw a scatterplot between two variables that has a large r^2 and a small, negative *b*. Then, draw another scatterplot that has a small r^2 and a large, positive *b*. Show the regression lines in each.**

____

____

____

____

____

____

5. **You are a program manager for a welfare agency and want to know if a relationship exists between the length of time that people receive welfare (in months) and their education (total years). Both variables are continuous. Table W14.1 is the hypothetical output from your simple regression model. Interpret the results.**

____

____

____

____

____

____

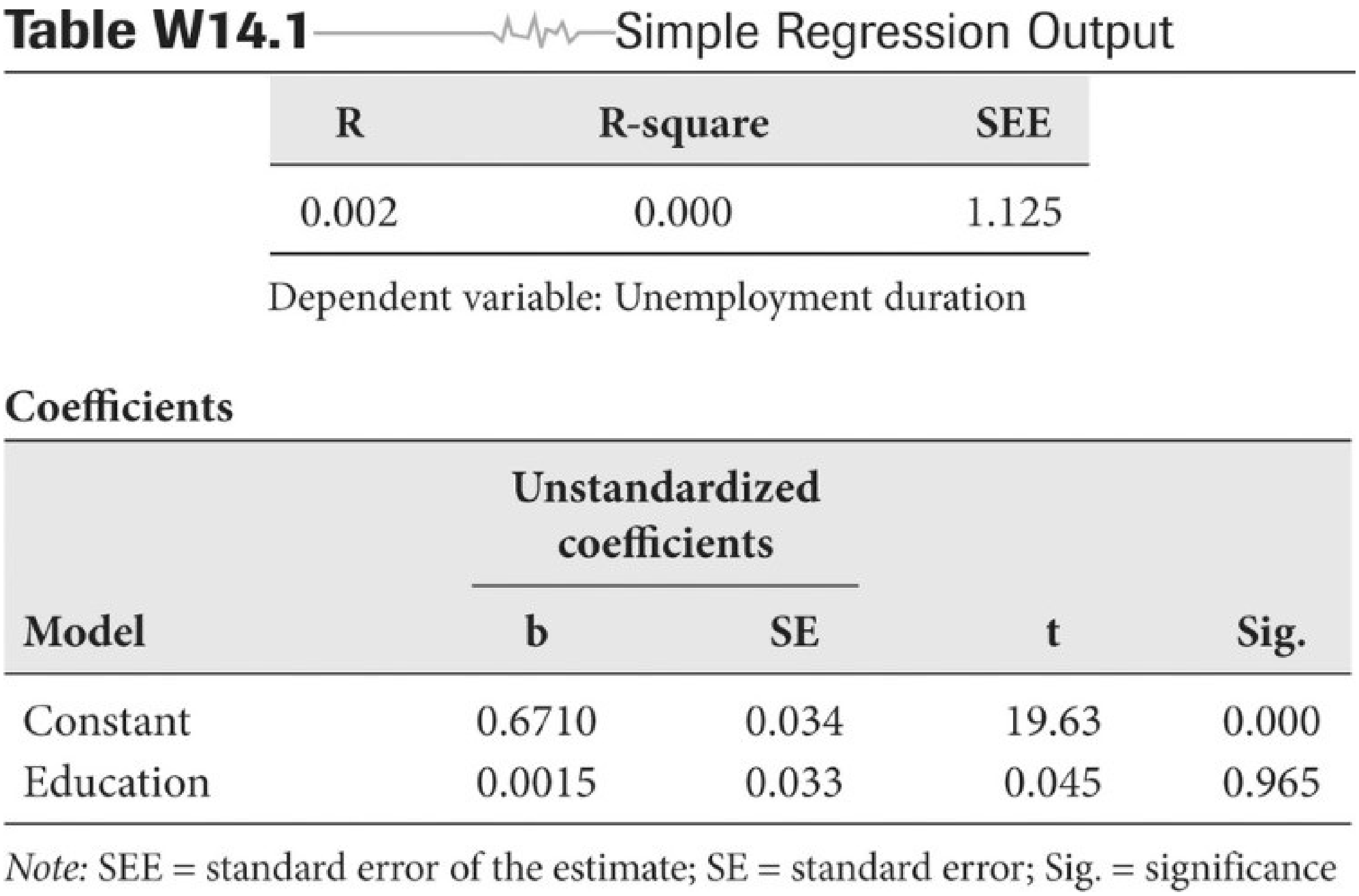


Table W14.1 Simple Regression Output

R	R-square	SEE
0.002	0.000	1.125

Dependent variable: Unemployment duration

Coefficients

	Unstandardized coefficients			
Model	**b**	**SE**	**t**	**Sig.**
Constant	0.6710	0.034	19.63	0.000
Education	0.0015	0.033	0.045	0.965

Note: SEE = standard error of the estimate; SE = standard error; Sig. = significance

6. **Explain how Spearman's rank coefficient transforms interval-level values into ranks. Then explain why Spearman's rank coefficient can also be used with ordinal-level variables.**

Data-Based Exercises

1. **Use the Productivity dataset. Examine the bivariate relationship between having authority (Jobauthr) and knowledge (Jobknowl) to do one’s job. Use simple regression. What are your findings?**

   ____

   ____

   ____

   ____

   ____

2. **Use the Productivity dataset. Examine the bivariate relationship between employees’ perception of productivity (Productivity) and the perceptions of having adequate authority to do one’s job (Jobauthr). Use both simple regression and Pearson’s correlation coefficient.**

   ____

   ____

   ____

   ____

   ____

3. **Examine the normality of variables used in the preceding exercises (see Chapter 12 in the text). Are Jobauthr, Productivity, and Jobknowl normally distributed? On what basis do you reach this conclusion? Is there any transformation that can make them normal?**

   ____

   ____

   ____

   ____

   ____

4. **Many managers believe that a government’s spending on the environment is a response**

to the environmental pressure such as population growth, pollution level, and economic development. (1) Use the dataset Data_FL_County to conduct a hypothesis test on the relationship between the environmental spending (EnvirnTotal08) and population growth (PopGrowth08) using the Pearson correlation. (2) Run a simple regression for this relationship.

____

____

____

____

____

5. **Use the Community Indicators dataset. Which variables are associated with median household incomes?**

____

____

____

____

____

6. **Use the Community Indicators dataset. Create the variables “Cancers per capita” and “Rapes per capita.” Which variables are associated with these variables?**

____

____

____

____

____

7. **Use the Productivity dataset. Compare employees’ perceptions of productivity (Productivity) with those of the outside consultant (Wkctrpro). What test statistic might you use for this comparison?**

____

____

8. **Regression is a useful tool in performance analysis. Choose a variable that you think influences your academic performance. (Many students would say that effort, attendance, instruction quality, level of difficulty, and study methods are keys to improving academic performance—choose one.) Create a database that includes all the courses you have taken in college. Recall your experiences and performance (for example, your grade) in those courses. For example, you could measure effort with the number of hours spent studying for the course weekly. Level of difficulty could be measured with “very difficult,” “difficult,” or “not difficult.” Use simple regression to analyze the relationship between the variable of your choice and your academic performance. What do you conclude? Write a report that summarizes your findings for a person who knows nothing about statistics.**

Further Reading

Many books mentioned in the previous chapters also discuss simple regression and the other statistics described here. By now you may have found one or two other statistics texts that are useful to you, and these are likely to discuss these statistics as well. In Chapter 15, we identify some further readings for multiple regression, and these books typically also discuss simple regression.

Chapter 15 Multiple Regression

Q & A

1. ***For what purpose is multiple regression used?***

   Multiple regression is used to examine the effect of control variables on the relationship between one dependent variable and one (or more) independent variable.
2. ***What is full model specification? How does full model specification relate to a nomothetic mode of explanation?***

   Full model specification means that analysts attempt to identify *all the variables* that affect a dependent variable. Fully specified models have two parts: (1) the identification of *the most important factors* that affect a dependent variable (these are independent variables) and (2) the identification of *all other factors* that affect the dependent variable, whose cumulative effects are contained in the error term. The identification of the most important factors is called a nomothetic mode of explanation.
3. ***What key assumption is made about the error term in regression? How is this assumption examined?***

   The error term is interpreted as the effect on y of all other influences on y that are *not included* as independent variables in the regression model. A key assumption is that the cumulative effect of these independent variables on the regression is zero. Examination of this assumption is based on plots of the standardized residuals against the standardized predicted variable. When the cumulative effect of the error term is zero, its pattern is randomly distributed around (0,0) of the error term plot.
4. ***How many independent variables are commonly identified in multiple regression? What is the relationship among independent variables?***

   Multiple regression models typically have five to seven independent variables, although some have many more. Independent variables should be the most important factors that affect the dependent variable. Correlations among independent variables should be low.
5. ***What is the interpretation of regression coefficients in multiple regression?***

   In multiple regression, the regression coefficients are interpreted as their effect on the dependent variable, controlled for the effect of all other independent variables included in the regression.
6. ***What undesirable property does the coefficient of determination, R-square* (R^2)*, have in multiple regression, and how is this problem overcome?***

   R-square is interpreted as the percentage of variation in the dependent variable that is explained by the independent variable(s). R-square has the undesirable property that it increases with the number of independent variables included in the regression model. Adjusted R-square (R–^{2}) controls for the number of independent variables in the regression and is always equal to or less than R-square.

7. ***What are standardized coefficients (beta values)? Why are they used?***

   Beta values (or betas) are standardized regression coefficients. Beta (β) is defined as the change produced in the dependent variable by a unit of change in the independent variable when both are measured in terms of standard deviation units. Betas are unitless, allowing analysts to compare the impact of different independent variables on the dependent variable. It is appropriate to compare betas across independent variables in the same regression model but not across different models.
8. ***What is the function of the global F-test in multiple regression?***

   The global F-test examines the overall effect of all independent variables jointly on the dependent variable. The null hypothesis is that the overall effect of all independent variables jointly on the dependent variables is statistically insignificant. The alternate hypothesis is that this overall effect is statistically significant. The null hypothesis implies that none of the regression coefficients is statistically significant, that is, $b_1 = b_2 = . . . = 0$; the alternate hypothesis implies that at least one of the regression coefficients is statistically significant.
9. ***What are dummy variables? Why are they useful? Give an example of the use of dummy variables.***

   Dummy variables allow researchers to use nominal-level variables in multiple regression. Dummy variables are variables that have values of either one or zero. See Table 15.2 in the textbook for an example of recoding. The number of dummy variables equals the number of measurement categories *minus one*.
10. ***Why are regression assumptions so important?***

    When the assumptions of multiple regression are violated, the results of multiple regression may be invalid.
11. ***Name six assumptions of multiple regression.***
    1. No observation is an outlier.
    2. No multicollinearity is present among the independent variables.
    3. The relationships between the independent variables and the dependent variable are linear.
    4. The error term is homoscedastic across the range of each independent variable.
    5. Error terms are not serially correlated (or autocorrelated).
    6. Variables are measured and specified accurately.
12. ***Explain the concept of outliers and how they may affect the results of multiple regression. Also, explain how outliers are detected and how the problem of outliers is remedied.***

    Outliers are observations with unusual values that may affect the statistical significance of regression coefficients, notably by influencing the regression slope. Outliers are detected by the size of their error term. Specifically, they are observations whose error terms either exceed +3 standard deviations or are less than –3 standard deviations. The problem of outliers usually is remedied by excluding such observations from analysis.
13. ***Explain the concept of multicollinearity and how it affects the results of multiple***

regression. Also, explain how multicollinearity is detected and how the problem of multicollinearity is remedied.

Multicollinearity is the problem of two or more independent variables that are so highly correlated that their individual effects on the dependent variable are statistically indistinguishable. Multicollinearity affects the results of multiple regression by causing the regression coefficients of multicollinear variables to be insignificant. Multicollinearity is usually first suspected when regression coefficients are insignificant, even though in bivariate analysis they are known to be highly significant. Multicollinearity is detected formally by variance inflation factor (VIF) scores that exceed 5 or 10. Multicollinearity is remedied by combining substantively related variables into a single index variable, by dropping a collinear variable, or by replacing one variable with a substantively similar but empirically dissimilar variable.

14. ***Explain the importance of linearity. How is curvilinearity detected, and how is it remedied?***

Multiple regression assumes that independent variables are *linearly* correlated with the dependent variable. When relationships are nonlinear (such as being curvilinear), regression coefficients underestimate the significance of the relationship. In some instances, the regression coefficient will be estimated as being insignificant when it is not. Diagnosis of curvilinear relationships centers on examining a curvilinear pattern of the error terms. Curvilinearity is typically corrected by transforming the independent variable with which the dependent variable is curvilinearly related.

15. ***What is heteroscedasticity? How does it affect the results of multiple regression? How can we detect and remedy heteroscedasticity?***

Heteroscedasticity is the problem of unequal variances of the error term. Unequal variances of the error term violate the assumption of the random distribution of the error term and cause the statistical significance of regression coefficients to be underestimated. Heteroscedasticity is graphically detected by examining the error term plot for unequal variances. Error terms are also plotted against each independent variable to determine which dependent-independent variable relationship is heteroscedastic. Often, a logarithmic transformation of both the dependent and independent variables sufficiently corrects the problem.

16. ***Explain the concept of autocorrelation. How can it affect the results of multiple regression? Also, explain how autocorrelation is detected and how it is remedied.***

Adjacent, time-ordered values of observations are usually highly correlated with each other: knowledge of today’s value is a good predictor of tomorrow’s. Autocorrelation (also called serial correlation) increases the significance of regression coefficients. Autocorrelation can be detected by plotting the error term against time. It is formally detected by testing the Durbin-Watson test statistic: values close to 2 indicate the absence of serial correlation, whereas values closer to 0 and 4 may indicate serial correlation. Two approaches to correcting for serial correlation are (1) to add a trend variable to the model and (2) to test the model in first-order difference form.

17. ***Why is accurate measurement important in multiple regression?***

    Multiple regression assumes that variables are measured accurately; variables should be substantively valid and free from any systematic biases. Accurate measurement is especially important for the dependent variable because inaccurate measurement may render it impossible for independent variables to achieve requisite levels of statistical significance.

18. ***Explain the importance of neither omitting relevant variables nor including irrelevant variables.***

    The effect of *omitting a relevant variable* is to inflate the value of t-test test statistics of independent variables that are included. The omitted variable is a relevant control variable. The effect of *including irrelevant variables* is the opposite, to understate the importance of other independent variables. Theoretically irrelevant variables cannot be justified, no matter how statistically significant they may be.

Critical Thinking

1. **Evaluate and explain the following statement: “Multiple regression is no substitute for bivariate analysis.”**

2. **You are a program manager for a welfare agency. In workbook Chapter 14, Critical Thinking Exercise 5, you examined the relationship between the length of time that people receive welfare (in months) and their education (total years). Now you want to consider other variables that might also affect the length of unemployment. Table W15.1 lists categories of different factors that are hypothesized to be the most important in affecting the length of unemployment. Interpret and write up your results.**

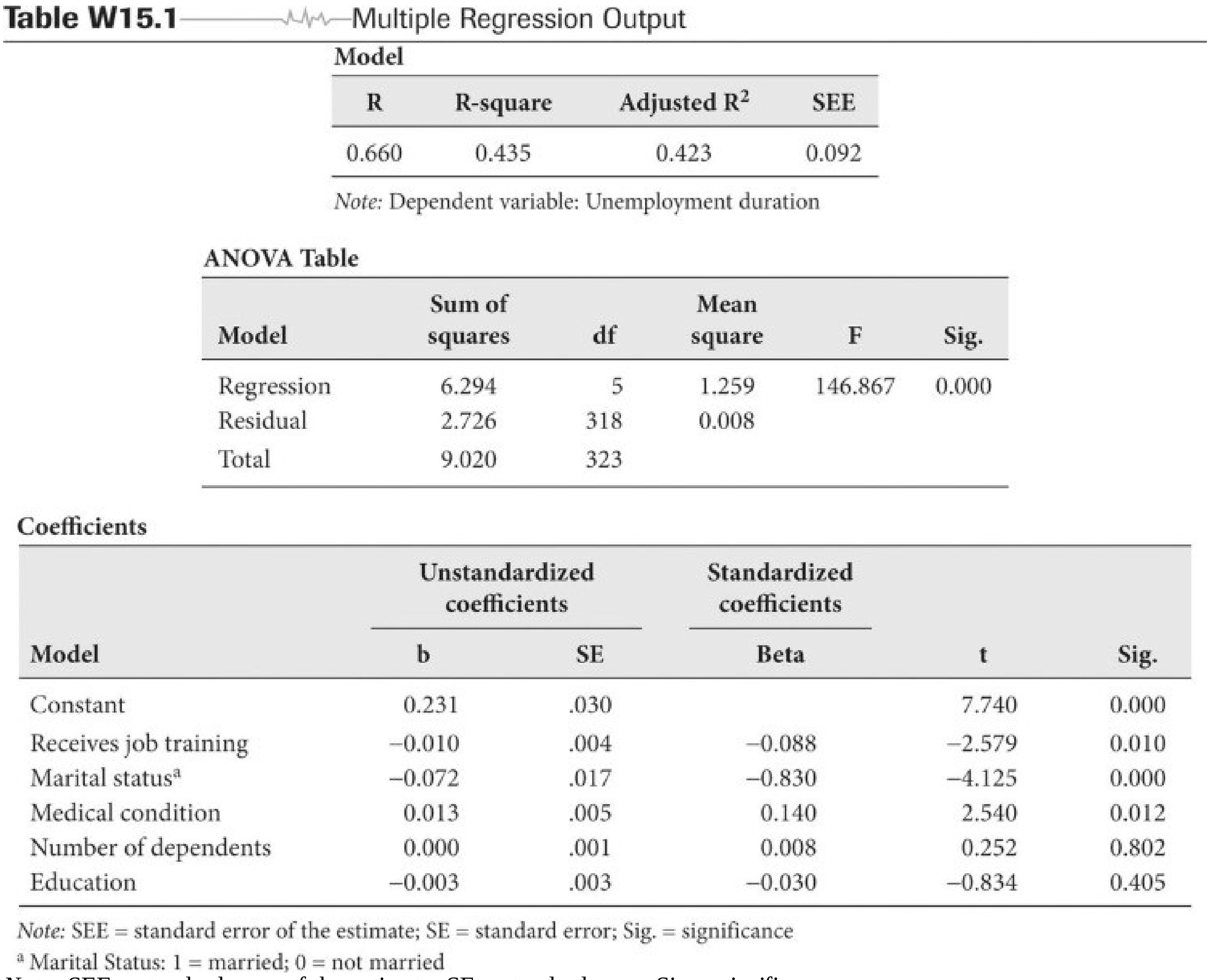


Table W15.1 Multiple Regression Output

Model

R	R-square	Adjusted R^2	SEE
0.660	0.435	0.423	0.092

Note: Dependent variable: Unemployment duration

ANOVA Table

Model	Sum of squares	df	Mean square	F	Sig.
Regression	6.294	5	1.259	146.867	0.000
Residual	2.726	318	0.008		
Total	9.020	323			

Coefficients

	Unstandardized coefficients		Standardized coefficients		
Model	b	SE	Beta	t	Sig.
Constant	0.231	.030		7.740	0.000
Receives job training	–0.010	.004	–0.088	–2.579	0.010
Marital status[a]	–0.072	.017	–0.830	–4.125	0.000
Medical condition	0.013	.005	0.140	2.540	0.012
Number of dependents	0.000	.001	0.008	0.252	0.802
Education	–0.003	.003	–0.030	–0.834	0.405

Note: SEE = standard error of the estimate; SE = standard error; Sig. = significance

[a] Marital Status: 1 = married; 0 = not married

Note: SEE = standard error of the estimate; SE = standard error; Sig. = significance

[a] Marital Status: 1 = married; 0 = not married

3. **Develop a fully specified model of the factors that affect fundraising at a nonprofit social services organization. Focus on different categories of factors that affect fundraising, and consider how some might be grouped as index variables.**

4. **Develop a fully specified model of the factors that affect teenage violence in high schools.**

____

____

____

____

____

____

5. **Explain how regression can be used for prediction. What problem do you see with predicting dependent variable values based on independent variable values that lie outside the range of observations that have been used to estimate the model?**

____

____

____

____

____

____

6. **Consider a hypothetical variable Race that is coded as 1 = Caucasian, 2 = Native American, 3 = African American, 4 = Asian or Pacific Islander, and 5 = Other. Recode this variable as a dummy variable that indicates whether the respondent is Caucasian.**

____

____

____

____

____

____

7. **Explain in your own words why the error term plot should show a random pattern. Also, which variables are shown on the axes?**

____

____

8. **Discuss, in practical terms, how analysts determine which observations are outliers. Discuss also how you would justify the deletion of observations that are outliers.**

9. **Draw an error term plot that shows the presence of an outlier (without peeking in the textbook!).**

10. **If two variables are highly correlated with each other (for example, *r*^{2} = 0.65), does it follow that they are multicollinear as well?**

____

____

____

____

____

____

11. **Explain the problem of including irrelevant variables.**

____

____

12. **Explain why time series data often lead to correlated error terms.**

Data-Based Exercises

1. **Consider the model shown in Table 15.1 of the textbook. In the Productivity dataset, the dependent variable is Productivity, and the independent variables are labeled, respectively, Teamwork, Jobknowl, Jobauthr, Wrkdyssk, and Recogawd. After you rerun the model shown in Table 15.1, create dummy variables for each of the four departments, and add to the model dummy variables for the first three departments. Are departments associated with employees' perceptions of productivity controlled for all other factors of the model? If yes, which departments are significantly associated with employees' perceptions of productivity? Record your answers on a separate sheet.**
2. **What contributes to a government's spending on conservation? Many people believe that socioeconomic and political pressures are the reasons why a government spends on conservation. They think that conservation spending is a response to population growth and economic development. Also, they believe that young educated people with Democratic party affiliation are more likely to vote for conservation spending. Use the dataset Data_FL_County to test this argument. Use the conservation spending in 2008 as the dependent variable (ConsTotal08). The dataset includes many sociopolitical and economic variables. You may want to consider to include measures of education, income, population growth, manufacturing industry base, and political voting behaviors. Record your answers on a separate sheet.**
3. **Use the Public Perceptions dataset. Examine a multiple regression model in which the dependent variable is Posview ("I have a positive view of Orange County"). The independent variables are Quality ("The quality of life in Orange County is good"), Interest ("I believe that the county is interested in what I have to say"), Respect ("The employees treated me with courtesy and respect"), Trust ("Do you trust Orange County to do what is right most of the time?"), and Works ("Do you believe that Orange County works efficiently?"). Examine the error term plot, and write up your findings. Record your answers on a separate sheet.**
4. **Regarding the model in Exercise 2, consider the effects of race and Hispanic origin on having a positive view. Do race and Hispanic origin affect overall perceptions when controlled for other variables mentioned in the earlier exercise? Record your answers on a separate sheet.**
5. **Use the Time dataset. This exercise illustrates the problem of outliers. In this exercise, we examine the effect of water pollution on the concentration of fish in a lake. The dataset contains two variables, Fishcon (the concentration of fish) and Contam1 (the concentration of a water pollutant). The observations are drawn from different parts of a large lake in order to test the hypothesis that alleged water pollution is affecting the stock of fish.**
    1. Examine and discuss whether the variables Fishcon and Contam1 are approximately normally distributed.

       ____

2. Make a scatterplot of Fishcon and Contam1 for the purpose of getting a visual read of the relationship. Is the relationship positive or negative? Is it strong or weak? Are there any possible outliers? Print the output, and write up your answers.

3. Based on your visual read, you wish to test the hypothesis that the two variables are related. Conduct a simple regression, with Fishcon as the dependent variable. Examine the error term plot for possible outliers. What do you conclude?

4. Identify the outlier, and eliminate it from subsequent analysis. Provide a write-up of the modified analysis that includes (1) the hypothesis, (2) the relationship between the variables ($y = ax + b$), (3) the identification and removal of outliers, (4) the correlation and percentage of variance explained (r and r^2), and (5) a plot of the standardized residuals against the predicted values of the dependent variable.

6. **Use the Time dataset. This exercise illustrates the use of dummy variables and includes a test of multicollinearity. The dataset contains observations from 35 hypothetical cities regarding the use of citizen focus groups in various departments (Focus). The data are based on a survey. Most variables are index variables taken from different survey questions. The variables are defined as follows:**

   Focus = A composite measure of the breadth and depth of the use of citizen focus groups in a city. Varies from 0 (low) to 20 (high).
   Mgrint = A measure of the interest of the city manager in obtaining citizen-based feedback. Varies from 1 (low) to 4 (high).
   Pubcompl = A measure of public complaints about the quality and effectiveness of a wide range of municipal services. Varies from 1 (low) to 8 (high).
   Budget = Indicates whether municipal budgets have increased in the past two years. Values are –1 = decrease in budget; 0 = no change in budget; 1 = increase in budget.
   Size = City size. Varies from 1 = small to 7 = large.
   Region = An indicator variable of the region in which the city is located. Values: 1 = Northeast; 2 = South; 3 = Midwest; 4 = West.
   In this analysis, you wish to understand which variables cause cities to use citizen-based focus groups.

   1. Briefly state your hypotheses.

   2. Provide a brief description of variables (univariate analysis). Also, examine the bivariate associations between Focus and Mgrint, Pubcompl, Budget, and Size. Examine the bivariate relationships that exist between all possible pairs of these variables.

   3. Examine whether Mgrint, Pubcompl, Budget, and Size are statistically associated

with the use of focus groups. For this purpose, conduct a multiple regression in which the dependent variable is Focus and the independent variables are Mgrint, Pubcompl, Budget, and Size. What do you conclude? (You need not write up these results.)

____

____

____

____

____

4. You now wish to include the effect of region in your analysis. To do this, you must make Region a dummy variable. Assume you wish to compare the effect of the different regions with Midwest. To make the dummy variables for the Northeast, South, and West, you must either recode the data for each region or enter the data manually. Through either of these techniques, add Northeast, South, and West to your model and rerun the analysis.

____

____

____

____

____

5. Provide a complete write-up of the regression. (1) Identify which variables are significant and at what level. Identify the adjusted R-square statistic (is it a strong or weak association?). Also, discuss the beta coefficients. What do you conclude? (2) Plot the standardized residuals against the standardized predicted values. What do you conclude? (3) To examine the possibility of multicollinearity, examine the bivariate correlations among the independent variables. What do you conclude? Given the significance of coefficients, is multicollinearity a problem here?

____

____

____

____

7. **Use the Time dataset. This exercise illustrates the problem of nonlinearity. It is commonly hypothesized that crimes are more frequent in large cities. The dataset contains the following variables:**

   Nvcrime = Index of nonviolent crimes in a given year.

   Citysize = City size.

   1. To test the above hypothesis, regress Nvcrime as a function of Citysize. Plot the standardized error term against the standardized predicted values and observe the curvilinear (parabolic) slope of the error terms.

   2. You suspect that this problem occurs because the relationship between the variables is nonlinear. To examine this possibility, plot Nvcrime against Citysize. Notice the faint rounded curve. Based on the nature of the curve, you decide that a transformation of Citysize is required. You decide to try a square root of Citysize. Make a square root transformation of Citysize. To evaluate the success of this transformation, plot the transformed variable, called TCity (on the X-axis), against Nvcrime (on the Y-axis). Is the relationship still nonlinear? Was the transformation successful?

   3. Next, rerun the regression using the transformed variable TCity. Hence, Nvcrime = *f*(TCity). Reevaluate the residuals by plotting the error term of the transformed model against the transformed independent variable, TCity. What pattern, if any, do you see? If the transformation was unsuccessful, try a log transformation. Do you see a pattern now?

8. (*optional*) **Refer to Data-Based Exercise 8 in Chapter 14. Add a second independent variable and run a multiple regression that examines the impact of these two variables on your academic performance (the dependent variable). Perform related tests of regression assumptions. Interpret the results. What do you conclude? Write a report that summarizes your findings for a person who knows nothing about statistics.**

Further Reading

Many statistics books discuss regression analysis, including those mentioned in previous chapters of this workbook. However, some of these discussions are a little superficial. For a brief introduction, see Larry Gonick and Woollcott Smith, *The Cartoon Guide to Statistics* (New York: HarperPerennial, 2005), or Sam Kash Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed. (New York: Radius Press, 1991).

For a more solid introduction, try William M. Mendenhall, *A Second Course in Statistics: Regression Analysis*, 7th ed. (Upper Saddle River, N.J.: Prentice Hall, 2011). Another good presentation is found in David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, and Eli S. Rosenberg, *Applied Regression Analysis and Multivariable Methods*, 5th ed. (Pacific Grove, Calif.: Brooks Cole, 2013). Try also McKee J. McClendon, *Multiple Regression and Causal Analysis* (Long Grove, Ill.: Waveland Press, 2002).

The following books are a little advanced but also quite useful. Two classic textbooks with an economics orientation are Robert Pindyck and Daniel Rubinfield, *Econometric Models and Economic Forecasts*, 3rd ed. (New York: McGraw-Hill, 1997), and Damodar Gujarati and Dawn Porter, *Basic Econometrics*, 5th ed. (New York: McGraw-Hill, 2008). A well-respected general social science book is Michael Kutner, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill, 2013).

Chapter 16 Logistic Regression and Time Series Regression

Q & A

1. ***When should logistic regression be used? Provide some examples.***

   Logistic regression should be used when the dependent variable is dichotomous, that is, when it has only two values. Some typical examples of a dichotomous dependent variable are whether or not someone got elected, whether or not a war occurred, and whether or not a medical event occurred. In these instances, either something happened or it didn't—a dichotomous situation.
2. ***What regression problem does logistic regression address?***

   Multiple regression assumes a continuous dependent variable. Logistic regression uses a dichotomous dependent variable, that is, one that has only two values, such as zero and one.
3. ***How does the estimation model of logistic regression differ from multiple regression?***

   Whereas the multiple regression estimation model uses a straight line, logistic regression uses a logistic curve, which is S-shaped. Both approaches select a model that best fits the observed observations.
4. ***What is the log likelihood value, −2LL?***

   The log likelihood value, −2LL, is a quantitative measure of how well the model predicts observed values. Better-fitting models have smaller values of |−2LL|.
5. ***How is a classification table used to assess the goodness of fit in logistic regression?***

   A classification table shows the percentage of corrected predicted observations. The minimum is 50 percent, indicating the lack of any useful prediction. Typically, standards of 80–85 percent indicate good model prediction.
6. ***What are Nagelkerke* $\mathbf{R}^2$ *and Cox and Snell* $\mathbf{R}^2$*?***

   These are measures of association. Higher values imply a better fit. Nagelkerke R^2 has a variance-explained interpretation.
7. ***What test is used for the statistical significance of logistic regression coefficients?***

   In logistic regression, Wald chi-square is used as a test statistic to determine the statistical significance of logistic regression coefficients.
8. ***How can logistic regression coefficients be used to calculate event probabilities?***

   The predicted values of logistic regression coefficients can be used to show the probability of an event occurring, by using the following formula:

   Prob(event) = $1/[1 + e^{-Z}]$, where $Z = a + b_1x_1 + b_2x_2 + \ldots$

   Z is also called the logit.
9. ***What is an odds ratio?***

An odds ratio is used to compare the probability of something occurring, as compared to it not occurring. One way of using this measure is to observe how the odds ratio changes when only one independent variable is changed, such as by one unit.

10. ***What are time series data?***

    Time series data are data that have been collected over time, such as periodic assessments of performance outcomes or public (or client) opinions.

11. ***Which regression assumption is usually violated when using time series data? What problem does this cause?***

    With time series data, the assumption of random distribution of error terms usually is violated. Specifically, the error term plot, when plotted against the sequence of time-ordered observations, typically exhibits a pattern. This is called autocorrelation or serial correlation. The problem with autocorrelation is that it severely exaggerates the statistical significance of variables, leading to the erroneous conclusion that variables are statistically associated when they are not.

12. ***How is autocorrelation detected?***

    Autocorrelation can be detected by plotting the error term against time. It is formally detected by testing the Durbin-Watson test statistic: values close to 2 indicate the absence of serial correlation, whereas values closer to 0 and 4 may indicate serial correlation.

13. ***How is autocorrelation addressed?***

    Two strategies are available for correcting serial correlation: the first strategy is to add a trend variable to the model, and the second strategy is to examine the relationship in so-called first-order differences. Relationships in first-difference form often eliminate problems of serial correlation because differenced data exhibit far more variability than do levels data. By contrast, adding a trend variable is a prophylactic strategy that attempts to control for the problem. Regression of first-order differences is considered a more stringent test.

14. ***How is time series regression used for evaluating the impact of policies?***

    Time series data are excellent for evaluating the impact of a policy or program. Levels of performance or service utilization are tracked and compared with the moment or period in which a policy is implemented.

15. ***What are policy variables?***

    Policy variables measure when and how policies affect the dependent variables. Four types of policy variables are those that model pulse, period, step, and increasing impacts.

16. ***What are lagged variables?***

    Lagged variables are independent variables that have a lagged effect on the dependent variable.

Critical Thinking

1. **Identify dichotomous variables that are relevant to your area of interest, and develop a model to predict them.**

2. **Explain the following statement: “The dichotomous nature of the dependent variable violates an assumption of multiple regression.”**

3. **The S-shaped logistic curve has values that lie between 0 and 1. What problem does this address?**

4. **Explain why a good model fit will show a Hosmer and Lemeshow test statistic that is**

insignificant.

5. **Verify the event probability calculations shown in Table 16.2 of the textbook.**

6. **Table W16.1 shows the output of a logistic model that predicts promotion within 5 years. Is the model adequate? Why or why not? What is the probability that a 37-year-old female who has a performance appraisal rating of 4 will be promoted? The performance variable (Appraisal) is measured on a scale of 5 = high to 1 = low, Gender is defined as 1 = male, 2 = female, and Age = employee's age (in years).**

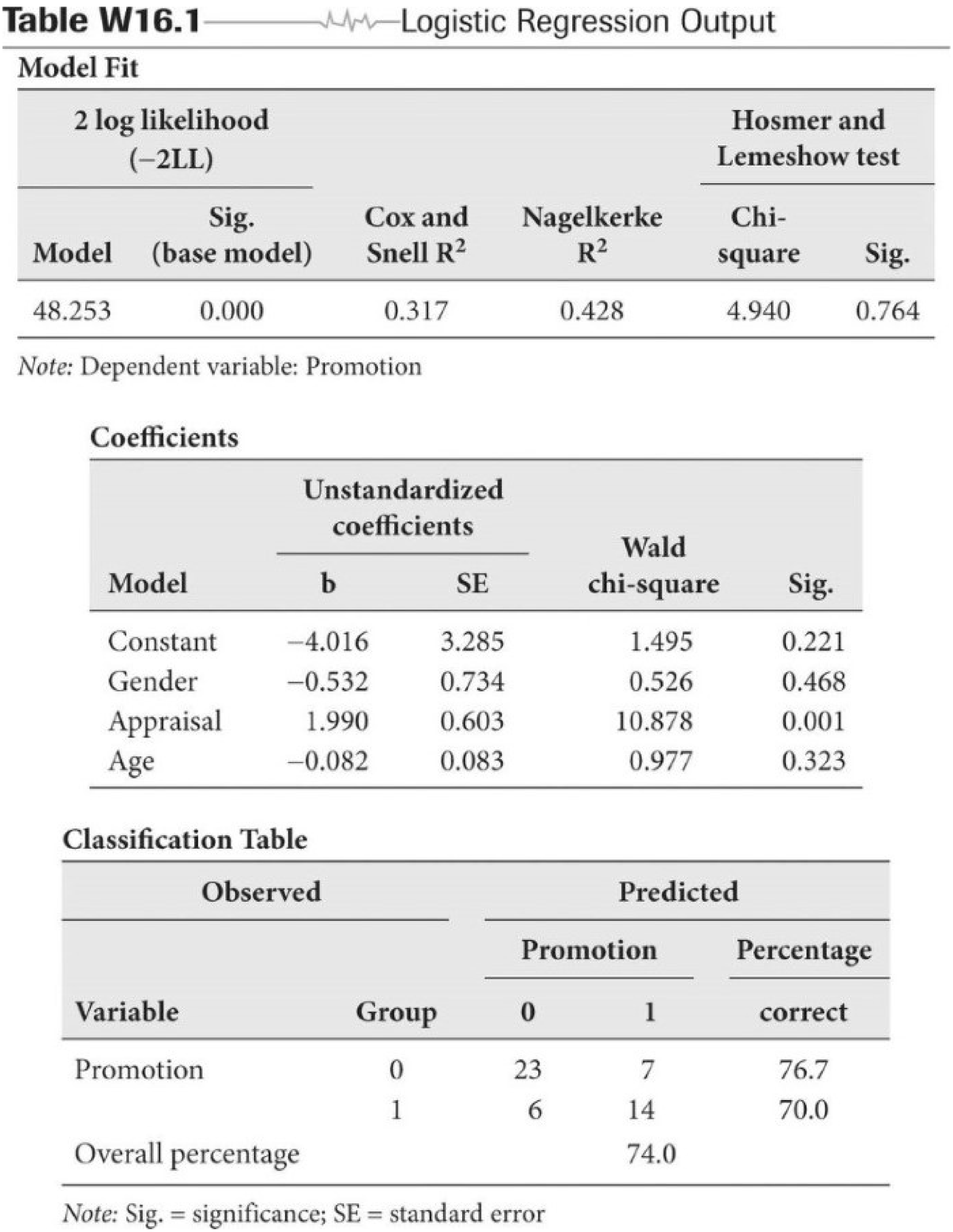


Table W16.1 Logistic Regression Output

Model Fit

2 log likelihood (−2LL)				Hosmer and Lemeshow test	
Model	Sig. (base model)	Cox and Snell R²	Nagelkerke R²	Chi-square	Sig.
48.253	0.000	0.317	0.428	4.940	0.764

Note: Dependent variable: Promotion

Coefficients

	Unstandardized coefficients			
Model	b	SE	Wald chi-square	Sig.
Constant	−4.016	3.285	1.495	0.221
Gender	−0.532	0.734	0.526	0.468
Appraisal	1.990	0.603	10.878	0.001
Age	−0.082	0.083	0.977	0.323

Classification Table

Observed		Predicted		
		Promotion		Percentage
Variable	Group	0	1	correct
Promotion	0	23	7	76.7
	1	6	14	70.0
Overall percentage			74.0	

Note: Sig. = significance; SE = standard error

Note: Sig. = significance; SE = standard error

7. **Give some examples of time series data in your area of interest. Discuss ways in which these data might be analyzed.**

8. **Explain why autocorrelation is almost always present when using regression analysis with time series data.**

9. **The Durbin-Watson test statistic for autocorrelation of a regression model with 35 observations and four independent variables is 1.98. What do you conclude? And what would you conclude if the Durbin-Watson test statistic were 1.08?**

10. **Explain the following statement: “The regression of first-order differences is considered a far more stringent test than adding a trend variable.”**

11. **Explain how policy variables help to evaluate program effectiveness. Does the magnitude of these dummy variables matter?**

12. **Give an example of an independent variable that you suspect might have a lagged impact on a dependent variable.**

Data-Based Exercises

1. **Use the Public Perceptions dataset. Predict Trust (“Trust in government”) as a function of Interest (“County officials are interested in what I have to say”), Works (“Do you believe the county government works efficiently?”), Quality (“The quality of life in Orange County is good”), and race. To this end, (1) create a dummy variable for race (1 = white, 0 = all others) from the variable Ethnic, and (2) recode the values of “DK” for the variables Trust and Works as missing. Use logistic regression, and report the results.**

   Trust = *f*(Interest, Works, Quality, White)

   1. Verify that the dependent variable is dichotomous.

      ____

      ____

      ____

   2. Does the model satisfy the standard for correctly predicted observations?

      ____

      ____

      ____

   3. Which variables are significant?

      ____

      ____

      ____

   4. Calculate event probabilities.

      ____

      ____

      ____

2. **Use the Public Perceptions dataset. Predict Manage (“Doing a good job managing growth”) as a function of Works (“Do you believe the county government works efficiently?”), Quality (“The quality of life in Orange County is good”), and Watch (“Watching Orange County TV”). Use logistic regression, and report the results.**

Manage = *f*(Works, Quality, Watch)

1. Verify that the dependent variable is dichotomous.

   ____

   ____

   ____

2. Does the model satisfy the standard for correctly predicted observations? If not, does it satisfy the Hosmer and Lemeshow test?

   ____

   ____

   ____

3. Which variables are significant?

   ____

   ____

   ____

4. Calculate event probabilities.

   ____

   ____

   ____

3. **Various universities make datasets of past research available. One such university is the University of California, which makes datasets available through SDA: Survey Documentation and Analysis (http://sda.berkeley.edu). Explore available datasets through the Archive link. For example, select any dataset (e.g., ANES Cumulative Datafile 1948–2000), and note the tab/option to run logit regression. However, you should first examine the variable descriptions by selecting "Open Extra Codebook Window." You will likely need to recode some of the variables first. You can also download a customized subset of data.**
4. **Use the Time dataset. This exercise illustrates the autocorrelation problem in the use of policy variables. Time series data are common in program evaluation and public policy. These data examine the impact of a law that, among other things, increases jail time for those convicted of driving under the influence (DUI). The dataset contains the following variables:**

   Fatal = Traffic fatalities per 100,000 miles driven.

Year = Year of traffic fatalities measured.

Short = A dummy variable identifying when the policy intervention occurred, namely, in 1980 when a law was passed that requires mandatory jail time for DUI. The values are 0 = pre–law adoption (pre-1980) and 1 = post–law adoption (post-1980).

Long = A dummy variable identifying the number of years of post–policy adoption. The values are 0 for pre–law adoption, 1 for 1980, 2 for 1981, 3 for 1982, and so forth. This variable gives weight to the long-term effect of policy.

Jailtime = Days of jail time served by offenders as a result of the law.

Note: Although increased jail time is the primary effect of the program, data from other aspects of the program (such as midnight checkpoints) are not available to the researcher. The variables Short and Long model these unknown effects.

1. You wish to evaluate whether the new law has been effective. You model the effect on fatalities and explore the possibility of autocorrelation in time series data. Hence, Fatal = *f*(Short, Long, Jailtime). Run the regression analysis, and examine the error term for autocorrelation. Specifically, examine the Durbin-Watson test statistic. What do you conclude?

   ____

   ____

   ____

   ____

   ____

2. Now, take the trend variable into account. Add Year as an independent variable to your list. Again, note the Durbin-Watson statistic. What do you conclude about adding the trend variable? Provide a complete write-up.

   ____

   ____

   ____

   ____

   ____

3. For further rigor, examine the relationship in first-order differences. Make first-order differences, and examine DFatal = *f*(Short, Long, DJailtime). What do you conclude? Also, try lagging each of the independent variables for up to two periods. What do you

conclude?

____

____

____

____

____

5. **Use the Crime dataset. The data are monthly observations of juvenile arrests in a city with a curfew for teenagers. The curfew prohibits teenagers from being outside between 8 p.m. and 1 a.m.**

   Note: These data are adapted from actual cities. Although these are monthly data, the same approach can be applied to weekly or daily data. In those cases, rather than controlling periodicity by month, control variables control for periodicity by week or day.

   1. Plot the number of juvenile arrests (Juvarsts) over time. What do you conclude about the impact of the curfew? What factors might underlie the periodicity?

      ____

      ____

      ____

      ____

      ____

   2. Run the regression of Curfew against Juvarsts. What do you conclude from the Durbin-Watson statistic? To attempt to correct for autocorrelation, include both the year and each of the months as control variables. Run the model in both levels and first-order difference forms, hence:

      Juvarsts = *f* (Curfew, Month 1 – Month 11, Year), and
      DJuvarsts = *f* (DCurfew, Month 1 – Month 11, Year).
      Which of these models deals with the problem of autocorrelation? Why is Month 12 excluded from the model?

      ____

      ____

      ____

      ____

____

3. It is sometimes noted in the literature that law enforcement displaces crime rather than reduces it. Rather than committing crimes between 8 p.m. and 1 a.m., juveniles will target other time periods to commit crimes. The dataset includes the variable Crimarst, which measures total monthly juvenile arrests. Run the same models as under item b, with Crimarst as the dependent variable. What do you conclude?

   ____

   ____

   ____

   ____

   ____

Further Reading

Numerous advanced texts are available on the topics discussed in this chapter. For a discussion of logistic regression, see Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson, *Multivariate Data Analysis*, 7th ed. (Upper Saddle River, N.J.: Prentice Hall, 2009). This is an advanced but nonmathematical introduction. Another text is David Hosmer and Stanley Lemeshow, *Applied Logistic Regression*, 3rd ed. (New York: Wiley, 2013). Many of the previously mentioned texts that discuss regression also cover logistic regression. Discussions of time series regression tend to be heavily mathematical, but many textbooks do provide a good conceptual foundation. See William M. Mendenhall, *A Second Course in Statistics: Regression Analysis*, 7th ed. (Upper Saddle River, N.J.: Prentice Hall, 2011). The research literature includes many examples of logistic and times series regression. You can easily search for them on Google Scholar, in journals of your interest, using the search terms “logistic regression” or “times series regression.”

Chapter 17 Survey of Other Techniques

Q & A

1. ***What is path analysis, and what advantage does it offer over multiple regression?***

   Path analysis is a causal modeling technique. Unlike multiple regression, path analysis allows for direct and indirect effects of variables. Path analysis is a recursive causal modeling technique, that is, one that does not allow feedback loops.
2. ***How are paths estimated?***

   Each path is estimated separately using regression (ordinary least squares, or OLS).
3. ***What is an exogenous variable? What is an endogenous variable? What purpose do these terms serve?***

   Causal modeling distinguishes between *exogenous* variables, which are variables that are unaffected by other variables in the model, and *endogenous* variables, which are affected by other variables. The distinction is useful because in causal models variables can be both independent and dependent.
4. ***What assumptions must path analysis models satisfy?***

   Path analysis must satisfy all OLS assumptions (see Chapter 1 in the textbook). In addition, all error terms must be uncorrelated with all exogenous variables. As in regression, the model must be theory based.
5. ***How are direct and indirect effects of variables calculated?***

   Direct effects are simply the beta coefficients of the variables that immediately affect another variable. Indirect effects are calculated as the product of beta coefficients of each pathway. See Table 17.1 in the textbook for an example of such calculations.
6. ***What is a structural equation model, and how is it different from path analysis?***

   Structural equation models simultaneously estimate relationships among observed variables and factor constructs. Unlike path analysis, structural equation models may involve feedback loops. Models with feedback loops are called nonrecursive.
7. ***What are censored data? Give an example.***

   Censored observations are those for which the specified outcomes have yet to occur. For example, in a study of student retention, some students have not (yet) dropped out of the program but may still do so before they graduate.
8. ***What is the purpose of a life table?***

   Life tables show the probability of a dichotomous event's occurring for each time period.
9. ***What is factor analysis?***

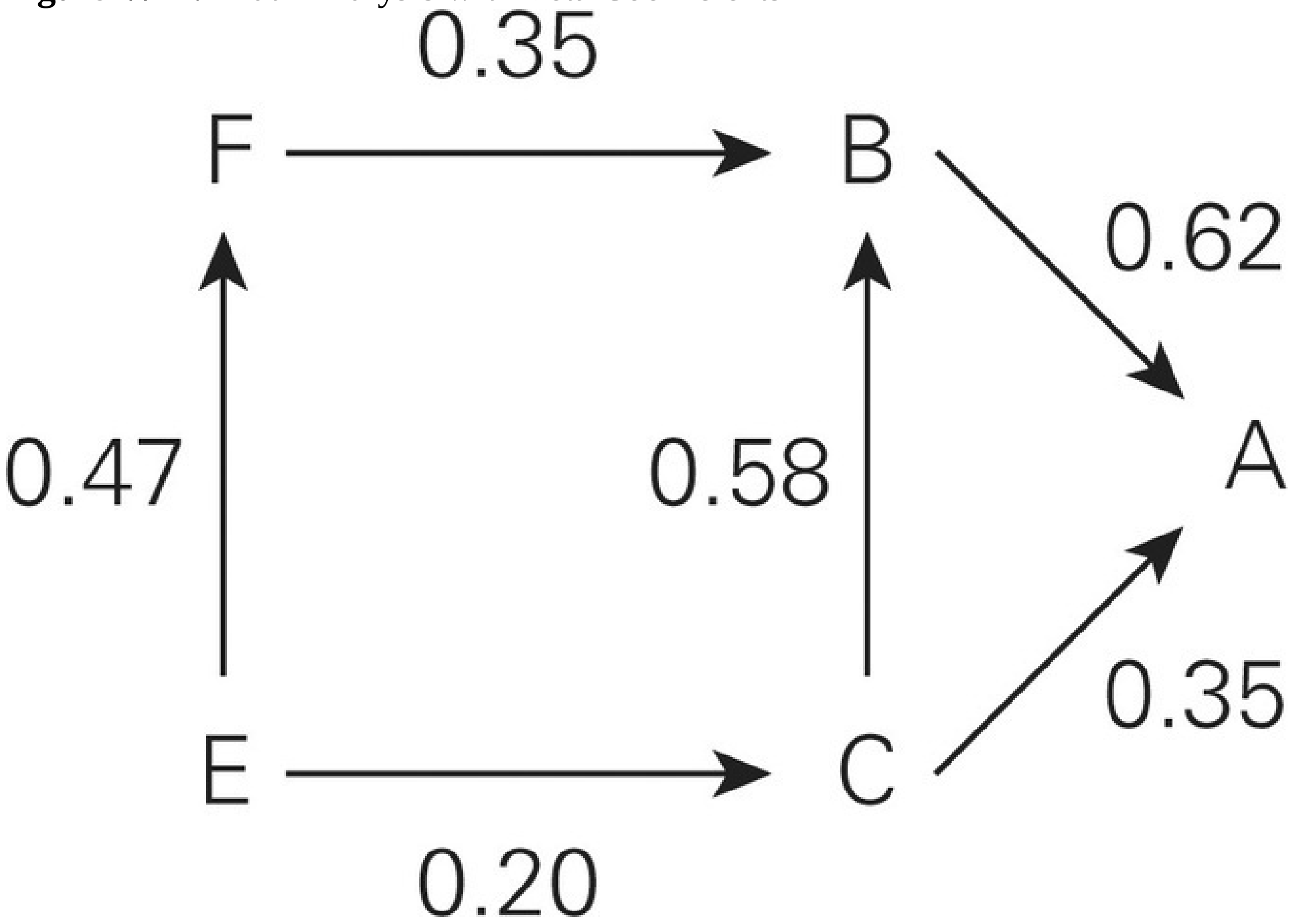
   Factor analysis is an exploratory technique that groups variables together based on their similarities and dissimilarities. Similarly grouped variables may suggest variables for subsequent index construction.

10. ***What four steps does factor analysis involve?***
    1. Determining that the group of variables has enough correlation to allow for factor analysis
    2. Determining how many factors should be used for classifying (or grouping) the variables
    3. Improving the interpretation of correlations and factors (through a process called rotation)
    4. Naming the factors and, possibly, creating index variables for subsequent analysis

Critical Thinking

1. **Draw a recursive causal model (path analysis) of factors that increase student success, measured as graduation. Which variables are direct influences, which are indirect influences, and which, if any, are both? Be sure to indicate clearly the direct and indirect influences on the dependent variable.**
2. **Figure W17.1 shows the results of a path analysis. The values above the paths are beta coefficients. Which variable has the largest influence?**

Figure W17.1 Path Analysis with Beta Coefficients



3. **Give three examples of situations that may involve censored data.**

4. **Do censored data necessarily involve data that are collected over time?**

____

____

____

____

____

____

5. **Give an example of a leading indicator of crime that might be used in regression-based forecasting.**

____

____

____

____

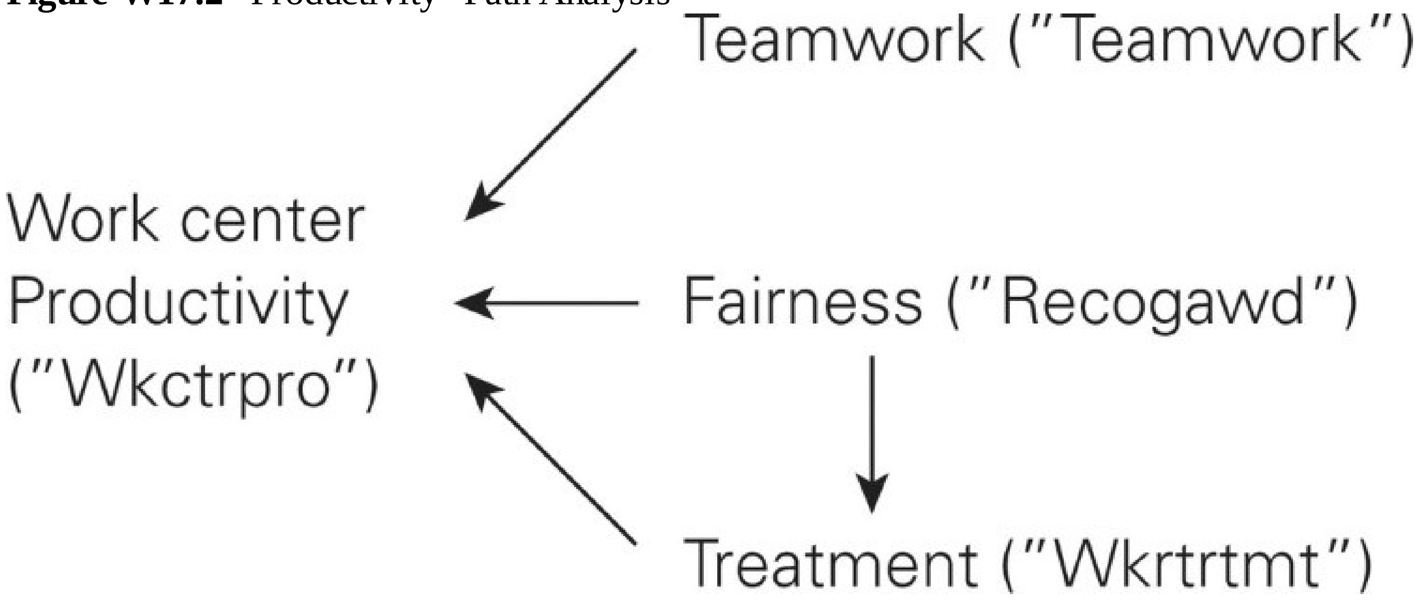
____

____

Data-Based Exercise

1. **Use the Productivity dataset. Create a path analysis of the model shown in Figure W17.2. Does the model meet the condition of not having feedback loops? Begin by specifying (that is, identifying) all the regression models that are to be estimated. For each, identify the dependent and independent variables. Then estimate ("run") the models, record the beta coefficients, and calculate direct and indirect effects.**

   Figure W17.2 "Productivity" Path Analysis



Further Reading

For a discussion of various exploratory techniques, see Sam Kash Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, 2nd ed. (New York: Radius Press, 1991). See also Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson, *Multivariate Data Analysis*, 7th ed. (Upper Saddle River, N.J.: Prentice Hall, 2009), for a more advanced discussion. This book also has an excellent, introductory chapter on ANOVA. Another book with a variety of statistical techniques is Gerald Miller and Kaifeng Yang, *Handbook of Research Methods in Public Administration*, 2nd ed. (New York: CRC Press, 2007). For survival analysis, see David Hosmer and Stanley Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data* (New York: Wiley, 2008), or search online for “IBM SPSS Advanced Statistics” or “IBM SPSS Regression 23.” For structural equation models, search online for “AMOS User’s Guide”; there are also YouTube videos that can help, such as those by James Gaskin. AMOS is very user-friendly; a trial software package is available. In recent years, Partial Least Squares has also become popular; see, for example, the SmartPLS software and the guide by Joseph F. Hair, G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)* (Thousand Oaks, Calif.: Sage, 2014).

Chapter 18 Excel User’s Guide

This chapter provides a guide on how to use Microsoft Excel in calculations and graphic presentations. We assume that you have basic knowledge about Excel, which is popular and easy-to-use spreadsheet software. If you are not comfortable with the software, visit the Microsoft Office website to take a tutorial or read an introductory guide to using Excel.

All Excel calculations introduced in this chapter are conducted with either Data Analysis ToolPak or the Insert Function (f_x) button. *Note:* Software versions change quickly these days, so perhaps some of the screenshots will be slightly different from your version.

Loading the Data Analysis ToolPak

To use the Data Analysis ToolPak, you need to load it first. To load it in Excel, click the File tab, and then click Options. Click Add-Ins, and then in the Manage box, select Excel Add-Ins. Click Go. In the Add-Ins available box, select the Analysis ToolPak check box, and then click OK. If Analysis ToolPak is not listed in the Add-Ins available box, click Browse to locate it. If you see a message that the Analysis ToolPak is not currently installed on your computer, click Yes to install it when prompted to do so. After you load the Analysis ToolPak, the Data Analysis command is available in the Analysis group on the Data tab. Loading the Analysis ToolPak is similar in older versions of Excel.

Locating the Insert Function (f_x)

You should be able to easily locate the Insert Function button, labeled with symbol f_x, next to the Formula Bar in an Excel sheet. If your Excel sheet does not show it, click the View tab, and check the Formula Bar button in the Show/Hide group. You should also find it in the Function Library group on the Formulas tab. For Excel 2003 users who can’t find it in an Excel sheet, click the Function button on the Insert tab.

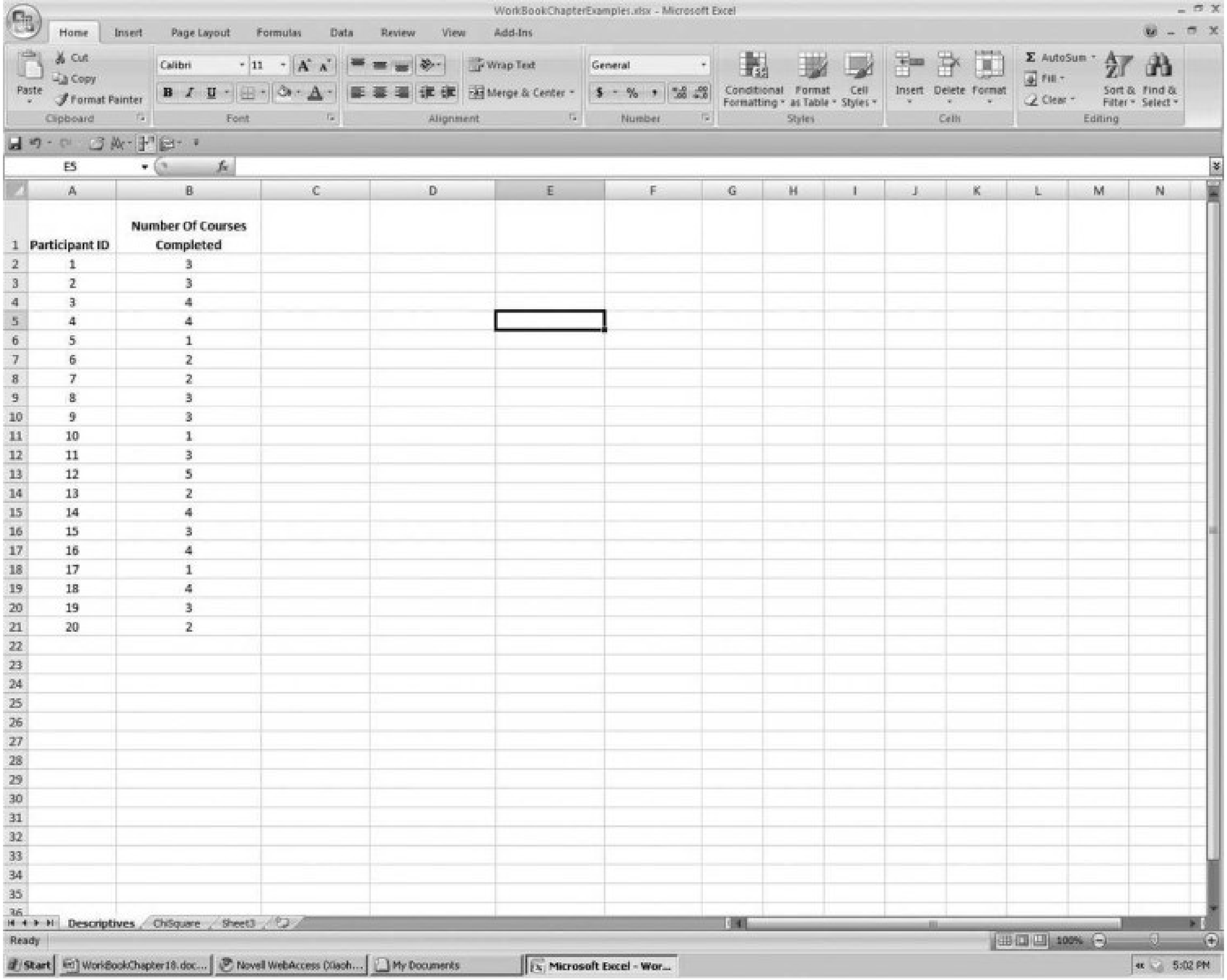
The Excel Sheet Screen and Descriptive Statistics Procedure

To use Excel, you need to open an Excel sheet. Notice that a sheet is designed for data input in columns and rows. Columns are named by letter; rows are named by number. A particular cell can be located by using a letter and a number such as A1, B2, and so on. Screen W18.1 shows the number of training courses completed by 20 participants in a job training program. The data are presented in a new Excel sheet. Notice that the first row of the sheet can be used for the labels of the variables, "Participant ID" in A1 and "Number of Courses Completed" in B1 in this example.

Excel makes it very easy to calculate and present the measures of central tendency and dispersion. You can use the Descriptive procedure in the Data Analysis ToolPak to obtain a group of such measures. Let's use the job training data as an example to illustrate this procedure. In a new Excel sheet, input the data on the number of courses completed in Column B as shown in Screen W18.2. Click the Data Analysis command in the Analysis Group on the Data tab if you use Excel 2010 or Excel 2007. If you use Excel 2003, click Data Analysis in the Tool menu. Select Descriptive Statistics in the Data Analysis window. Select the data in B1 to B21 in the Input Range (i.e., $B$1:$B$21). Click the Label in First Row box to show the number of courses completed as the title of your output. Select an Output Range that does not overlap with the data ($D$20 in this example). If you want your results in a new worksheet, select New Worksheet Ply instead of Output Range. Select Summary Statistics. Click OK. The results, shown in Screen W18.2, include the mean, median, mode, standard deviation, standard error, sample variance, range, sum, kurtosis, and skewness.

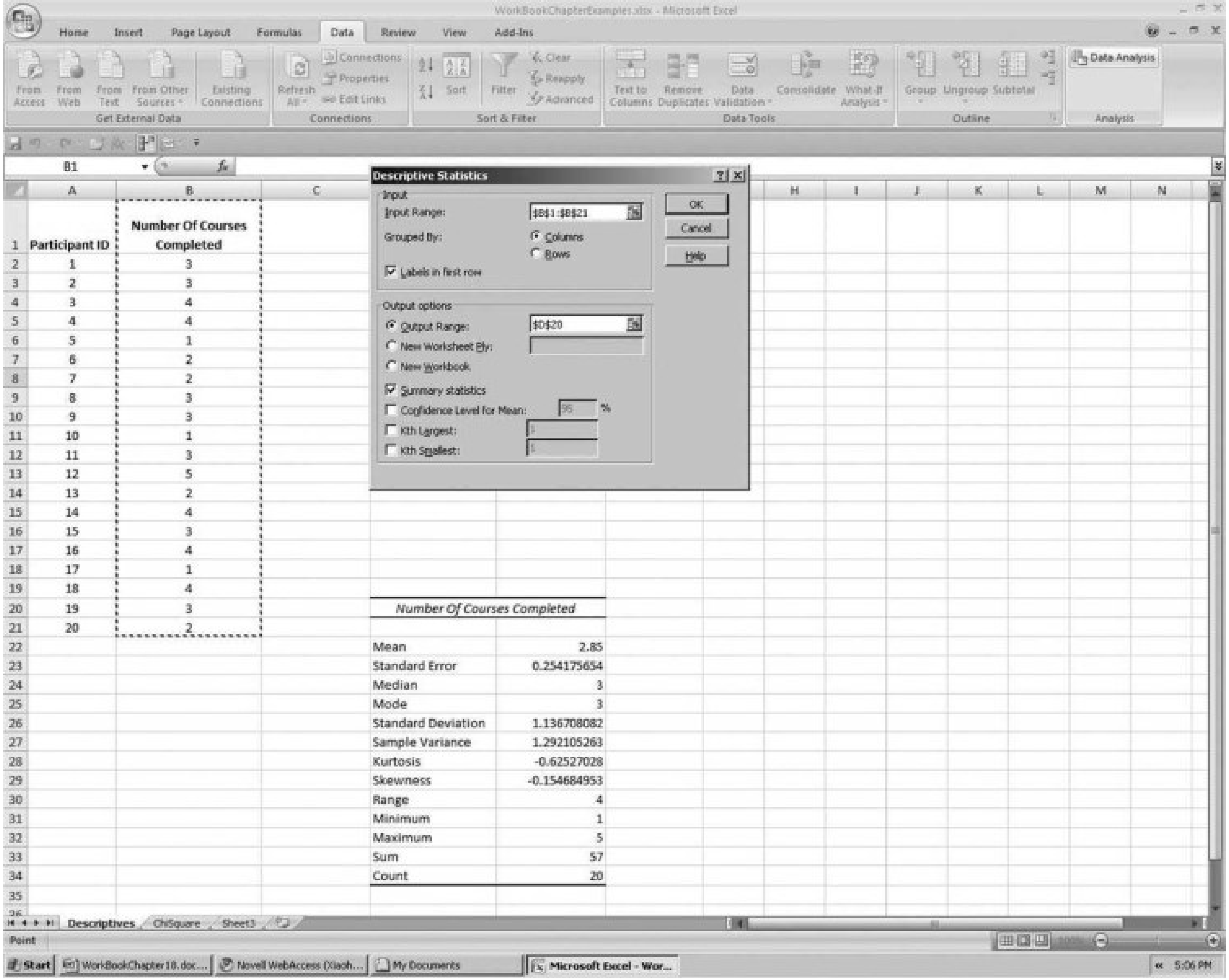
The Data Analysis Descriptive Statistics procedure gives a group of descriptive measures. If you want to obtain just one measure, such as the mean, Excel Insert Function (f_x) provides a much easier and more convenient way to do so. Let's say that you want to calculate the mean for the number of courses completed in the preceding example. Click the Insert Function (f_x) button, next to the formula bar (FORMULAS → Insert Function), to open the Insert Function dialogue box, shown in Screen W18.3. In the Insert Function dialogue box, if you know the name of the function you want to use, type the name in the Search for a Function window. If you don't know the name, select a category of formulas you want to work with from the Select a Category window. If you are not sure which category you should use, select the All category. Then click a function from the Select a Function window and read the definition of the function given below the window. Because you want to calculate the average of the data, scroll to the AVERAGE function. Select it and then click OK.

Screen W18.1 A Screen of Excel Sheet



Participant ID	Number Of Courses Completed
1	3
2	3
3	4
4	4
5	1
6	2
7	2
8	3
9	3
10	1
11	3
12	5
13	2
14	4
15	3
16	4
17	1
18	4
19	3
20	2

Screen W18.2 The Excel Descriptive Statistics Procedure



	A	B
1	Participant ID	Number Of Courses Completed
2	1	3
3	2	3
4	3	4
5	4	4
6	5	1
7	6	2
8	7	2
9	8	3
10	9	3
11	10	1
12	11	3
13	12	5
14	13	2
15	14	4
16	15	3
17	16	4
18	17	1
19	18	4
20	19	3
21	20	2

Descriptive Statistics

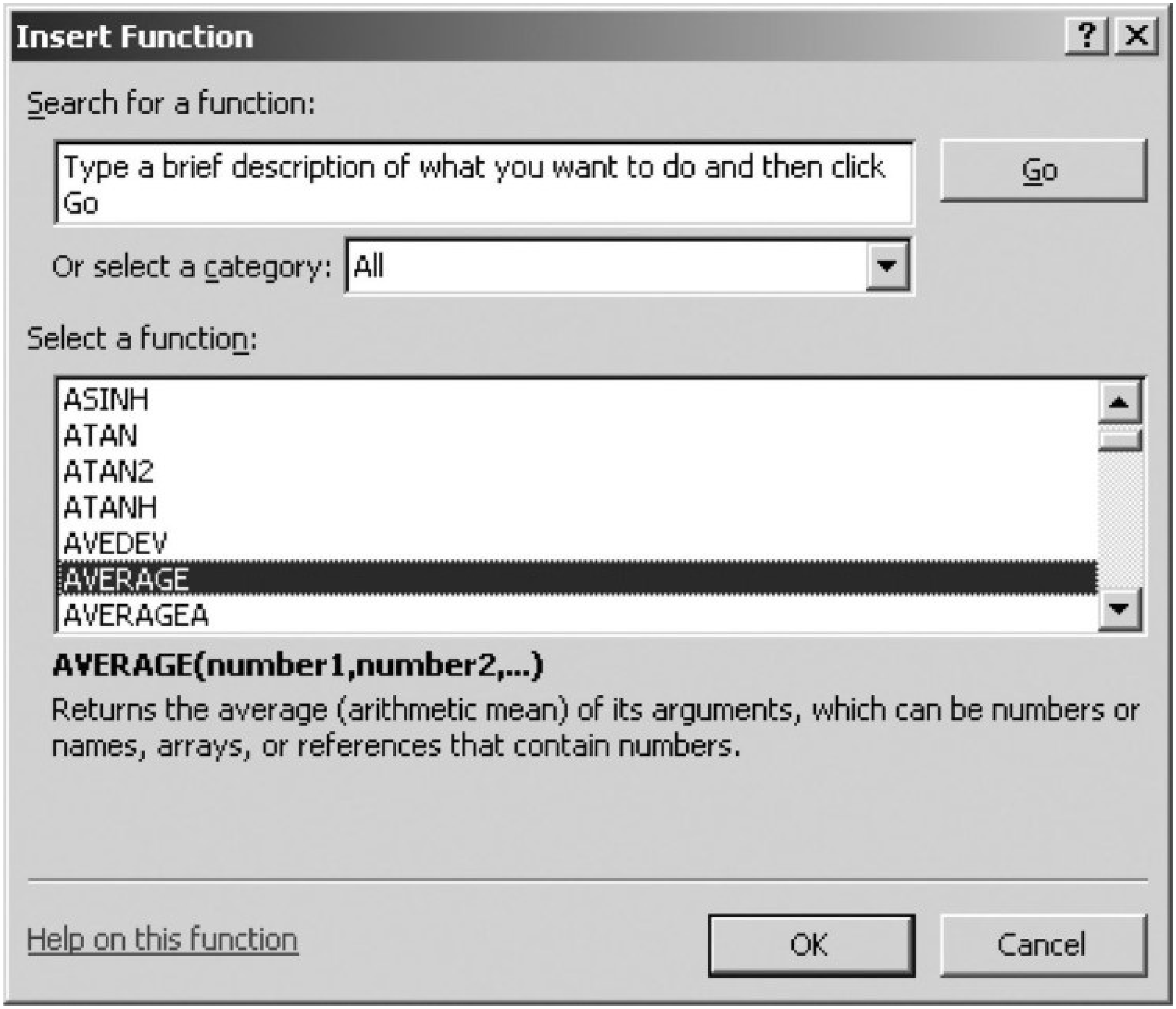
Input
Input Range: $B$1:$B$21
Grouped By: ● Columns ○ Rows
☑ Labels in first row

Output options
● Output Range: $D$20
○ New Worksheet Ply:
○ New Workbook
☑ Summary statistics
☐ Confidence Level for Mean: 95 %
☐ Kth Largest: 1
☐ Kth Smallest: 1

OK
Cancel
Help

Number Of Courses Completed	
Mean	2.85
Standard Error	0.254175654
Median	3
Mode	3
Standard Deviation	1.136708082
Sample Variance	1.292105263
Kurtosis	-0.62527028
Skewness	-0.154684953
Range	4
Minimum	1
Maximum	5
Sum	57
Count	20

Screen W18.3 Insert Function Dialogue Box



Search for a function: Type a brief description of what you want to do and then click Go

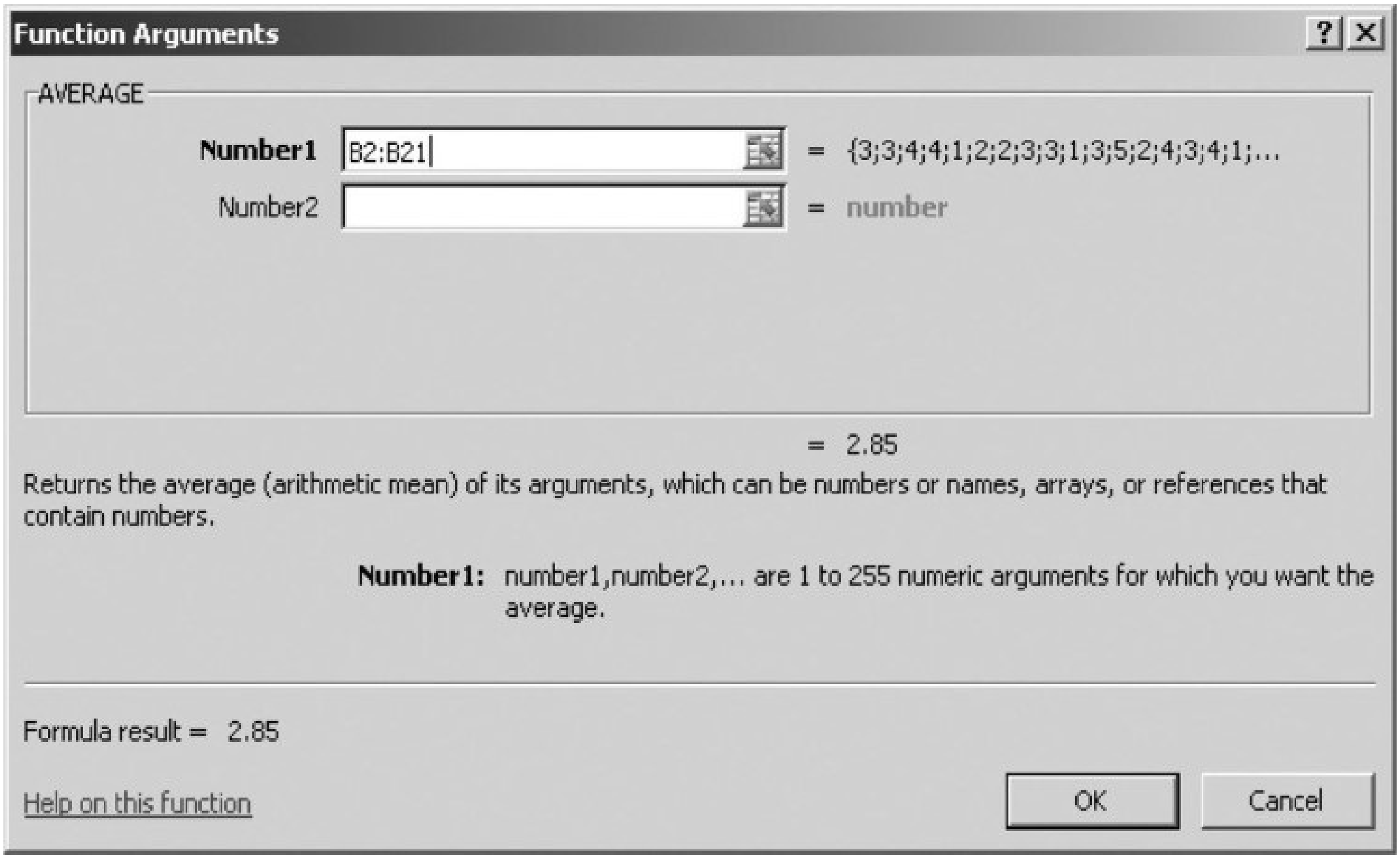
Or select a category: All

Select a function: AVERAGE

AVERAGE(number1,number2,...)

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Screen W18.4 Insert Function Argument Box



Function Arguments

AVERAGE

Number1 B2:B21 = {3;3;4;4;1;2;2;3;3;1;3;5;2;4;3;4;1;...

Number2 = number

= 2.85

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

Number1: number1,number2,... are 1 to 255 numeric arguments for which you want the average.

Formula result = 2.85

In the Function Arguments box, select data in cells from B2 to B21 (i.e., B2:B21) in the Number 1 window (the only calculation in this example). You should be able to see the 2.85 in the Formula Result, shown as in Screen W18.4. Click OK. The result should appear in a cell of your choice. You should also see the = AVERAGE (B2:B21) in the Formula Bar. In other words, you can use the Formula Bar directly to get your answer if you know the name of the function and the location of the data. Make sure you type the symbol "=" each time you use the Formula Bar and place parentheses "( )" for the data.

The Excel Insert Function is a very useful calculation tool. You can use it for many calculations in this book. Table W18.1 lists the names of some popular Insert Functions.

Excel offers another easy way to access statistical functions. Click Formulas and More Functions, as shown in Screen W18.5, to choose the statistical procedure you want. If you right-click on Statistical, you add a quick access button to Quick Access Toolbar.

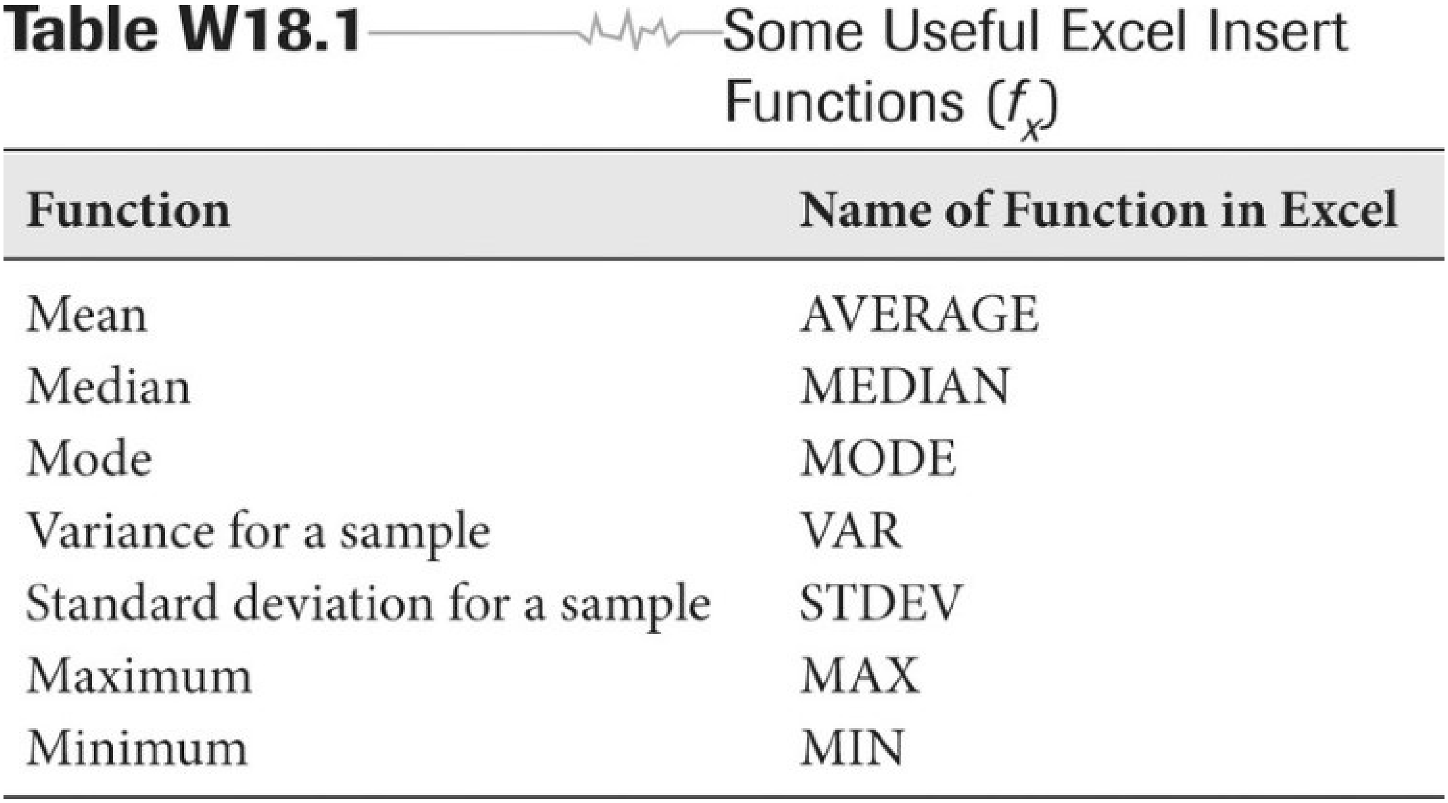
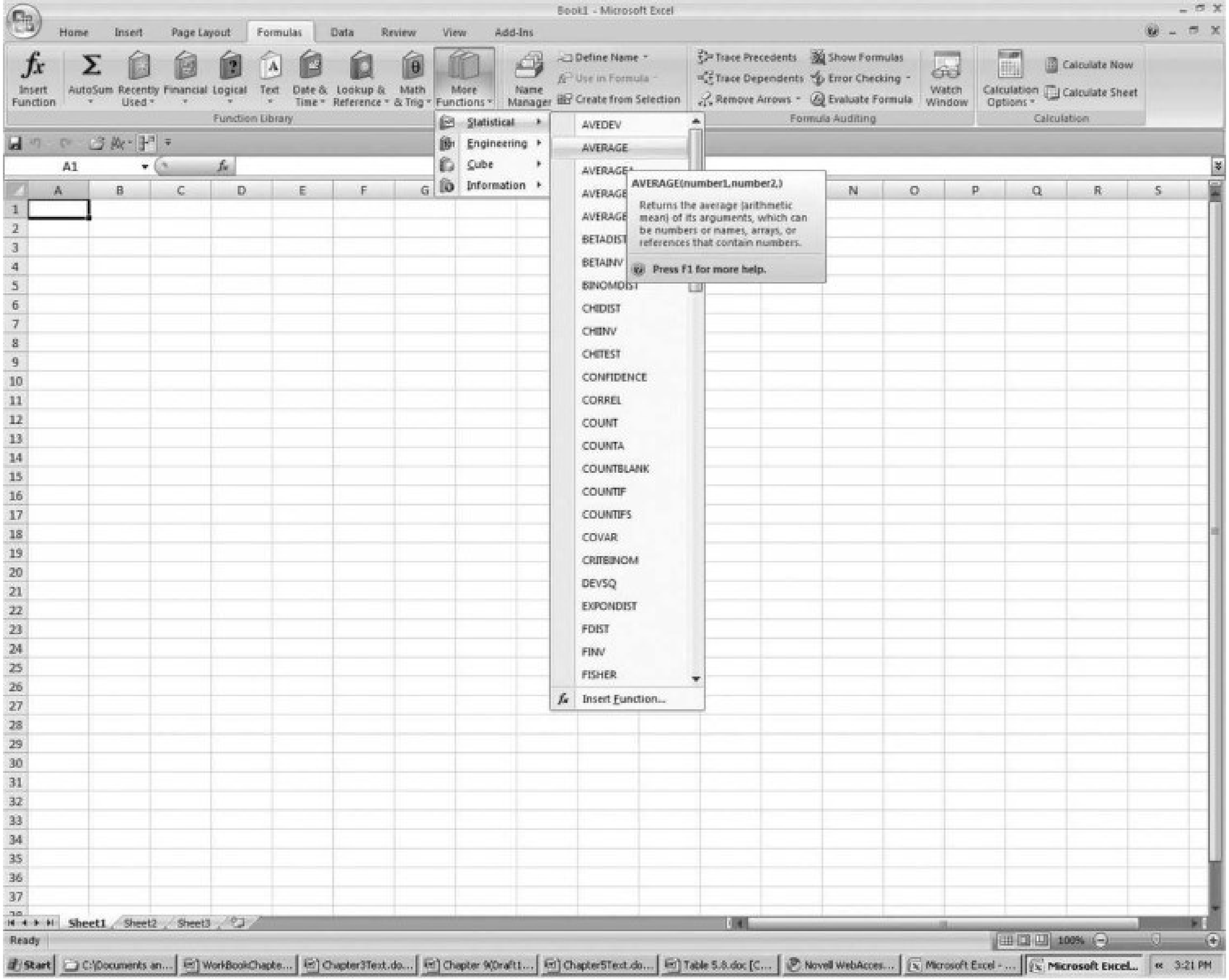


Table W18.1 Some Useful Excel Insert Functions (f_x)

Function	Name of Function in Excel
Mean	AVERAGE
Median	MEDIAN
Mode	MODE
Variance for a sample	VAR
Standard deviation for a sample	STDEV
Maximum	MAX
Minimum	MIN

Screen W18.5 Quick Access to Statistical Functions in Excel



AVERAGE(number1,number2,...)

Returns the average (arithmetic mean) of its arguments, which can be numbers or names, arrays, or references that contain numbers.

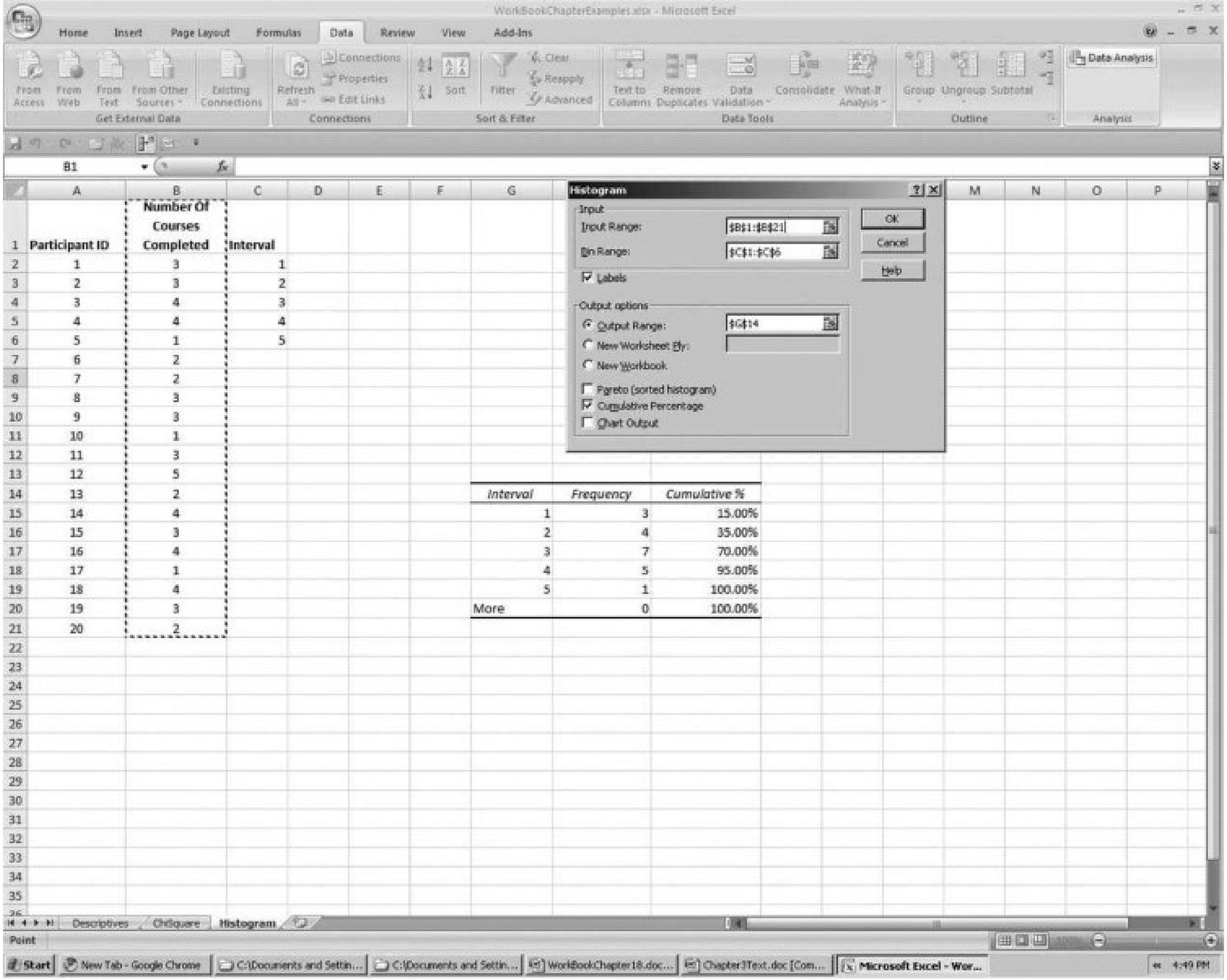
Press F1 for more help.

Frequency Distributions

The Excel Data Analysis Histogram procedure allows easy calculation of frequencies and cumulative percentages. Let's use the job training course completion data in the preceding example to illustrate the procedure. In the data sheet, as shown in Screen W18.6, click the Data Analysis command in the Analysis Group on the Data tab. Select Histogram in the Data Analysis window. Select the number of courses completed in the Input Range (i.e., $B$1:$B$21). Determine the intervals in which you want to calculate the frequencies. In our example, we want the frequencies of all possible numbers in the number of courses completed (i.e., 1, 2, 3, 4, and 5). We create a new variable (Interval) to hold these numbers and Bin in the Histogram Procedure is the place to define the intervals. Select an Output Range that does not overlap with the data. If you want your results in a new worksheet, select New Worksheet Ply instead of Output Range. Click the Label box. This allows you to show the title of your output. (*Important:* If you do not select the title cells, B1 and C1 in this case, in the Input Range and Bin Range, then you do not want to check the Label box.) Select Cumulative Percentage. Click OK. You should see the output as shown in Screen W18.6. Excel calculates the frequencies and cumulative percentages. Notice that it also creates in the "Interval" column of the frequency and cumulative percentage chart a "More" category for more than 5 courses completed.

If you want, you can also calculate a set of ranges for the number of courses completed, such as 0 to 3, 4 to 5, and above 5. Keep in mind that you need to enter the upper bound, not the lower bound, of an interval because Excel interprets the upper bound of a range as the lower bound of the next range. Screen W18.7 shows frequencies and cumulative percentages of such an example. The result shows that 70 percent (14 of 20) of participants completed three or fewer courses in the program.

Screen W18.6 Calculating Frequencies and Cumulative Percentages



	A	B	C
1	Participant ID	Number Of Courses Completed	Interval
2	1	3	1
3	2	3	2
4	3	4	3
5	4	4	4
6	5	1	5
7	6	2	
8	7	2	
9	8	3	
10	9	3	
11	10	1	
12	11	3	
13	12	5	
14	13	2	
15	14	4	
16	15	3	
17	16	4	
18	17	1	
19	18	4	
20	19	3	
21	20	2	

Histogram

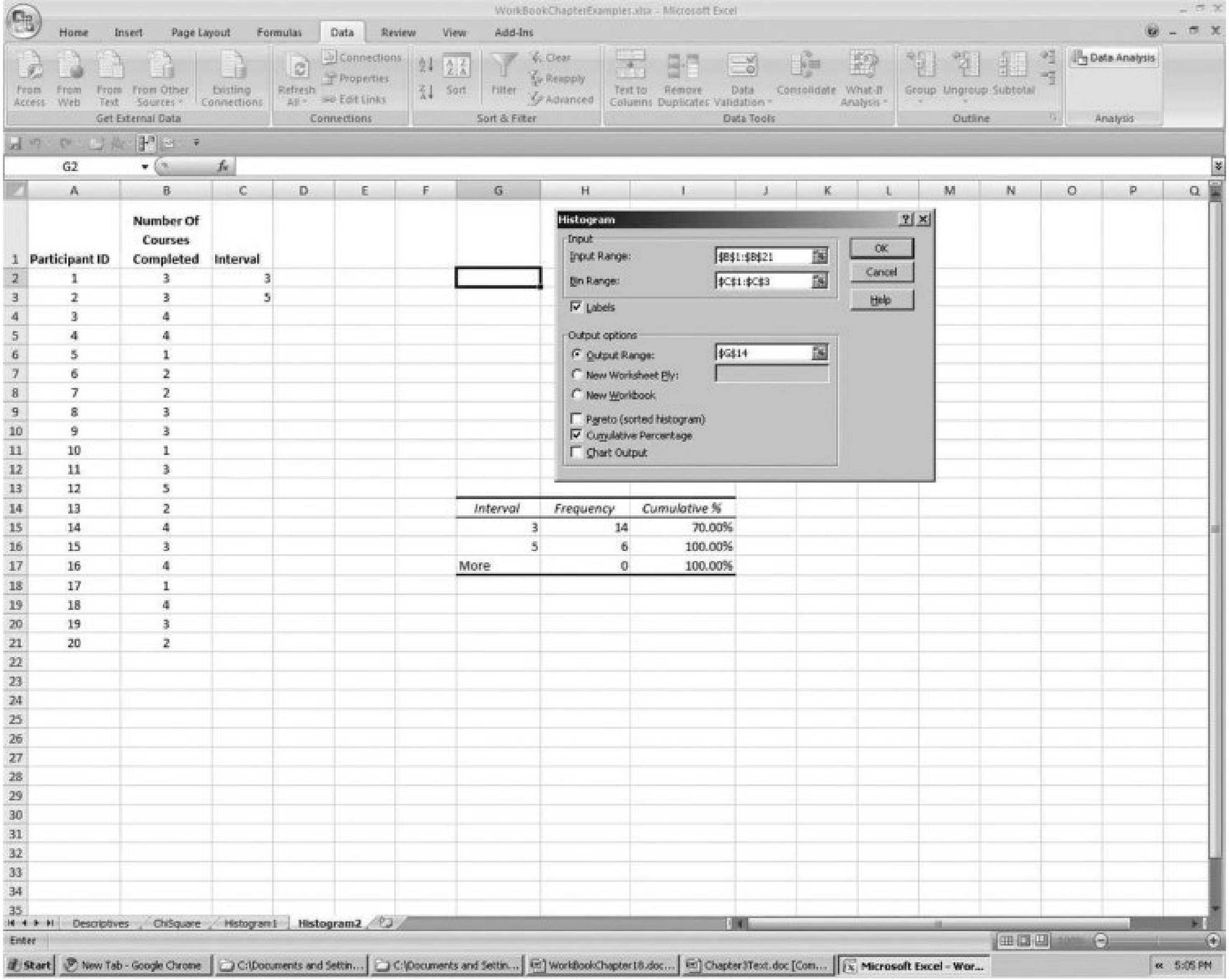
Input
Input Range: $B$1:$B$21
Bin Range: $C$1:$C$6
☑ Labels

Output options
● Output Range: $G$14
○ New Worksheet Ply:
○ New Workbook
□ Pareto (sorted histogram)
☑ Cumulative Percentage
□ Chart Output

OK
Cancel
Help

Interval	Frequency	Cumulative %
1	3	15.00%
2	4	35.00%
3	7	70.00%
4	5	95.00%
5	1	100.00%
More	0	100.00%

Screen W18.7 Calculating Frequencies and Cumulative Percentages of Ranges



	A	B	C
1	Participant ID	Number Of Courses Completed	Interval
2	1	3	3
3	2	3	5
4	3	4	
5	4	4	
6	5	1	
7	6	2	
8	7	2	
9	8	3	
10	9	3	
11	10	1	
12	11	3	
13	12	5	
14	13	2	
15	14	4	
16	15	3	
17	16	4	
18	17	1	
19	18	4	
20	19	3	
21	20	2	

Histogram

Input

Input Range: $B$1:$B$21

Bin Range: $C$1:$C$3

☑ Labels

Output options

● Output Range: $G$14

○ New Worksheet Ply:

○ New Workbook

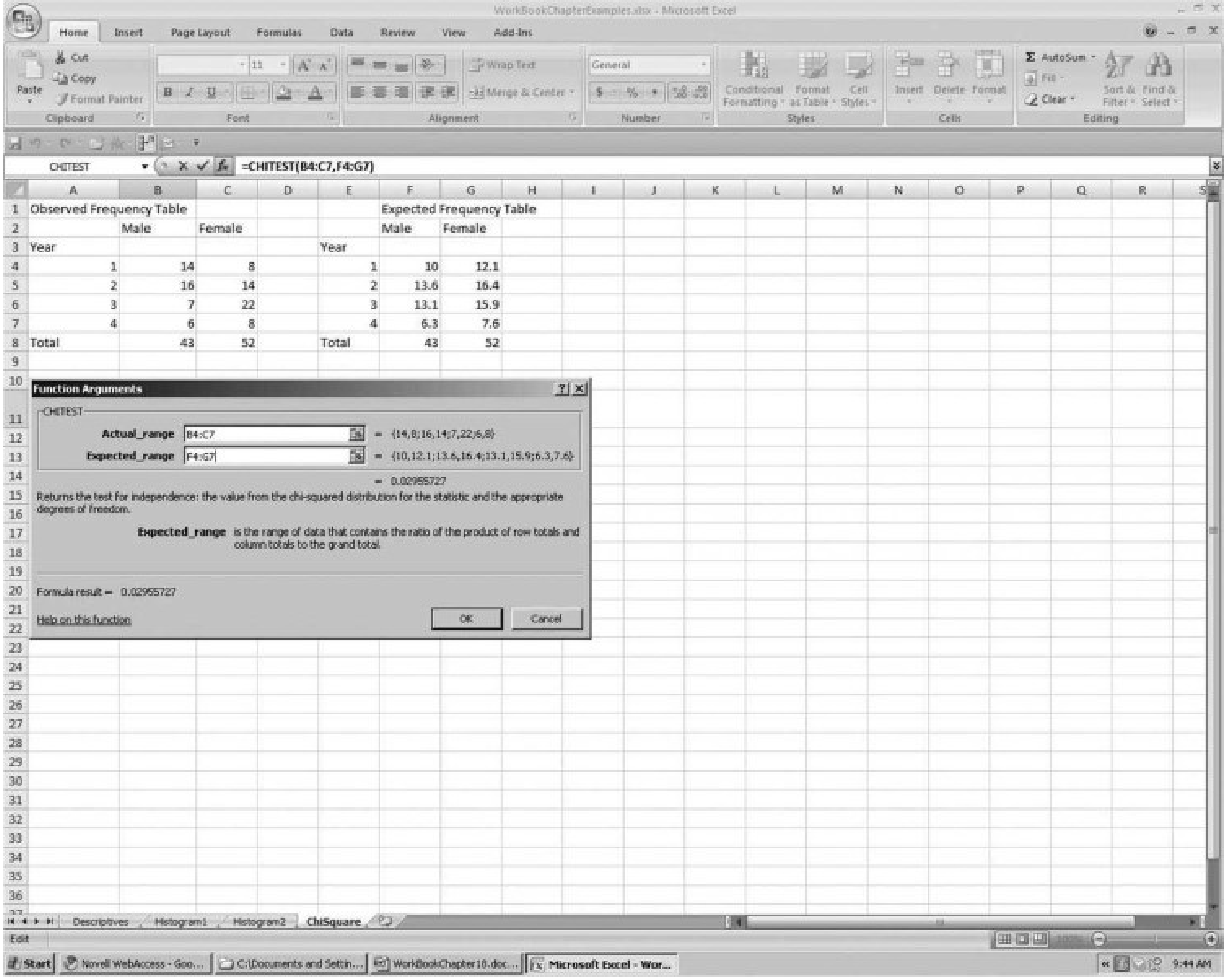
☐ Pareto (sorted histogram)

☑ Cumulative Percentage

☐ Chart Output

Interval	*Frequency*	*Cumulative %*
3	14	70.00%
5	6	100.00%
More	0	100.00%

Screen W18.8 Using CHITEST to Perform Chi-Square Test



CHITEST =CHITEST(B4:C7,F4:G7)

	A	B	C	D	E	F	G	H
1	Observed Frequency Table					Expected Frequency Table		
2		Male	Female			Male	Female	
3	Year				Year			
4	1	14	8		1	10	12.1	
5	2	16	14		2	13.6	16.4	
6	3	7	22		3	13.1	15.9	
7	4	6	8		4	6.3	7.6	
8	Total	43	52		Total	43	52	

Function Arguments

CHITEST

Actual_range: B4:C7 = {14,8;16,14;7,22;6,8}

Expected_range: F4:G7 = {10,12.1;13.6,16.4;13.1,15.9;6.3,7.6}

= 0.02955727

Returns the test for independence: the value from the chi-squared distribution for the statistic and the appropriate degrees of freedom.

Expected_range is the range of data that contains the ratio of the product of row totals and column totals to the grand total.

Formula result = 0.02955727

Help on this function

OK Cancel

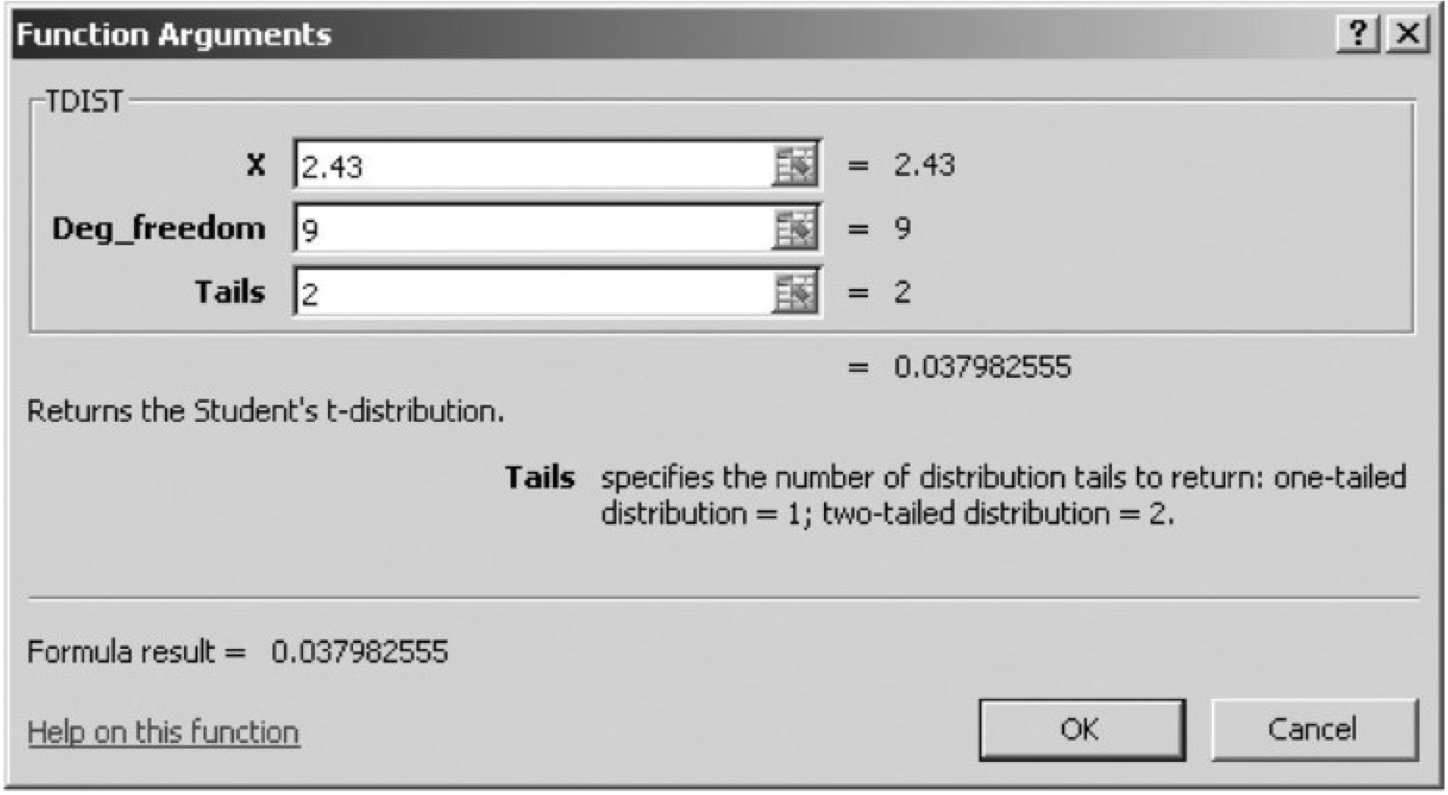
Chi-Square Test

The CHITEST procedure in Excel Insert Function (f_x) allows easy calculation of the probability associated with a critical value, facilitating the chi-square hypothesis testing process. To illustrate this procedure, let’s use the data on promotion by gender shown in Table 11.3 of the textbook. To use CHITEST, you need to obtain observed frequencies and expected frequencies as presented in Screen W18.8. In the CHITEST function arguments window, select observed frequencies in Actual_range (B4:C7 in our example) and expected frequencies in Expected_range (F4:G7 in our example). The formula result of 0.0296 is the probability associated with the critical value of the chi-square test. Because this value is smaller than 0.05, the result is statistically significant at the 5 percent level. Although the CHITEST procedure does not present the critical value, the associated probability provides the same information on the test result. It indicates that the relationship between promotion and gender exists.

One-Sample T-Test

Excel has procedures that allow you to perform one-sample and two-sample t-tests. The TDIST procedure in Insert Function can be used to calculate the probability associated with a critical t-value and can be used in a one-sample t-test. Screen W18.9 shows the procedure to obtain the probability using an example given in the discussion of one-sample testing in Chapter 12 in which the t-value is 2.43 with degrees of freedom of 9. In the TDIST procedure, the t-value needs to be placed in the X window (2.43, in this example). The result shows that the probability is 0.0380 with a two-tailed test. Excel does not take negative t-values, for some reason. However, because the t-distribution is symmetric, the probability value is the same for either a positive or a negative value. Thus you can use a positive value (say, 3.330) to replace the negative value of the same number (–3.330).

Screen W18.9 One-Sample T-Test



X 2.43 = 2.43

Deg_freedom 9 = 9

Tails 2 = 2

= 0.037982555

Returns the Student's t-distribution.

Tails specifies the number of distribution tails to return: one-tailed distribution = 1; two-tailed distribution = 2.

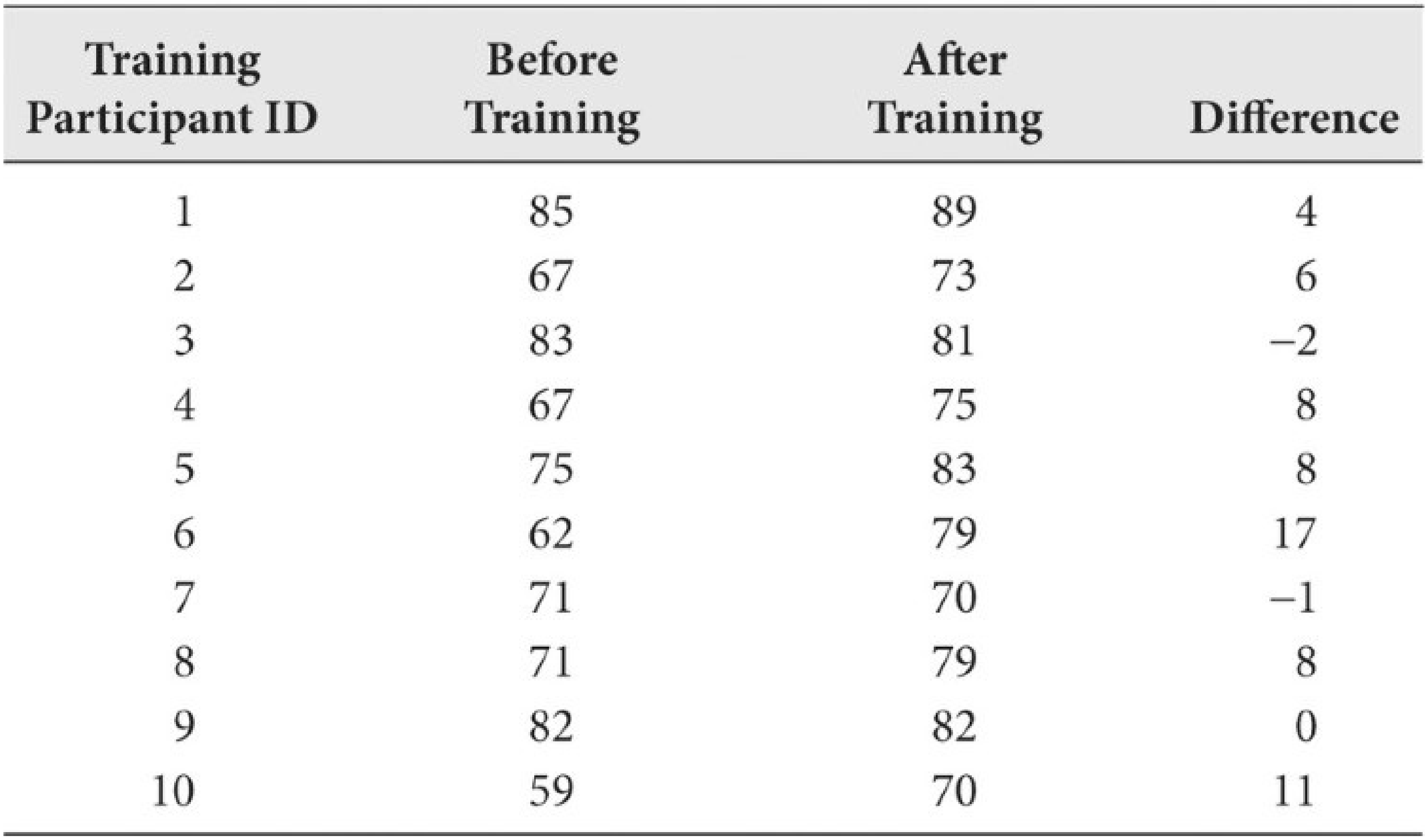
Formula result = 0.037982555

Two-Samples T-Test

Excel Insert Function TTEST procedure is designed to perform t-tests of two sample means. Table W18.2 shows the psychosocial performances of 10 school-age participants in a nonprofit program designed to improve at-risk participants' psychosocial behaviors. The training performance is measured by a standardized test score on psychosocial behaviors; 100 is the best possible score and 50 is the worst.

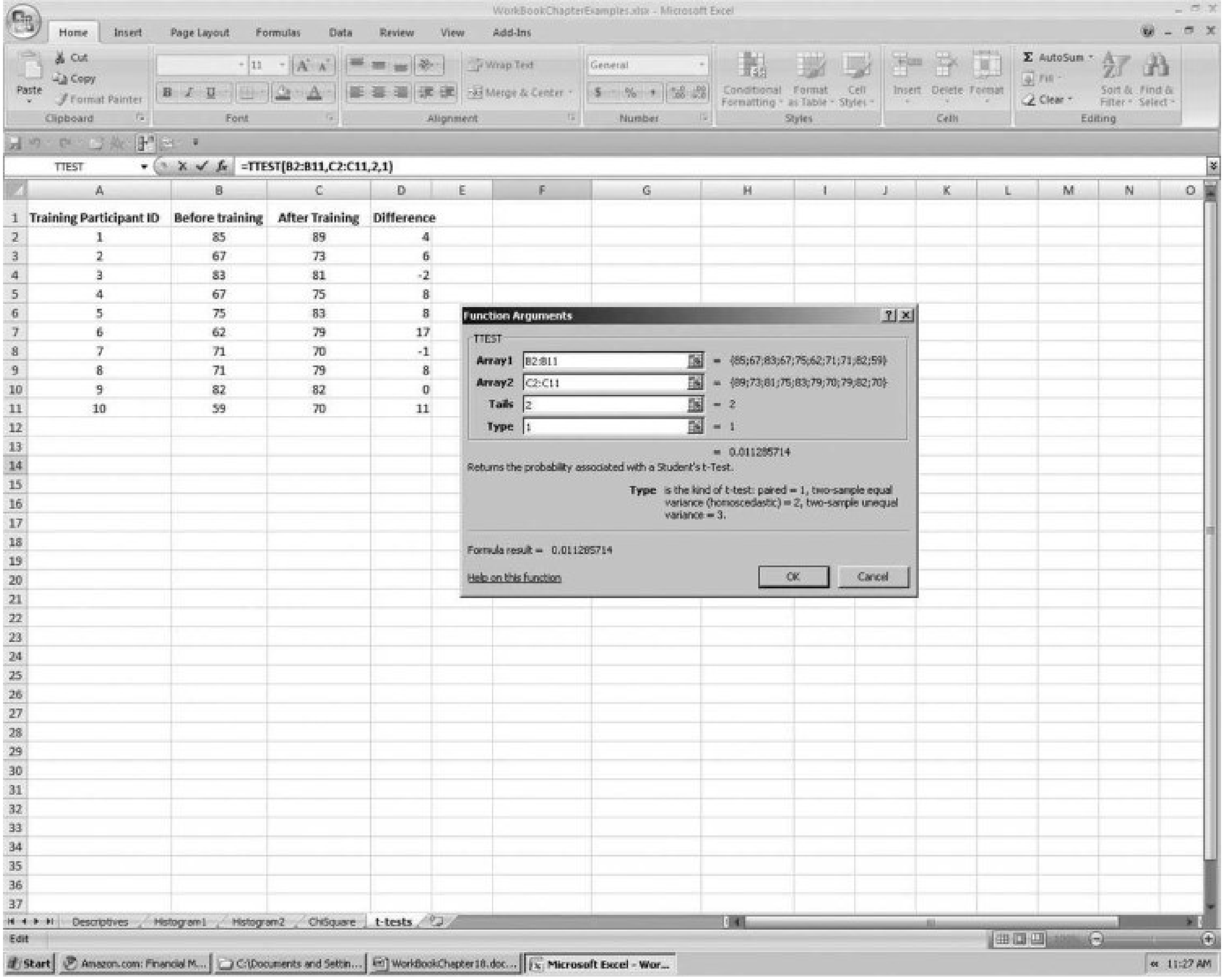
Because the tests were performed on the same subjects, the results constitute a paired sample. Screen W18.10 shows the TTEST procedure with this sample. In the TTEST function arguments window, include before-training scores in Array 1 and after-training scores in Array 2. Choose a two-tailed test if you don't know the direction of the test. Type "1" in the window that says "Type" for the paired-samples test. The result, probability or p = .0113, indicates that the difference between the average scores before and after the training is statistically significant at the 5 percent level.

Table W18.2 Training Performance for a Behavior Improvement Program



Training Participant ID	Before Training	After Training	Difference
1	85	89	4
2	67	73	6
3	83	81	–2
4	67	75	8
5	75	83	8
6	62	79	17
7	71	70	–1
8	71	79	8
9	82	82	0
10	59	70	11

Screen W18.10 TTEST for a Paired-Samples T-Test



TTEST =TTEST(B2:B11,C2:C11,2,1)

Training Participant ID	Before training	After Training	Difference
1	85	89	4
2	67	73	6
3	83	81	-2
4	67	75	8
5	75	83	8
6	62	79	17
7	71	70	-1
8	71	79	8
9	82	82	0
10	59	70	11

Function Arguments

TTEST

Array1 B2:B11 = {85;67;83;67;75;62;71;71;82;59}

Array2 C2:C11 = {89;73;81;75;83;79;70;79;82;70}

Tails 2 = 2

Type 1 = 1

= 0.011285714

Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result = 0.011285714

Help on this function

OK Cancel

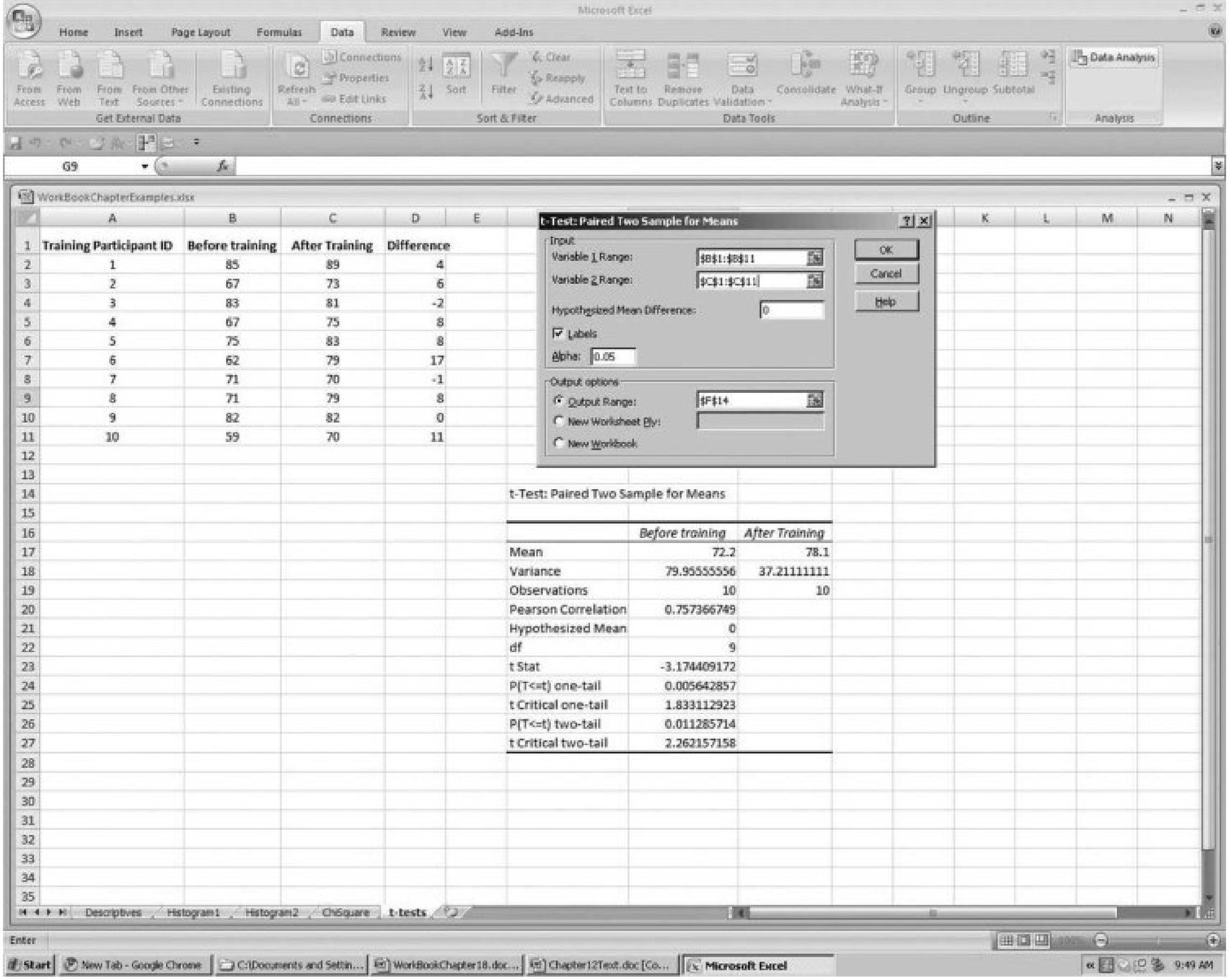
The Excel Data Analysis t-Test procedure can also be used in hypothesis testing of two samples. Choose Data Analysis from the Data tab. Choose t-Test: Paired Two Sample for Means in the Data Analysis window. As shown in Screen W18.11, use the Variable 1 Range and Variable 2 Range windows for before and after training scores, respectively. The Hypothesized Mean Difference is 0 (meaning that there is no difference between the mean scores before and after the training). Check the Labels box if you include the variable names in the data input ranges. We set the alpha, the significance level, at .05. The output of the Data Analysis t-Test procedure provides much richer information than the Insert Function TTEST procedure. In addition to the means and variances of the samples, the results include the t-value of the two samples, the critical t-values for one-tailed and two-tailed tests, and their associated p-values.

T-tests for two independent sample means can be performed similarly. In the Insert Function TTEST procedure, as shown in Screen W18.12, choose 2 in the "Type" window from the function arguments window if equal variance is assumed for the two samples (homogeneity of variance) or choose 3 if unequal variance is assumed (heterogeneity of variance). The Data Analysis t-Test procedure allows you to choose t-Test with Two-Sample Assuming Equal Variance or t-Test with Two-Sample Assuming Unequal Variance. The Excel procedure for these two tests is the same as that for the paired-samples test.

ANOVA

If you have more than two sample means and want to perform a one-way ANOVA, you can use the Excel Data Analysis ANOVA: Single Factor procedure. Consider an example. A city administration organizes a water-conservation campaign to educate residents on the need to conserve water. The campaign is conducted in three residential areas of the city. To determine the effectiveness of different educational instruments, the city provides flyers to the residents in area 1, flyers and a television ad in area 2, and only a television ad in area 3. Residents in the three areas are then tested to gauge their awareness of the need for water conservation. On the 10-point test used, a score of 10 corresponds to the highest level of awareness and a score of 0, to the lowest level. A sample from each area is drawn and the data are shown in Table W18.3.

Screen W18.11 Data Analysis t-Test for Paired Samples



	A	B	C	D
1	Training Participant ID	Before training	After Training	Difference
2	1	85	89	4
3	2	67	73	6
4	3	83	81	-2
5	4	67	75	8
6	5	75	83	8
7	6	62	79	17
8	7	71	70	-1
9	8	71	79	8
10	9	82	82	0
11	10	59	70	11

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $B$1:$B$11

Variable 2 Range: $C$1:$C$11

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

● Output Range: $F$14

○ New Worksheet Ply:

○ New Workbook

OK

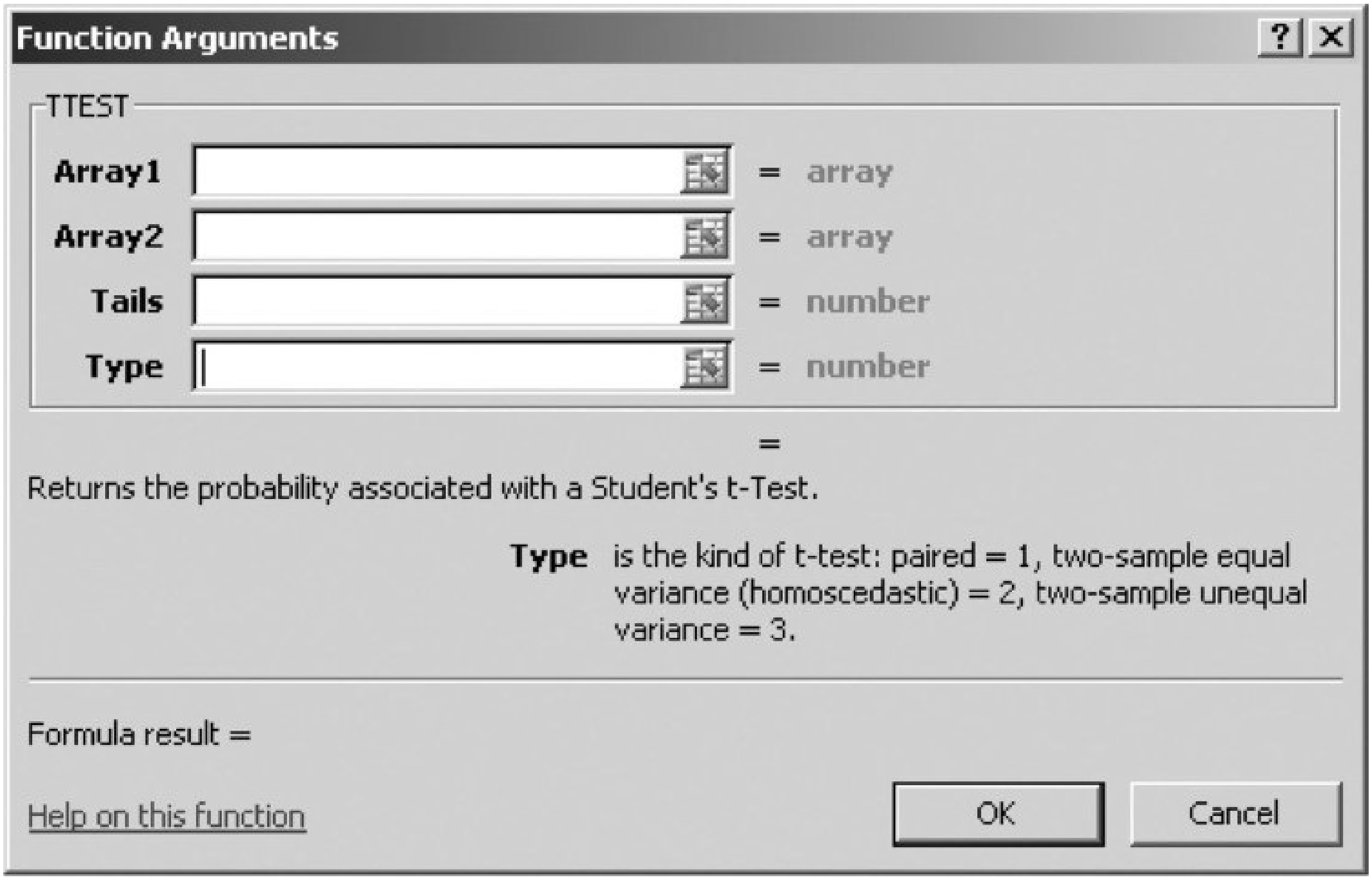
Cancel

Help

t-Test: Paired Two Sample for Means

	Before training	After Training
Mean	72.2	78.1
Variance	79.95555556	37.21111111
Observations	10	10
Pearson Correlation	0.757366749	
Hypothesized Mean	0	
df	9	
t Stat	-3.174409172	
P(T<=t) one-tail	0.005642857	
t Critical one-tail	1.833112923	
P(T<=t) two-tail	0.011285714	
t Critical two-tail	2.262157158	

Screen W18.12 TTEST Function Arguments Window



Returns the probability associated with a Student's t-Test.

Type is the kind of t-test: paired = 1, two-sample equal variance (homoscedastic) = 2, two-sample unequal variance = 3.

Formula result =

In the Excel Data Analysis Window, select ANOVA: Single Factor. Screen W18.13 shows the Excel procedure. It is important that you include all the data when selecting the Input Range as shown in Screen W18.13. The samples' descriptive statistics are presented in a Summary table and the results of the hypothesis test are included in an ANOVA table (see Screen W18.13). The null hypothesis of this test is that the three population means are the same and the alternate hypothesis is that they are different. The p-value (0.1298) of the F-test indicates that the difference is not statistically significant at the 5 percent level, which means that evidence is insufficient to indicate that the average awareness levels of the residents in the three areas are different.

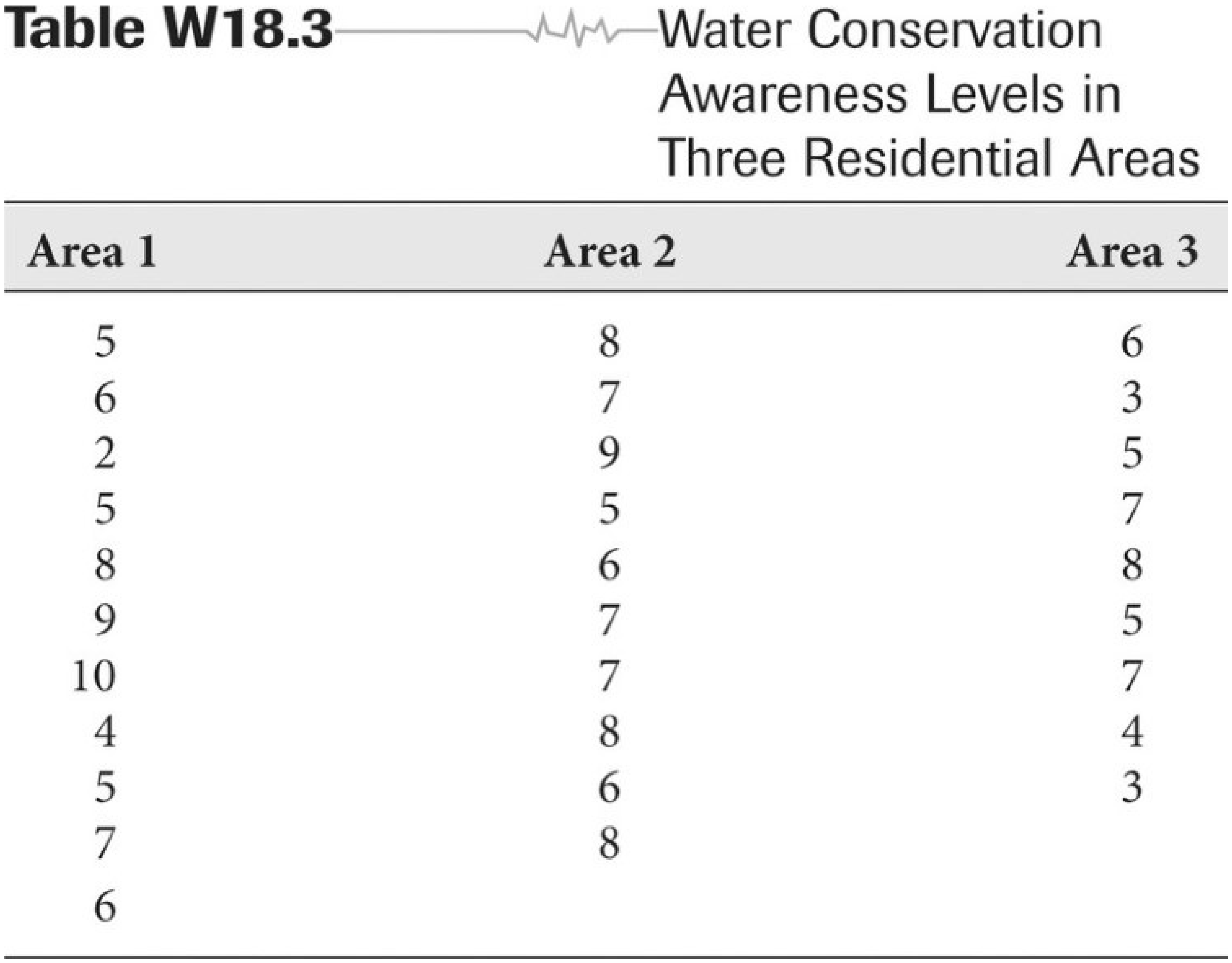
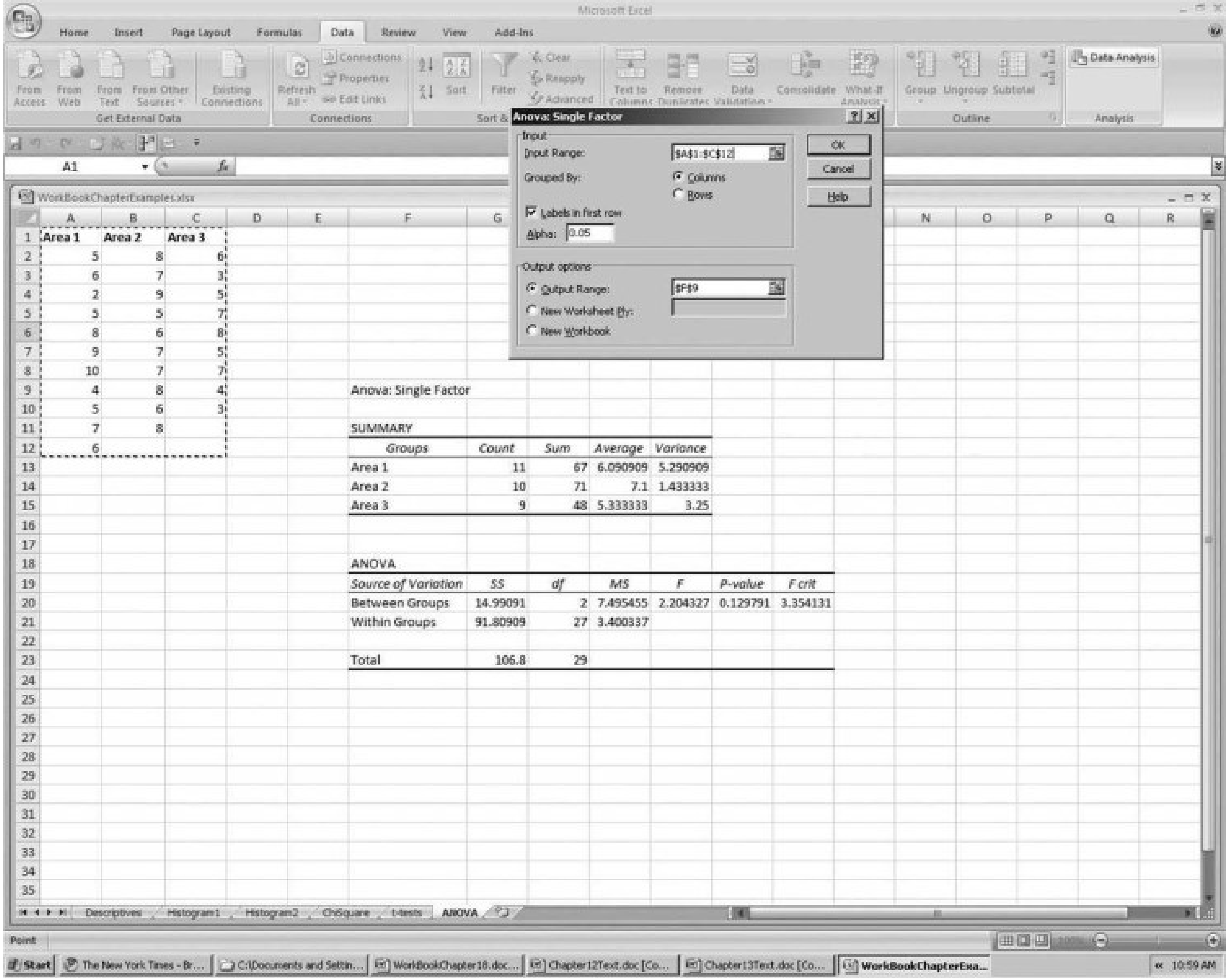


Table W18.3 Water Conservation Awareness Levels in Three Residential Areas

Area 1	Area 2	Area 3
5	8	6
6	7	3
2	9	5
5	5	7
8	6	8
9	7	5
10	7	7
4	8	4
5	6	3
7	8	
6		

Screen W18.13 Excel Data Analysis ANOVA: Single Factor



Anova: Single Factor

Input

Input Range: $A$1:$C$12

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Alpha: 0.05

Output options

● Output Range: $F$9

○ New Worksheet Ply:

○ New Workbook

OK Cancel Help

Area 1	Area 2	Area 3
5	8	6
6	7	3
2	9	5
5	5	7
8	6	8
9	7	5
10	7	7
4	8	4
5	6	3
7	8	
6		

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Area 1	11	67	6.090909	5.290909
Area 2	10	71	7.1	1.433333
Area 3	9	48	5.333333	3.25

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	14.99091	2	7.495455	2.204327	0.129791	3.354131
Within Groups	91.80909	27	3.400337			
Total	106.8	29				

Regression

The Excel Data Analysis Regression procedure provides easy access to regression analysis. Let's look at an example of multiple regression to illustrate the procedure. The procedure for simple regression is the same. Environmental managers have long used environmental spending level, measured by environmental spending as a percentage of total government spending, to gauge a government's efforts in environmental protection and development. Environmental scientists speculate that this spending level is affected by population growth and economic activities. A regional planning agency has collected data on environmental spending, population densities, and manufacturing employment growth for 12 cities in the region (Table W18.4).

To use the Data Analysis Regression procedure, select the Data Analysis command in the Analysis group on the Data tab, and click Regression from the Data Analysis window. As shown in Screen W18.14, the dependent variable *Y* is the environmental spending level. The independent variable *X* range is the population density and manufacturing employment growth rate. Check the Labels box if you include variable names in your data ranges. Select an output range that does not overlap with the data.

The Excel Summary Output has three tables, also shown in Screen W18.14. The first table (Summary Output) provides R-square, adjusted R-square, and the standard error of the estimate. It also presents "multiple R," which is the correlation coefficient between the actual values and the predicted values of the dependent variable *Y*. The second table (ANOVA) presents the F-test result of the regression model. It shows that the model is statistically significant at the 10 percent level but not at the 5 percent level. The last table gives coefficients (unstandardized) of the independent variables and the constant (known as Intercept in Excel printout). In our example, the regression model is as follows: Environmental Spending Level = 0.0818 + 0.000235*Population Density + 0.0009175*Manufacturing Employment Growth Rate. The t-test results of slopes are also presented. The results show a possible positive impact of population density on environmental spending and that the relationship is statistically significant at the 5 percent level (p = .0452). Nevertheless, there is no evidence of an impact of manufacturing growth on environmental spending (p = .2833). The output also presents 95 percent confidence intervals (as the default) for the estimates. You can specify the level of confidence interval estimates (for example, 99 percent).

Excel Data Analysis Regression also allows you to plot residuals against predicted values of the dependent variable as well as values of each independent variable in your effort to check the assumption of homoscedasticity. Screen W18.15 shows the procedure to obtain the plot. The plot of residuals by the predicted values of the dependent variable in our environmental spending example is shown in Figure W18.1.

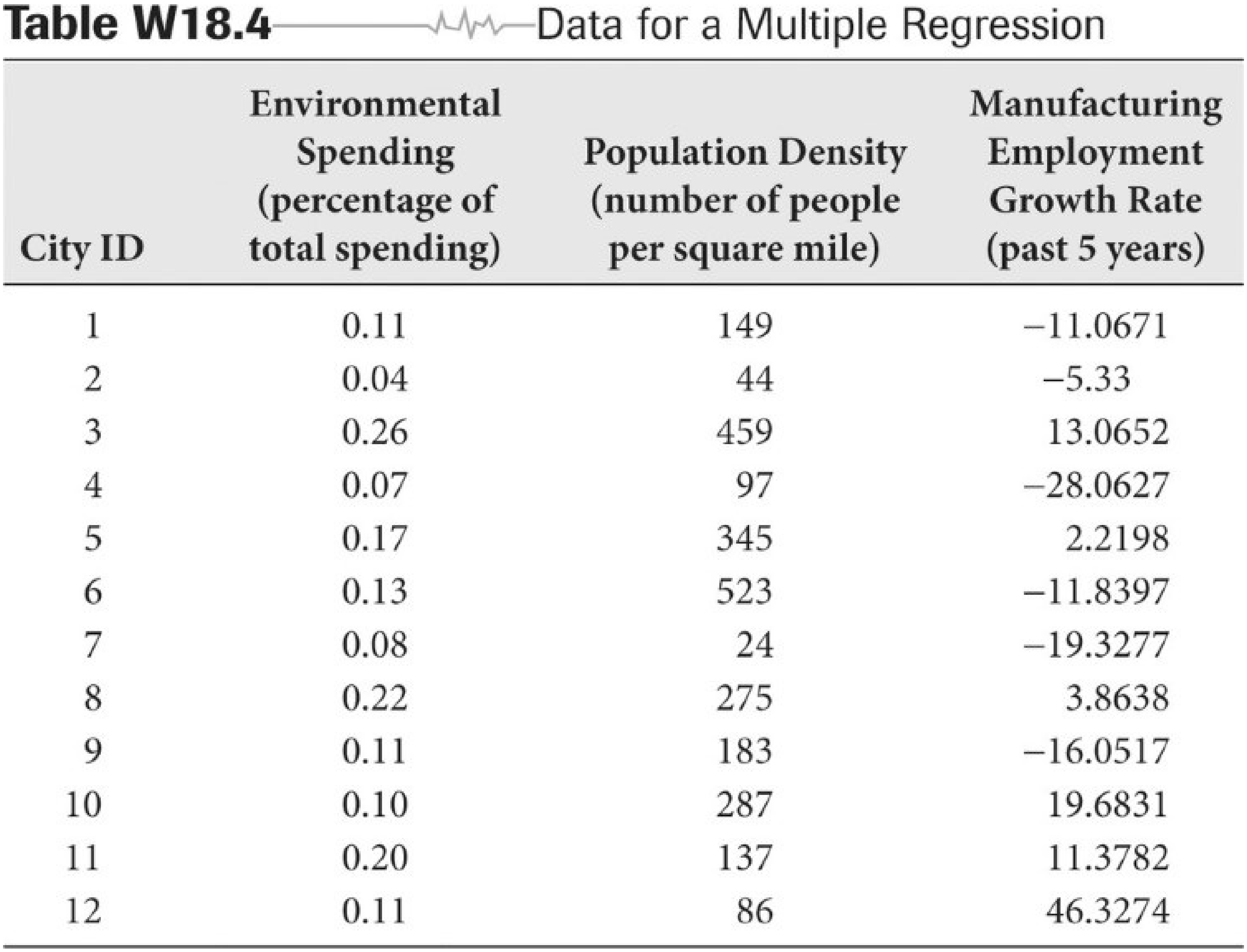
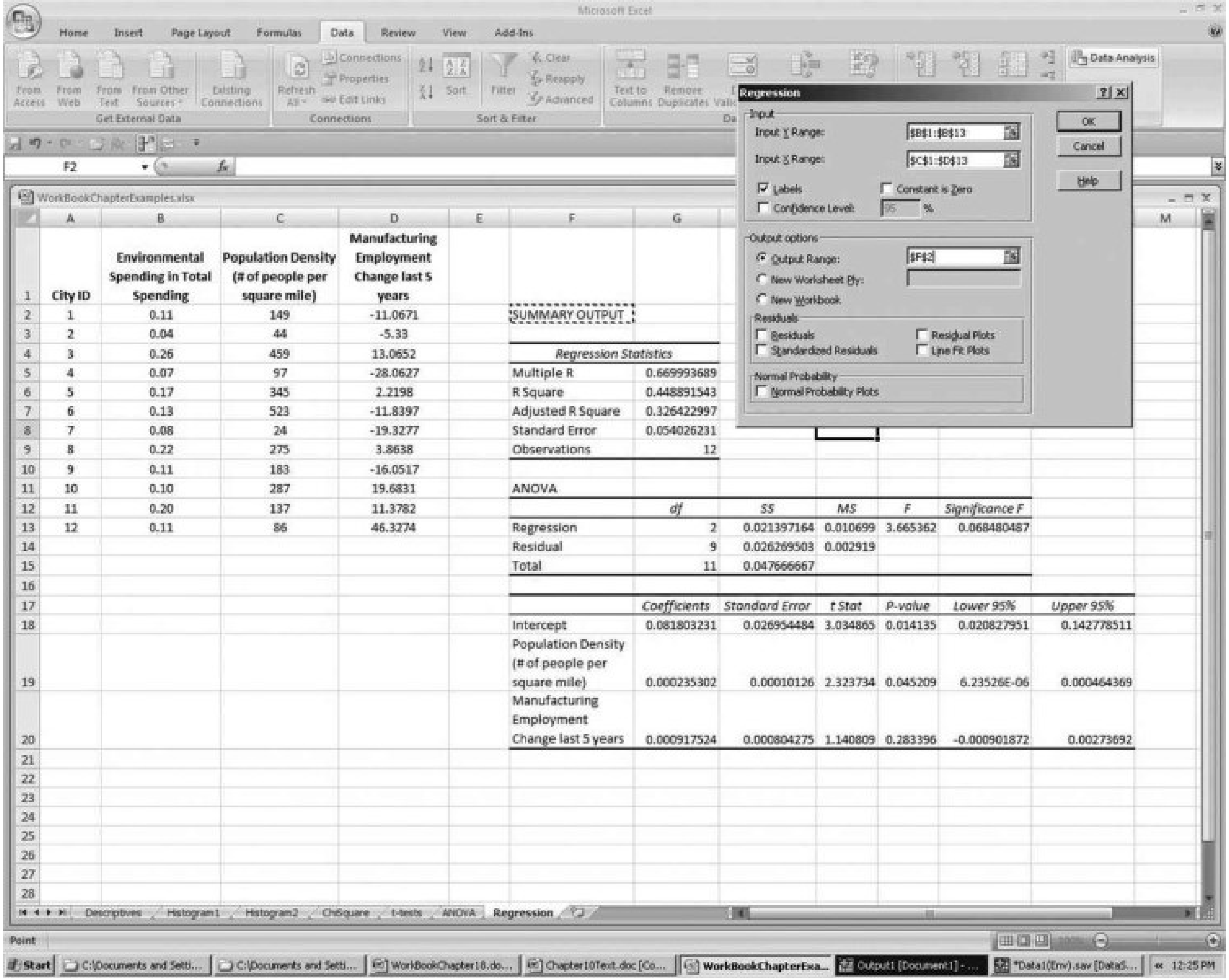


Table W18.4 Data for a Multiple Regression

City ID	Environmental Spending (percentage of total spending)	Population Density (number of people per square mile)	Manufacturing Employment Growth Rate (past 5 years)
1	0.11	149	−11.0671
2	0.04	44	−5.33
3	0.26	459	13.0652
4	0.07	97	−28.0627
5	0.17	345	2.2198
6	0.13	523	−11.8397
7	0.08	24	−19.3277
8	0.22	275	3.8638
9	0.11	183	−16.0517
10	0.10	287	19.6831
11	0.20	137	11.3782
12	0.11	86	46.3274

Screen W18.14 Excel Data Analysis Regression



	City ID	Environmental Spending in Total Spending	Population Density (# of people per square mile)	Manufacturing Employment Change last 5 years
2	1	0.11	149	-11.0671
3	2	0.04	44	-5.33
4	3	0.26	459	13.0652
5	4	0.07	97	-28.0627
6	5	0.17	345	2.2198
7	6	0.13	523	-11.8397
8	7	0.08	24	-19.3277
9	8	0.22	275	3.8638
10	9	0.11	183	-16.0517
11	10	0.10	287	19.6831
12	11	0.20	137	11.3782
13	12	0.11	86	46.3274

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.669993689
R Square	0.448891543
Adjusted R Square	0.326422997
Standard Error	0.054026231
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.021397164	0.010699	3.665362	0.068480487
Residual	9	0.026269503	0.002919		
Total	11	0.047666667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.081803231	0.026954484	3.034865	0.014135	0.020827951	0.142778511
Population Density (# of people per square mile)	0.000235302	0.00010126	2.323734	0.045209	6.23526E-06	0.000464369
Manufacturing Employment Change last 5 years	0.000917524	0.000804275	1.140809	0.283396	-0.000901872	0.00273692

Regression

Input

Input Y Range: $B$1:$B$13

Input X Range: $C$1:$D$13

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $F$2

○ New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

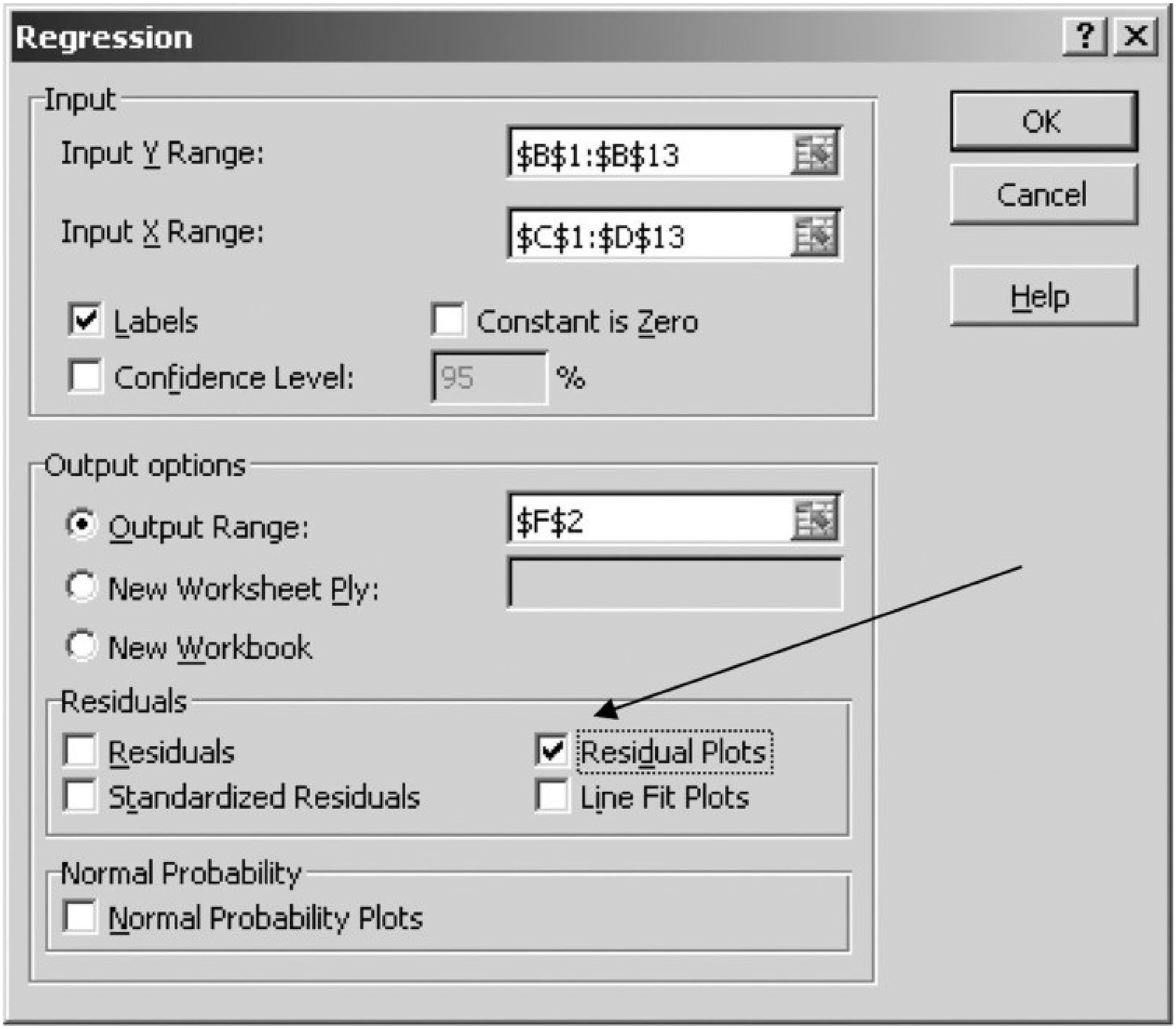
☐ Normal Probability Plots

OK

Cancel

Help

Screen W18.15 The Excel Procedure for Residual Plots in Regression



Regression

Input

Input Y Range: $B$1:$B$13

Input X Range: $C$1:$D$13

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $F$2

○ New Worksheet Ply:

○ New Workbook

Residuals

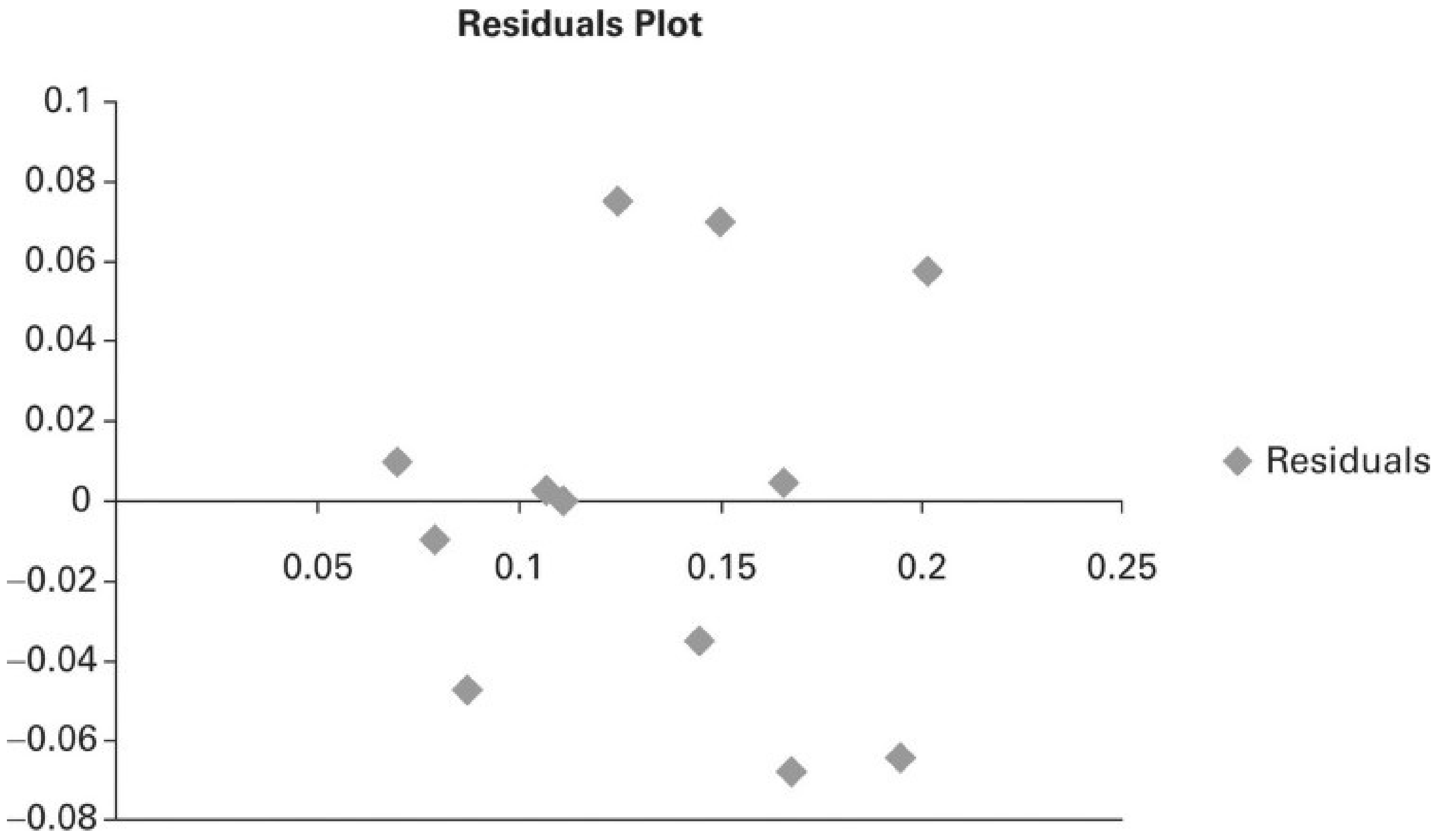
☐ Residuals ☑ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

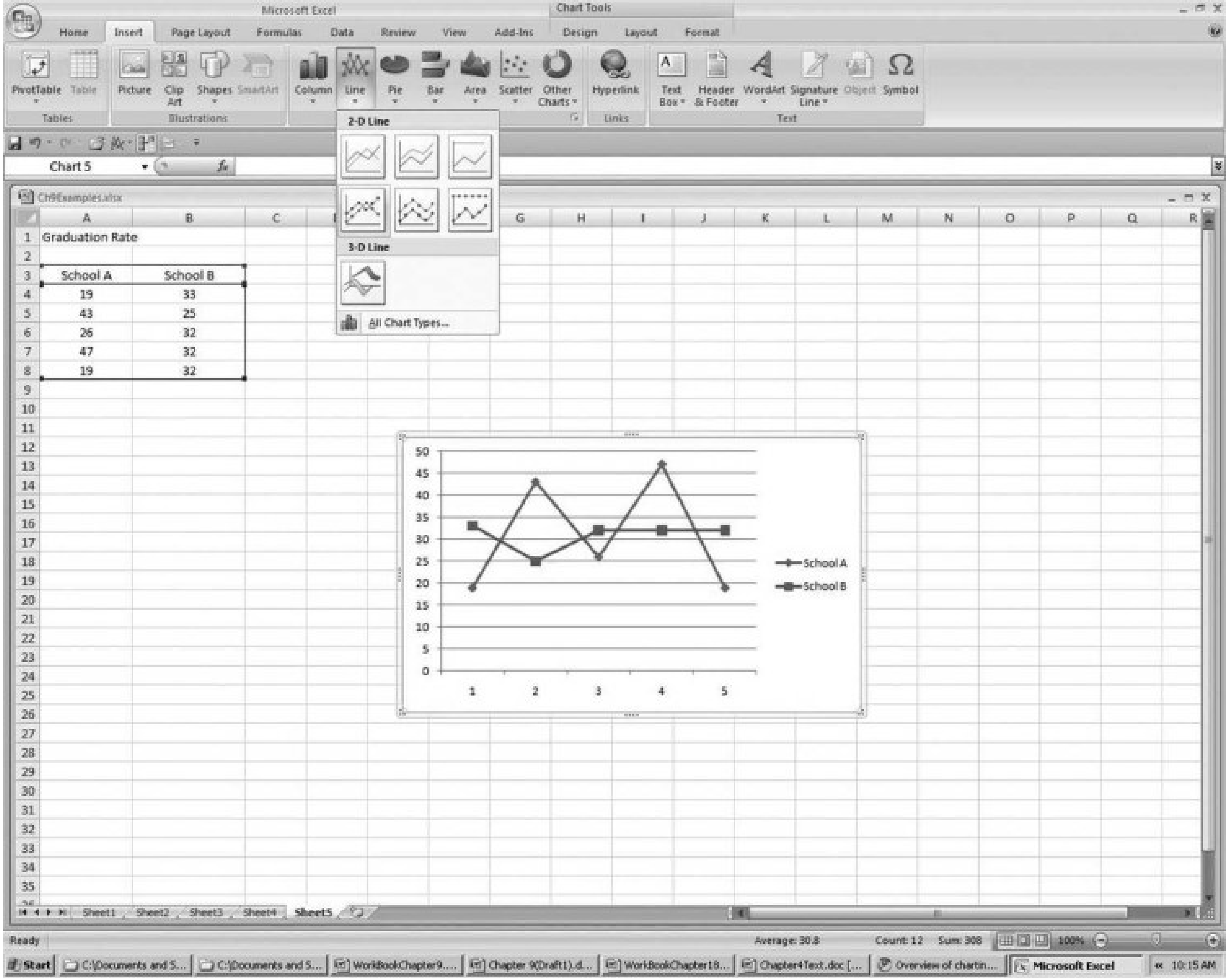
Figure W18.1 The Plot of Residuals by Predicted Values


Residuals Plot
0.1
0.08
0.06
0.04
0.02
0
–0.02
–0.04
–0.06
–0.08
0.05
0.1
0.15
0.2
0.25
◆ Residuals

Creating Charts in Excel

Finally, we would like to introduce the Excel Chart function. It is very easy to create charts with Excel. You can do it by clicking the chart type in the Charts group on the Insert Tab of an Excel sheet. To create a professional-looking chart that displays the details that you want, you can modify the chart, apply predefined styles and layouts, and add various formatting features. For example, to create a line chart similar to the one in Figure 9.4 of the textbook, select the data for the presentation first, click the Line command in the Charts group on the Insert tab. (*Note:* If you use Excel 2003, click the Chart button on the Insert tab and you will be shown a list of chart types in the Chart Wizard window.) Select the Line with Markers on the Line tab. You should see a graph similar to that shown in Screen W18.16. Use the Design menu and the Layout menu to edit the graph.

Screen W18.16 Creating an Excel Line Chart



Graduation Rate

School A	School B
19	33
43	25
26	32
47	32
19	32

Average: 30.8 Count: 12 Sum: 308

Chapter 19 SPSS User’s Guide

This chapter guides you through the operating characteristics of the Statistical Package for the Social Sciences (SPSS), now sold by IBM as “IBM SPSS.” This popular software is an invaluable tool for social scientists because of its capacity for handling large datasets and performing a wide range of statistical tests. Across many social science professions, including public administration and public policy analysis, SPSS software offers professionals, practitioners, and students alike the ability to perform myriad statistical procedures and data evaluation processes. Even better, it handles these analyses and operations without requiring users to perform intricate mathematical operations. Luckily, SPSS is also extremely easy to use. For those already familiar with Excel, SPSS looks and feels quite similar. If you don’t have previous experience with spreadsheets, there is no need for apprehension, since SPSS is truly user-friendly. Through the initial familiarization exercises, you will quickly gain an appreciation of just how simple it is to use this software as a data analysis tool.

Of course, understanding the statistical concepts that underlie the tasks SPSS will perform is crucial. By analogy, if SPSS is viewed as a vehicle, then you are the driver and need to know the rules of the road. The textbook *Essential Statistics for Public Managers and Policy Analysts* is about the rules and practice of using statistics; studying it will help you develop a sense of what you need to look out for—such as data characteristics, causality patterns in relationships, and the assumptions and limitations of various statistical procedures. This chapter helps users get started with SPSS and shows them how to apply concepts discussed in the textbook. Enough said. Now, let’s get started!

Note: Software versions change quickly these days, and perhaps some of the screenshots will be slightly different from your version. Still, we think the following will be helpful. As needed, you can always use the “Help → Tutorial” that comes with your version for further answers and guidance, too. SPSS is available on most (if not nearly all) university campuses these days, and some universities make this software available for download to their students. At this time of writing, student versions can also be leased for 6 to 24 months (search online for “buy SPSS Statistics GradPack”).

SPSS Screens

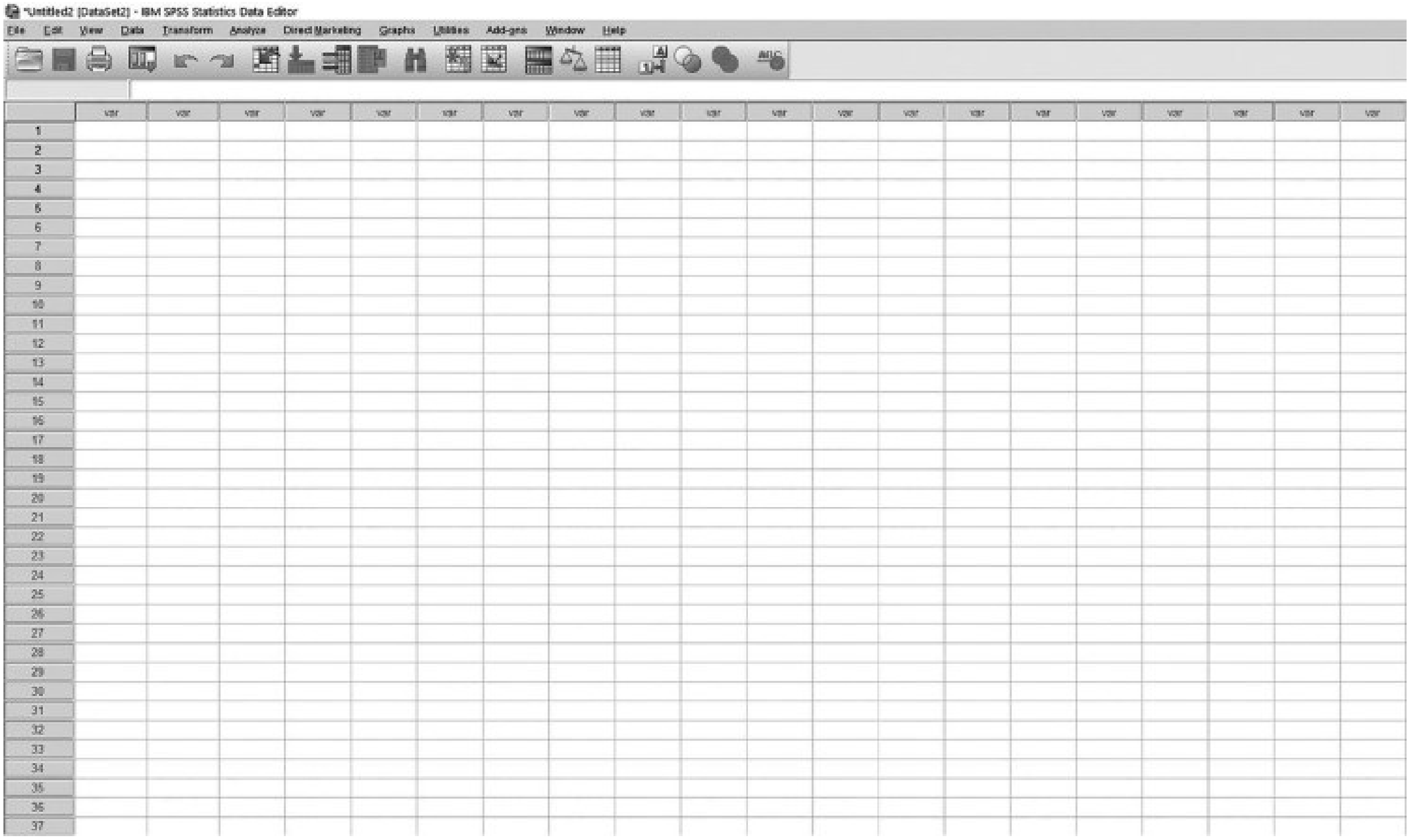
After you open SPSS, the Data Editor screen will appear as shown in Screen W19.1. This screen will display the data, values, and labels for each dataset variable. Presently, no data are displayed because you have not yet created or opened a file in SPSS.

At the top of the screen is the toolbar from which SPSS commands are selected. These commands are the principal way of getting your data in shape for subsequent analysis. Note the Help function, which contains a tutorial to introduce you to the capabilities of SPSS. Investigating this tool is certainly worth the time investment and is highly recommended.

Take time to explore each of the toolbar menus and familiarize yourself with the options and functions of each. Don’t worry if the terminology seems unfamiliar at first. As you progress through this guide, you will work your way through it.

At the bottom left of the SPSS Data Editor screen are two tabs, called Data View and Variable View. Select the Variable View tab. The screen displayed contains no information beyond the fields comprising that screen because no file has been opened yet. This screen will be explained and demonstrated when you create your first variable.

Screen W19.1 Data Editor Screen



To return to the Data View screen, merely select the tab at the bottom of the screen. SPSS also has an Output screen that will become familiar to you after you complete your first analysis.

Creating a Variable

Chapter 6 of the textbook uses sample data in the discussion of the mean, median, and mode. We will use those data to create your first variable and then see how easy it is for SPSS to perform statistical calculations.

Note: Throughout this guide, the symbol → means "execute the next command immediately following the arrow." For example, Select Variable View → Name means "open the Variable View screen and then select Name by placing the cursor over the Name option and left clicking." Practice this command now.

To create the first variable, we will give it a name and a label. Go to the Variable View screen, and place the cursor in the first row of the cells in the Name column. In the first cell, type "variablx" as shown in Screen W19.2. We could have used a longer name than variablx, but variable names in SPSS have been limited historically to eight characters and many analysts still follow this practice. Variable names must start with a letter and may contain only letters and numbers, as well as periods when not used as the first or last character; thus, "variable x" is not a valid variable name (it contains an empty space), but "variablx" and "var.x" are. Don't worry about remembering what each variable name means; the purpose of the Label column is to fully define the variable name. We will show you how to create labels, too. (Once the label is entered, SPSS will use the label, rather than the variable name, in subsequent output tables.)

Entering the variable name automatically produces some additional settings, as shown in Screen W19.2. The next column allows selection of the type of variable. The default setting for Type is a numeric variable; your variable is numeric since you will be entering numbers. To examine other options, highlight the cell in the first row of the Type column. This will bring up an icon in the right-hand sector of the data cell. When you click on that icon, you will see a pop-up menu, a typical feature of SPSS, from which you select the type of variable. In the case of variablx, select Numeric. This pop-up dialogue box is shown in Screen W19.3. (On newer versions of SPSS, the dialogue box may include the information that "The numeric type honors the digit grouping setting." This simply means that any numbers such as 1, 2, 3, and so on could also be used to refer to, say, groups that are, in fact, nominal (e.g., groups of subjects, the first group, second group). This is just an advisory so that you don't forget or later assume that every numerically coded variable is in fact interval or ratio.

Continuing across, the first cell in the Width column identifies the maximum number of characters of this variable; the default is eight. The Decimals column indicates the number of decimals; the default is two. Hence, the default settings accommodate any value of variablx up to 99999.99. The Label column provides user-defined fields to further describe each variable. Type in the label "Data for calculating mean, median and mode." (Double clicking on the Label data box allows label entry and editing of entries.) In later sections, you will learn more about the Values and Missing columns. The latter column is to the right of the screen, found by using the horizontal scroll bar at the bottom of the screen. The default for both is "None." The column labeled

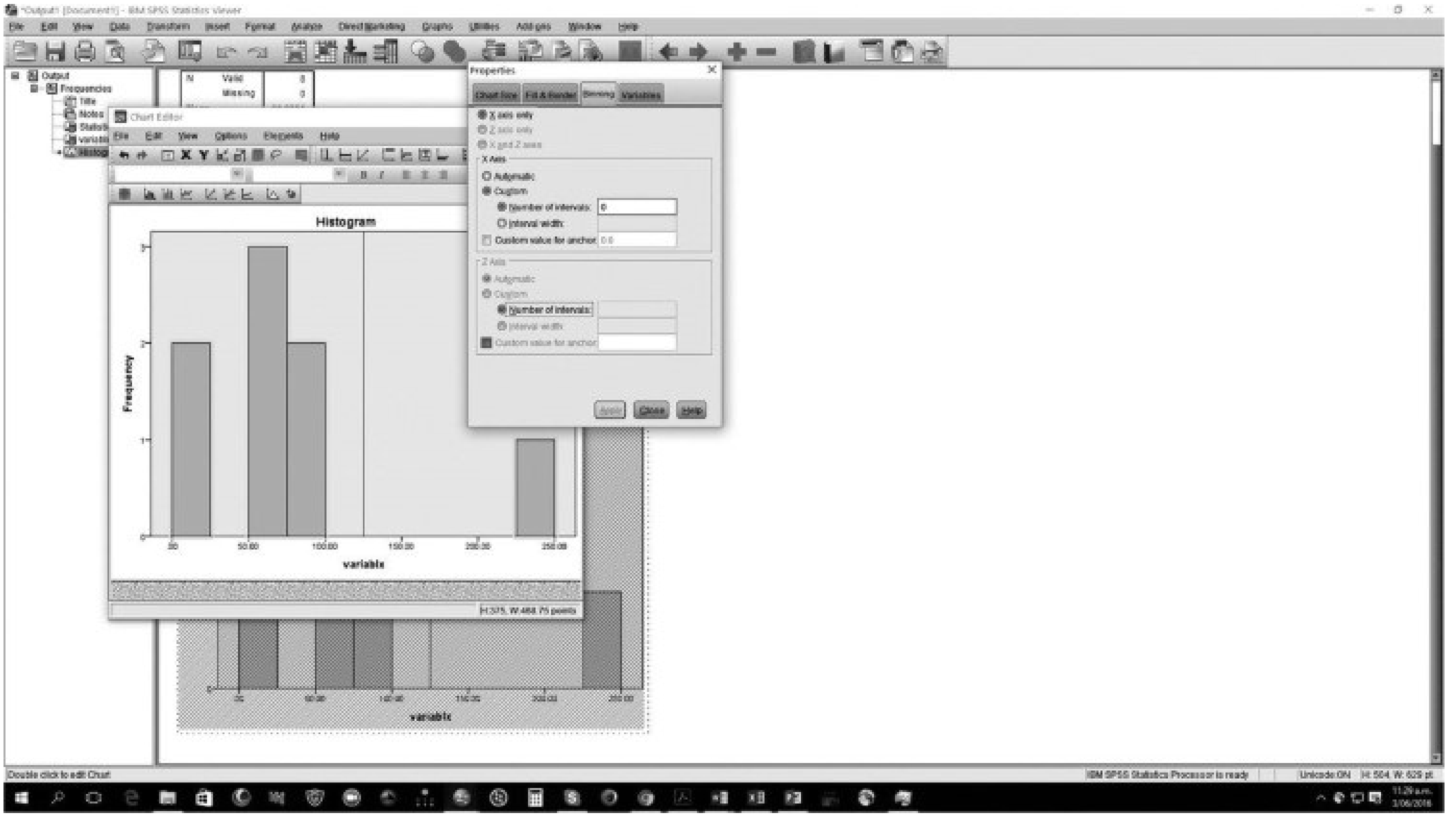
“Columns” adjusts the widths of the columns in the Data View screen. The widths of the columns in the Variable View screen can also be adjusted by placing the cursor between two columns until you see the icon that is used to adjust the widths of the columns, which looks like this:

⇹

On the right-hand side of the Variable View screen, you will see two additional columns. The Align column aligns the data in the Data View screen to the right, left, or center (this is somewhat analogous to the alignment of text in word-processing programs). The rightmost column is labeled “Measure.” Variables can be labeled as being nominal, ordinal, or scale (that is, continuous). This label is of no consequence in using most SPSS features later, nor does the label appear on most SPSS outputs. Rather, the label shows up in dialogue boxes from which variables are selected, such as in Screen W19.5. This label may serve as a reminder to analysts of the measurement level of their variables. Here, we define variablx as a scale—the default label, which appears with a ruler icon to the left of the variable name in Screen W19.5. Many analysts ignore the option of labeling their measurement levels. Ordinal and nominal labels show up with other icons, namely as bar charts and Venn diagrams (concentric circles). Again, these are just user-assigned labels that have no effect on our ability to use any statistical procedure later.

Return now to the Data View screen, where variablx is displayed. Place the cursor on variablx, and view the drop-down label defining the variable. The next step is to enter values for variablx. Let’s enter the data from Chapter 6 in the textbook (i.e., 20, 20, 67, 70, 71, 80, 90, and 225) by selecting a data cell, entering a value, and then pressing Enter. (The data are also recorded by simply selecting the keyboard down arrow after entering the numeric data value in each cell.) In this manner, you should reproduce Screen W19.4. You have just completed the data entry for your first variable. Congratulations!

Screen W19.2 Naming Variables



Properties

Chart Size | Fill & Border | Binning | Variables

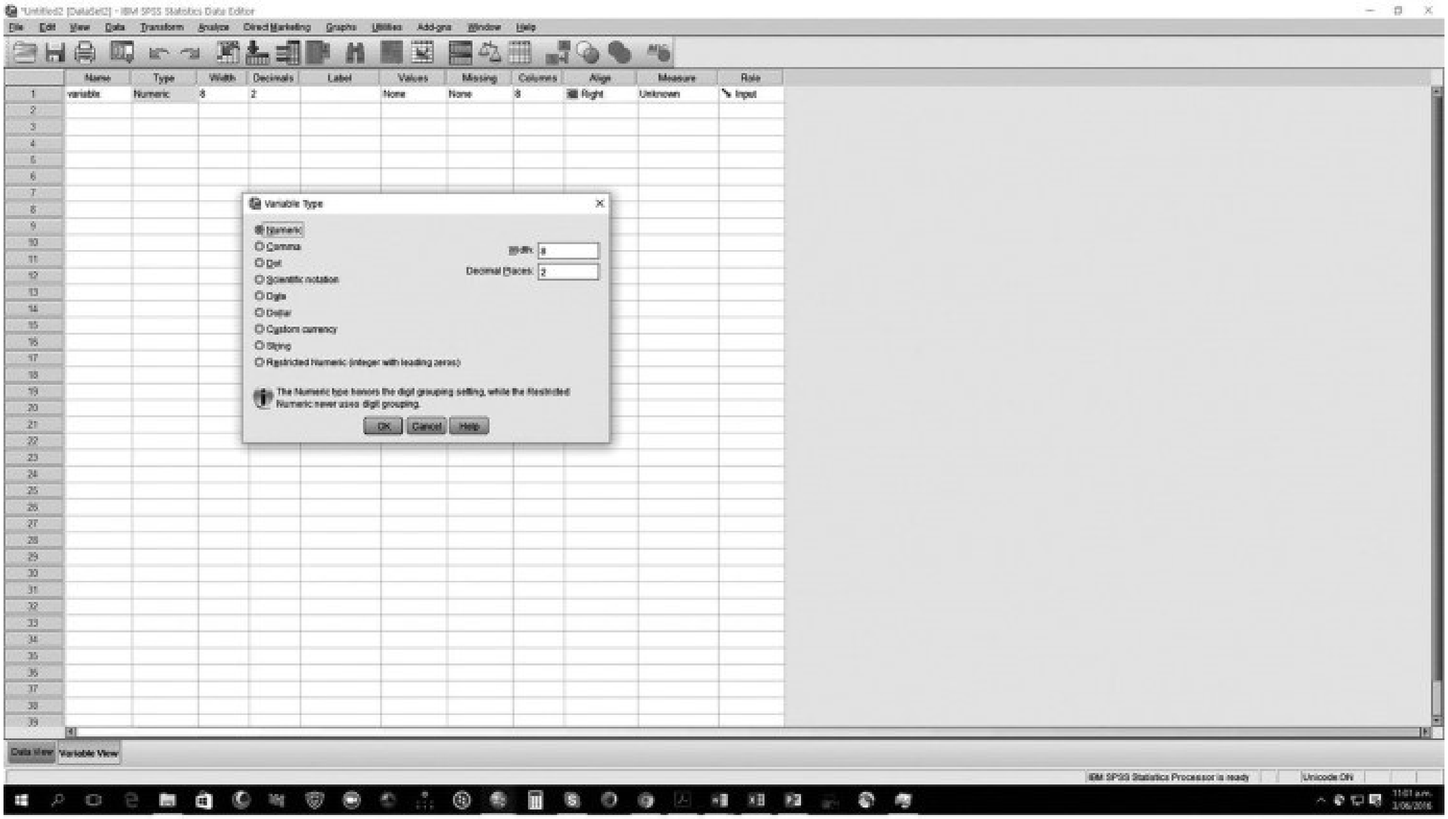
● X axis only
○ Z axis only
○ X and Z axes

X Axis
○ Automatic
● Custom
● Number of intervals: 6
○ Interval width:
□ Custom value for anchor: 0.0

Z Axis
● Automatic
○ Custom
● Number of intervals:
○ Interval width:
■ Custom value for anchor:

Apply | Close | Help

Screen W19.3 Selecting Variable Type



*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2		None	None	8	Right	Unknown	Input

Variable Type

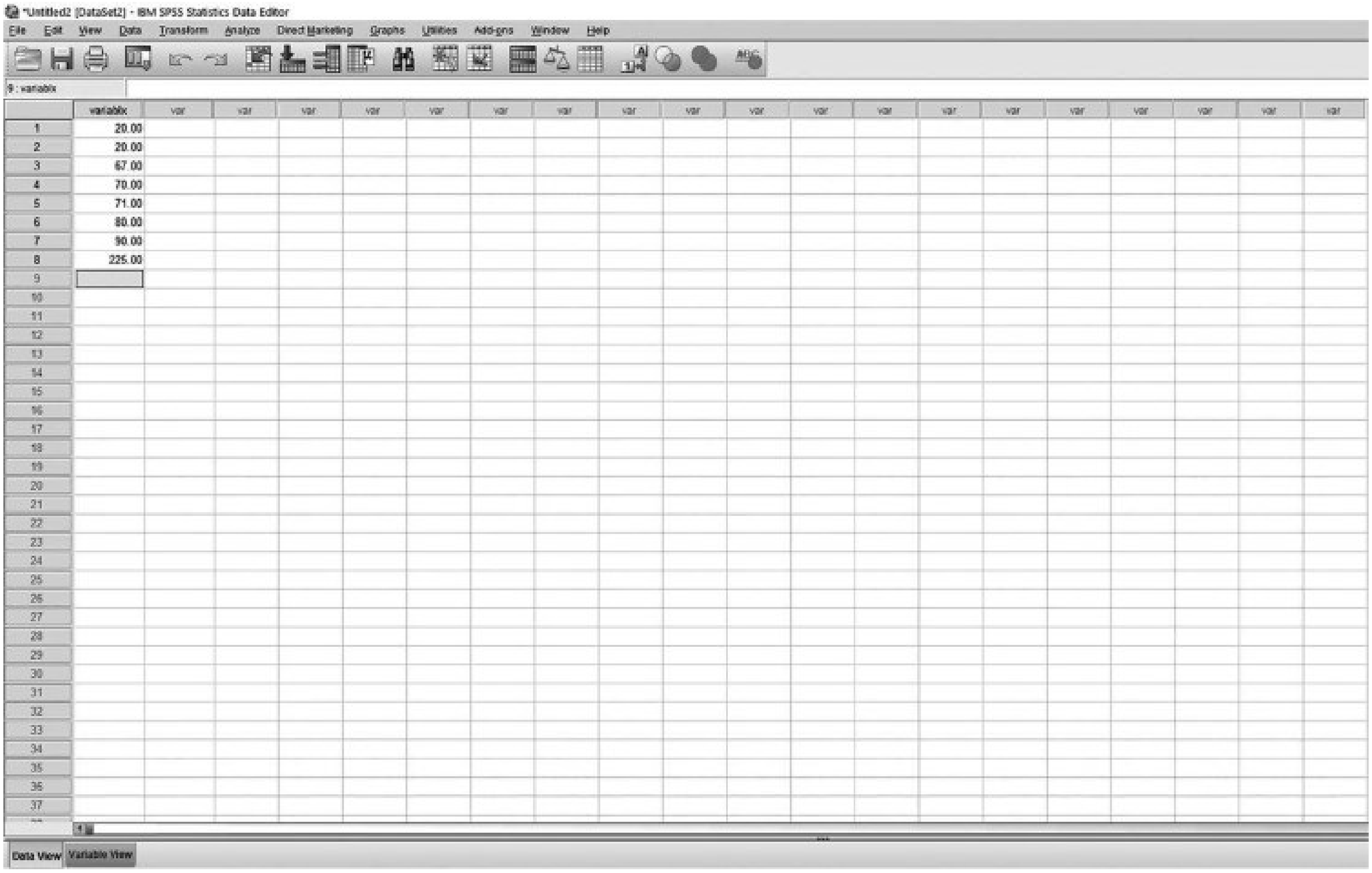
● Numeric
○ Comma
○ Dot
○ Scientific notation
○ Date
○ Dollar
○ Custom currency
○ String
○ Restricted Numeric (integer with leading zeros)

Width: 8
Decimal Places: 2

The Numeric type honors the digit grouping setting, while the Restricted Numeric never uses digit grouping.

OK | Cancel | Help

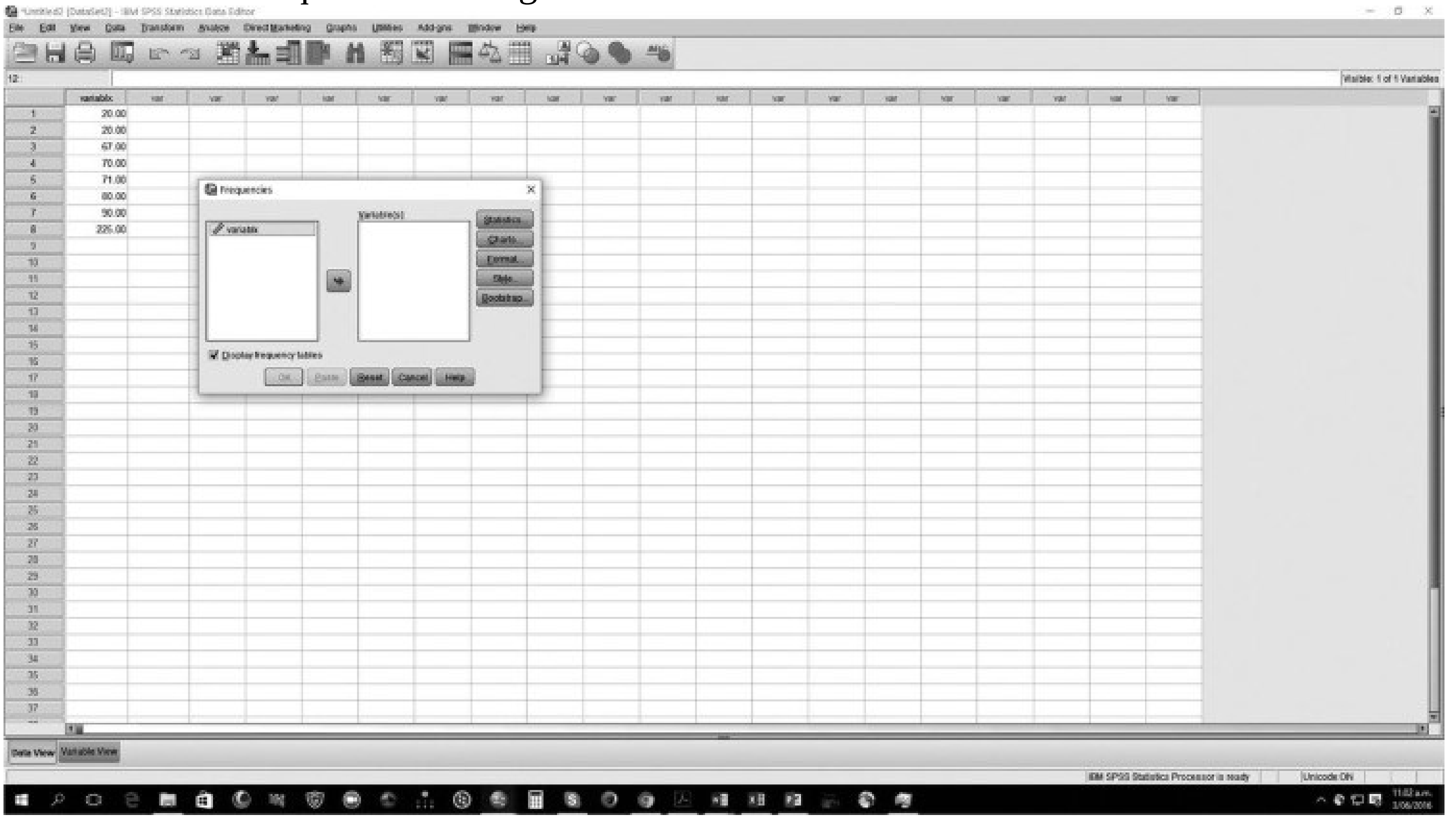
Screen W19.4 Entering Data



*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
9 : variablx
Data View
Variable View

	variablx
1	20.00
2	20.00
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	225.00

Screen W19.5 Frequencies Dialogue Box



*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
12:
Visible: 1 of 1 Variables
Frequencies
Variable(s):
Statistics...
Charts...
Format...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode ON
11:02 a.m.
3/06/2016

	variablx
1	20.00
2	20.00
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	225.00

variablx

☑ Display frequency tables

Univariate Analysis: Means and Frequency Distributions

You are now ready to perform your first SPSS data analysis. From the toolbar, select Analyze → Descriptive Statistics → Frequencies, which produces the dialogue box seen in Screen W19.5. Select variablx and then select the center arrow to move variablx to the Variable(s) box. Next, select the Statistics . . . button at the bottom of the dialogue box. This action produces another dialogue box, from which you can now select Mean, Median, and Mode, as shown by the checkmarks in Screen W19.6. Then select Continue.

You can also select other analyses. From the Frequency dialogue box, select Charts . . . → Histograms → Show normal curve, as shown in Screen W19.7. Now select Continue to proceed.

After you return to the Frequencies screen, select OK in the Frequencies dialogue box. SPSS will then perform the commands you just specified. Screen W19.8—called the Output screen (or Output Viewer)—will be produced. The scroll bar on the right-hand edge of the screen can be used to view the entire output. (*Note:* This screen also shows the syntax that SPSS shows, which is used to provide the output. To not see this, select Edit→ Options →Viewer → and deselect (turn off) “Display commands in the log” in the lower left-hand corner. That dialogue box has some other useful features you might want to use later, such as displaying variables name rather than variable, which we discuss later.) Specifically, the results shown in Table W19.1 and Figure W19.1 are produced. (*Note:* You can cut and paste these results into a word-processing program, which can be useful when preparing reports and slide shows. Table W19.1 and Figure W19.1 were reproduced in this way.) The number of bars is set by default.

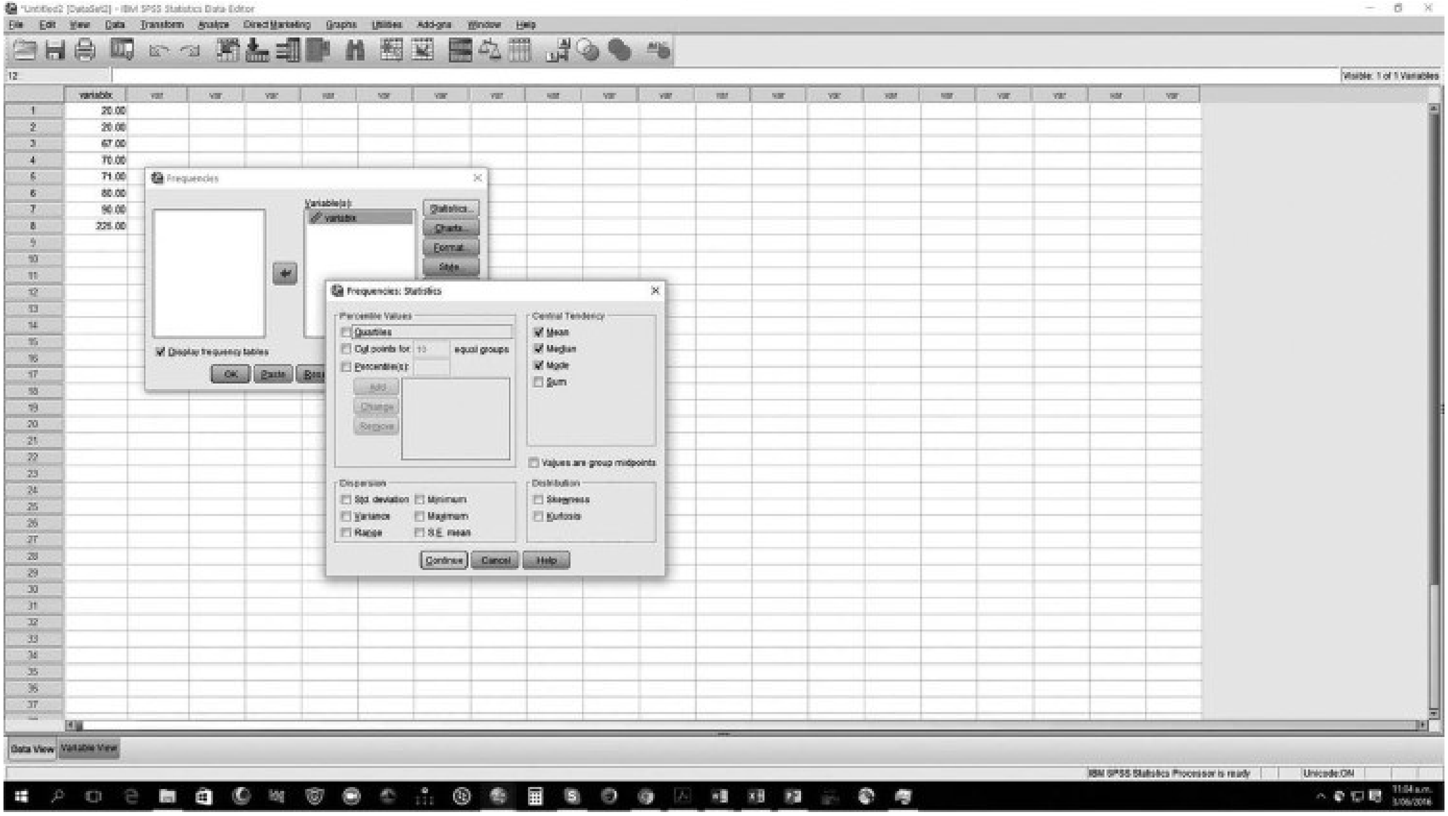
(*Note:* You can also adjust the number of bars shown in the histogram in Figure W19.1 by opening the Chart Editor. You can do this either by placing the cursor over the histogram on the Output screen and double clicking it or by selecting Edit → SPSS Chart Object → Open.) Once you see the Chart Editor dialogue box, select Options → Un-bin Elements (click again, then called “Bin Elements”) → Properties: Binning → change from “automatic” to “custom” and set the number of intervals. Select, for example, Number of intervals: 10, and then click Apply. (On older versions of SPSS, the commands are Edit → Select X-Axis → Histogram Options: Bin Sizes → Custom.) Screen W19.9 shows how this selection changes the display.

If you had created a bar chart rather than a histogram, a separate bar would be created for each value of the variable. To create a bar chart, select Analyze → Descriptive Statistics → Frequencies → Charts: Bar charts. You might want to open up the Chart Editor and then explore options for changing the bar chart.

Next, extend this example in the following way. The appendix to Chapter 7 (Appendix 7.1) in the textbook discusses how boxplots are used to determine whether individual observations are outliers. You might wonder whether the data point 225 is an outlier, as discussed in the textbook. SPSS can create a boxplot to find out. First, select from the toolbar Graphs → Legacy Dialogues → Boxplots. The dialogue box shown in Screen W19.10 will appear. (*Note:* recent versions of

SPSS now have guided dialogue boxes to build graphs, but quite a few folks find the legacy dialogues much faster! The choice is yours.)

Screen W19.6 Frequencies: Statistics Dialogue Box



	variablx
1	20.00
2	20.00
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	225.00

Frequencies

Variable(s): variablx

☑ Display frequency tables

Frequencies: Statistics

Percentile Values
□ Quartiles
□ Cut points for: 10 equal groups
□ Percentile(s):

Central Tendency
☑ Mean
☑ Median
☑ Mode
□ Sum

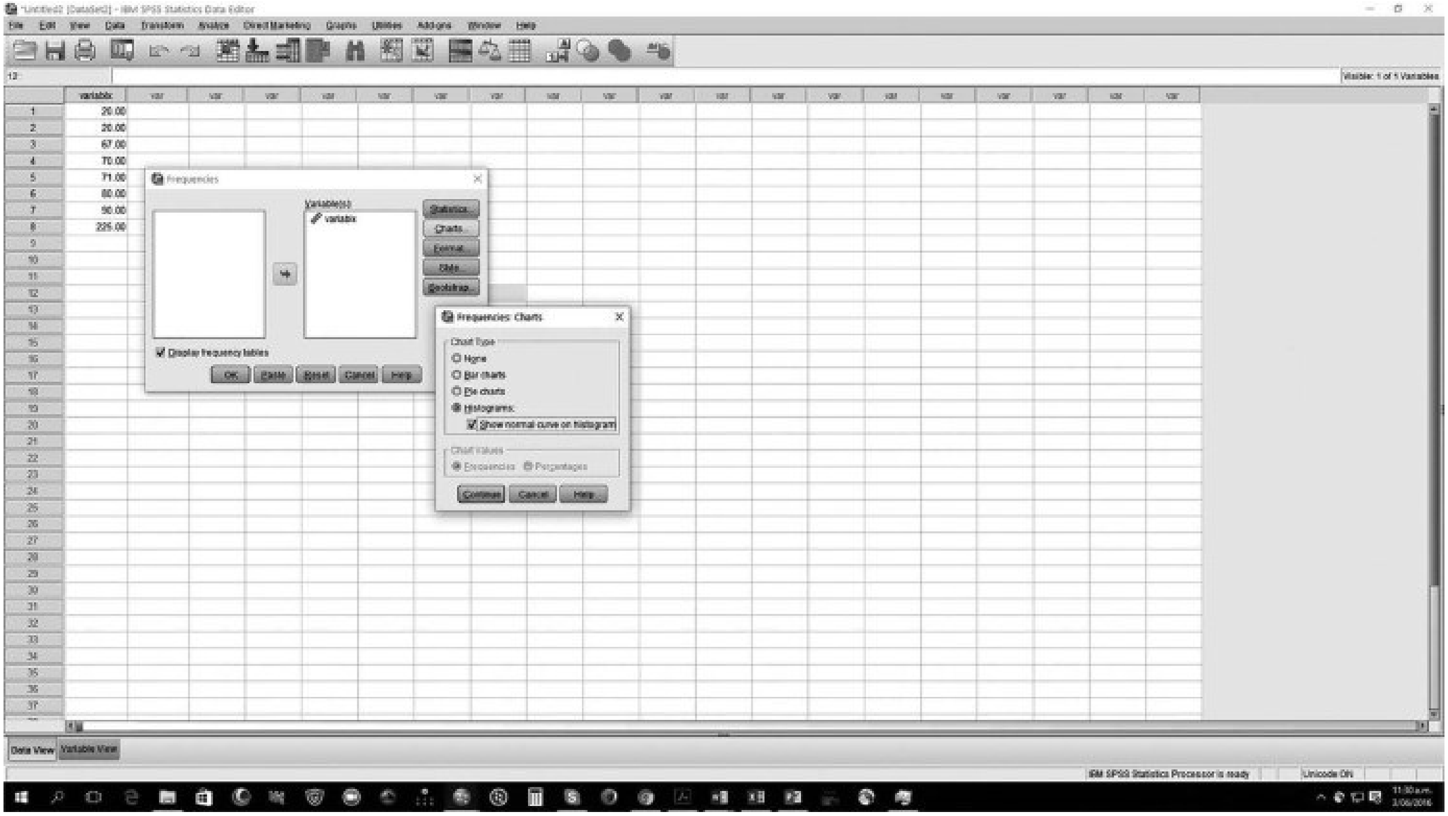
□ Values are group midpoints

Dispersion
□ Std. deviation □ Minimum
□ Variance □ Maximum
□ Range □ S.E. mean

Distribution
□ Skewness
□ Kurtosis

*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 1 of 1 Variables
Statistics...
Charts...
Format...
Style...
OK
Paste
Add
Change
Remove
Continue
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Screen W19.7 Creating a Histogram



Frequencies

Variable(s): variablx

☑ Display frequency tables

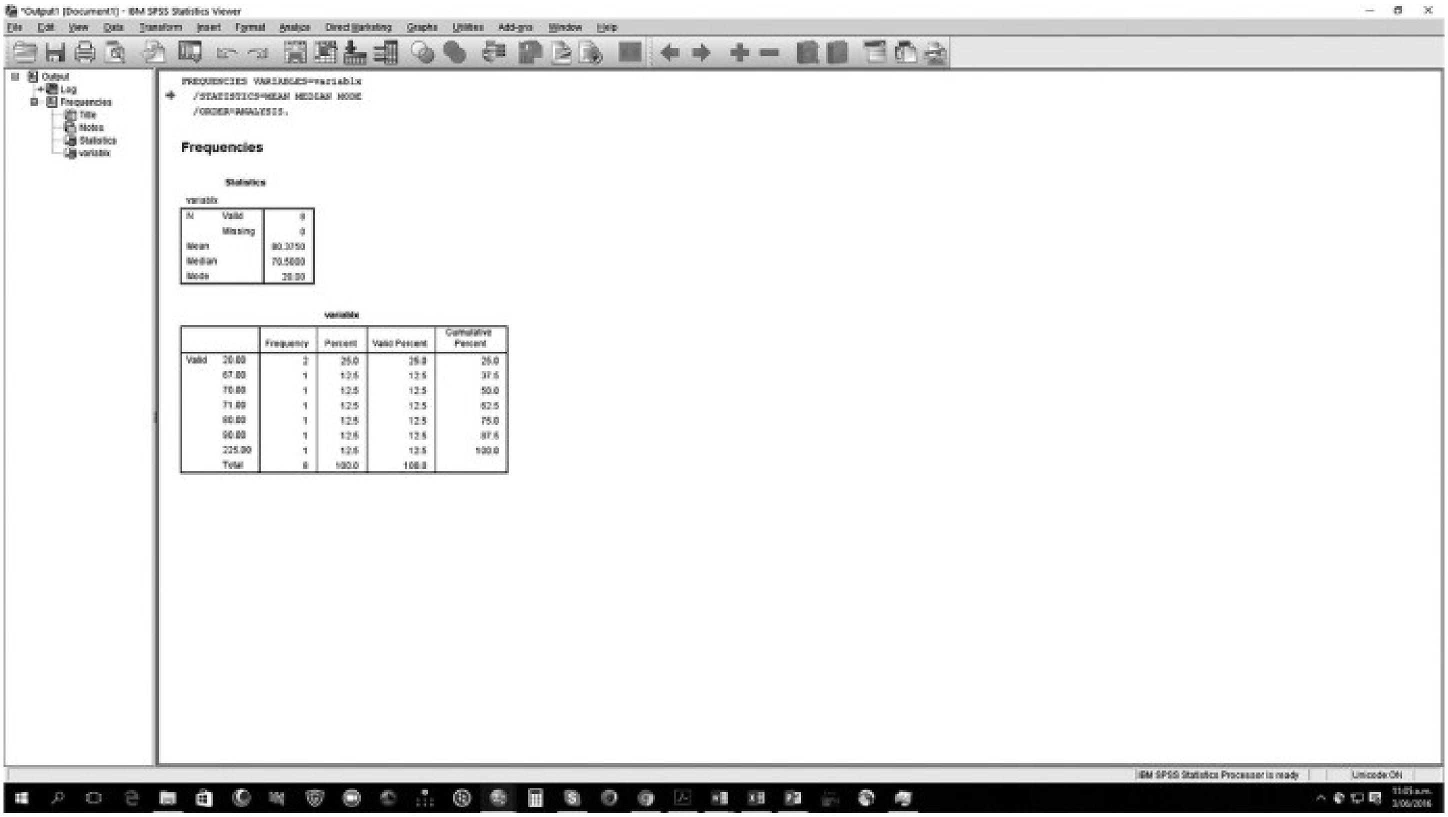
Frequencies: Charts

Chart Type
○ None
○ Bar charts
○ Pie charts
● Histograms:
☑ Show normal curve on histogram

Chart Values
● Frequencies ○ Percentages

*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 1 of 1 Variables
Statistics...
Charts...
Format...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Continue
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Screen W19.8 Output Screen



```
FREQUENCIES VARIABLES=variablx
  /STATISTICS=MEAN MEDIAN MODE
  /ORDER=ANALYSIS.
```

Frequencies

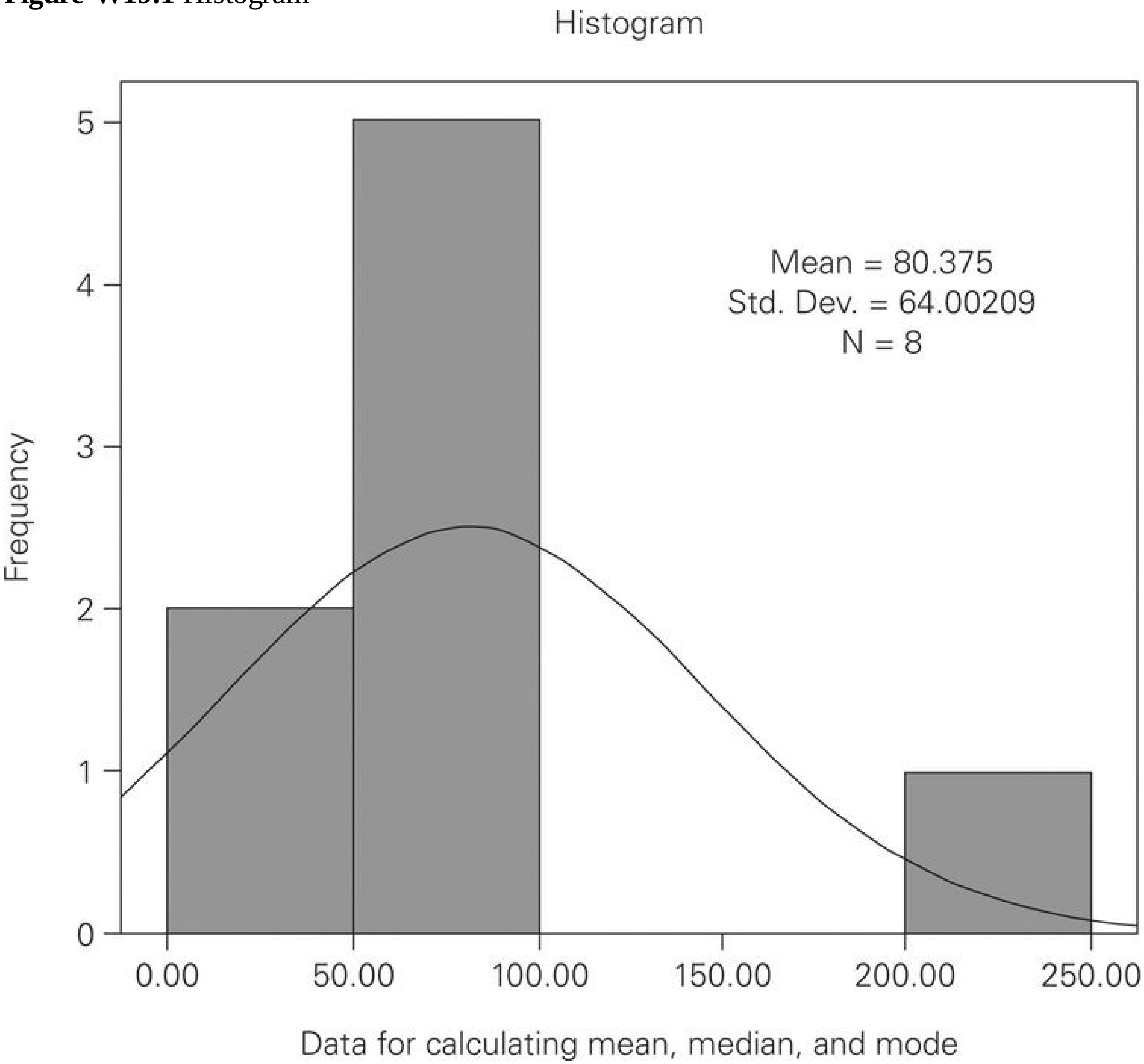
	Statistics	VARIABLX	
	N	Valid	8
		Missing	0
	Mean		80.3750
	Median		70.5000
	Mode		20.00

VARIABLX		Frequency	Percent	Valid percent	Cumulative percent
Valid	20.00	2	25.0	25.0	25.0
	67.00	1	12.5	12.5	37.5
	70.00	1	12.5	12.5	50.0
	71.00	1	12.5	12.5	62.5
	80.00	1	12.5	12.5	75.0
	90.00	1	12.5	12.5	87.5
	225.00	1	12.5	12.5	100.0
	Total	8	100.0	100.0	

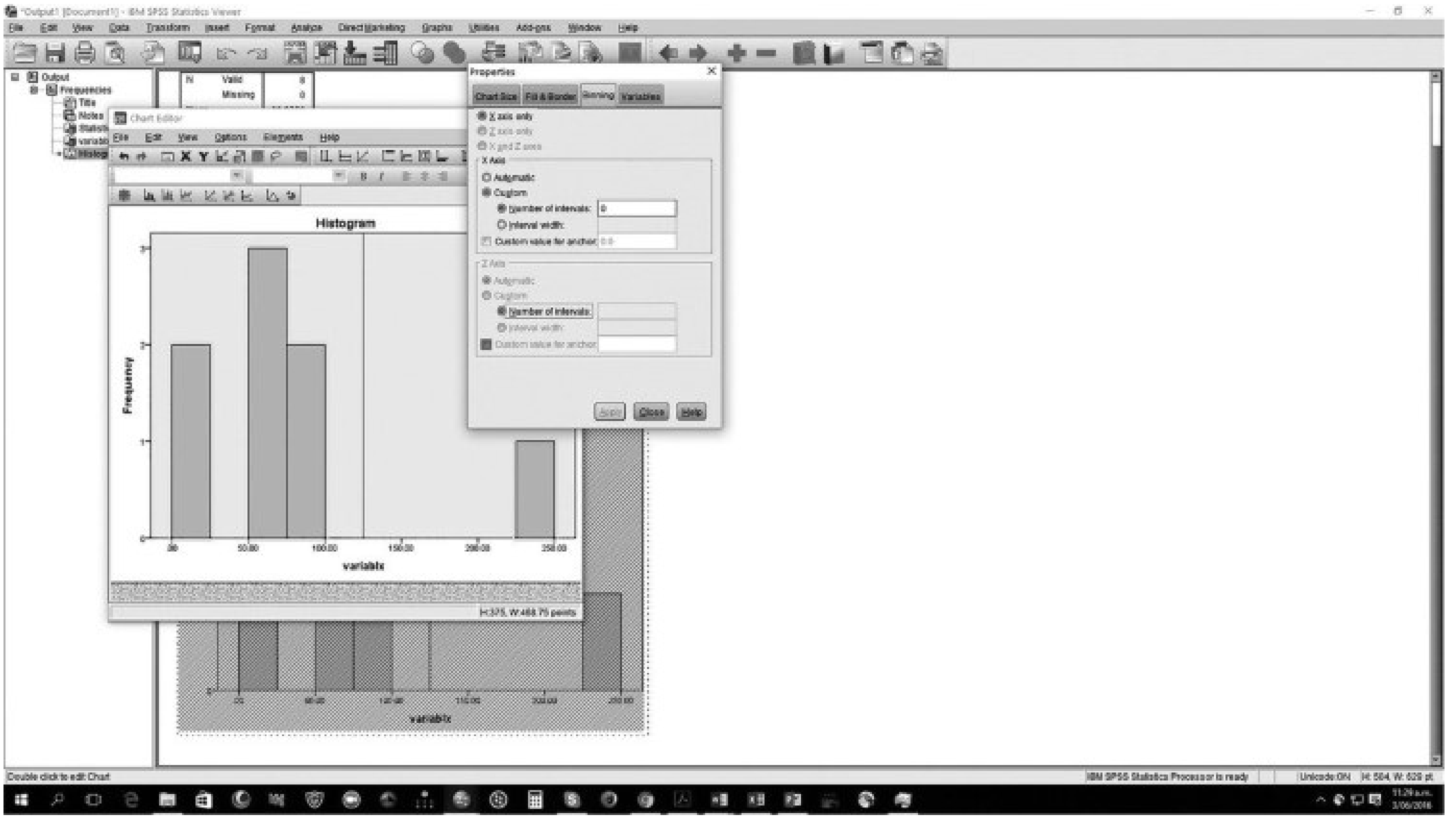
Select Summaries of separate variables → Define. The dialogue box shown in Screen W19.11 appears. Select the variable "variablx," and highlight the arrow to the left of the box labeled

"Boxes Represent" so that you can drag the variable to the Boxes Represent box. Then, to continue, select OK. The output is produced in the Output screen, as shown in Screen W19.12. Examine the display. Notice that the X-axis depicts the number of cases as being "8." Also note that the eighth observation in the dataset is an outlier. As part of examining a dataset, you should make a note of this observation because it may have an undue influence on further analyses. In this example, there are no observations with small values.

Figure W19.1 Histogram



Screen W19.9 Chart Editor

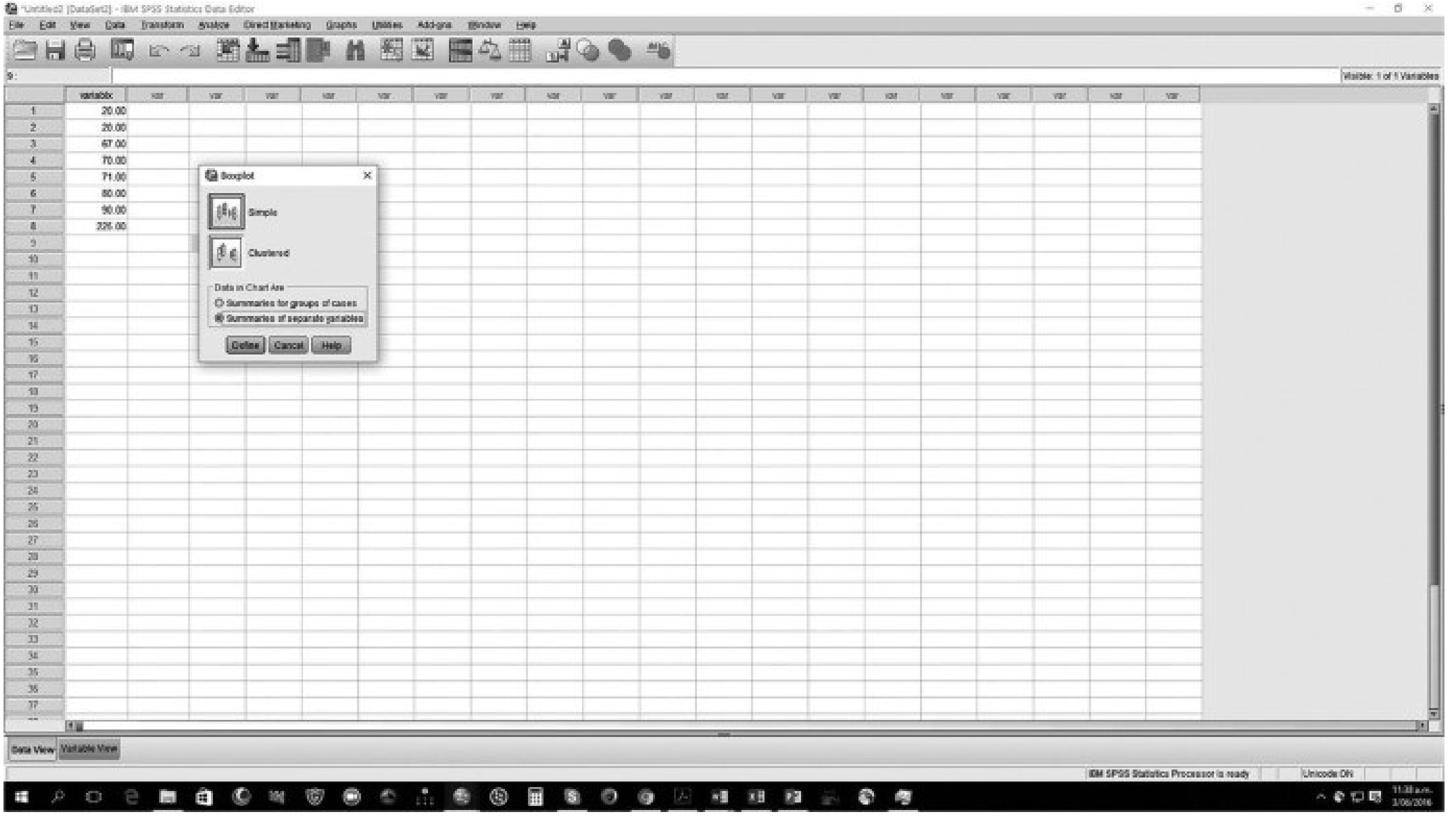


● X axis only
○ Z axis only
○ X and Z axes

X Axis
○ Automatic
● Custom
● Number of intervals: 8
○ Interval width:
□ Custom value for anchor: 0.0

In SPSS, the same output can often be produced in different ways. For example, to produce the boxplot shown in Screen W19.12, you could have used Analyze → Explore → Variable: variablx → OK. No one way is better than any other; both result in the same output. Just choose whatever method works best for you.

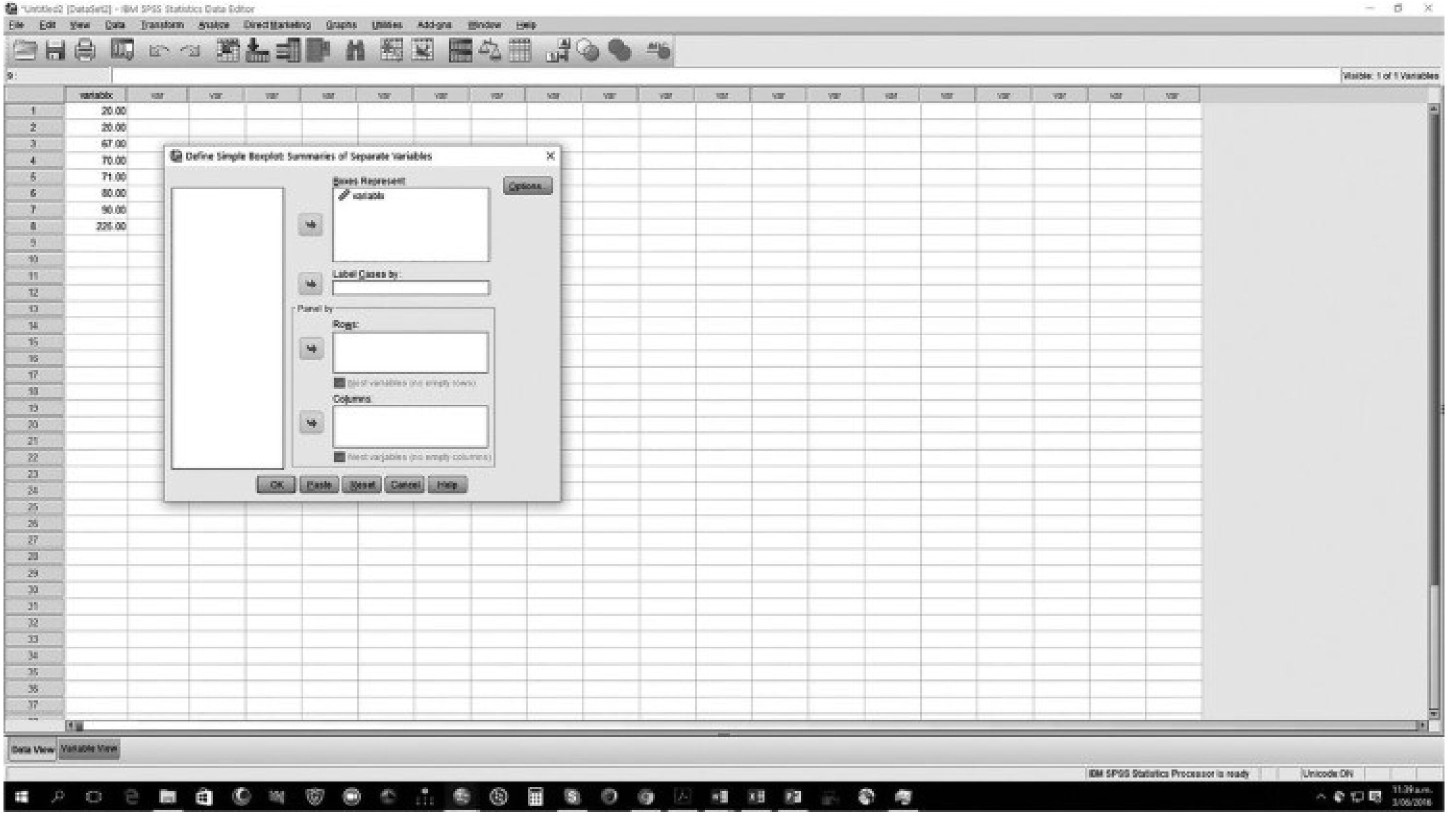
Screen W19.10 Boxplot Dialogue Box



	variablx
1	20.00
2	20.00
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	226.00

Data in Chart Are
○ Summaries for groups of cases
● Summaries of separate variables

Screen W19.11 Defining Boxplot Variables



	variable
1	20.00
2	20.00
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	225.00

Define Simple Boxplot: Summaries of Separate Variables

Boxes Represent: variable

Label Cases by:

Panel by

Rows:

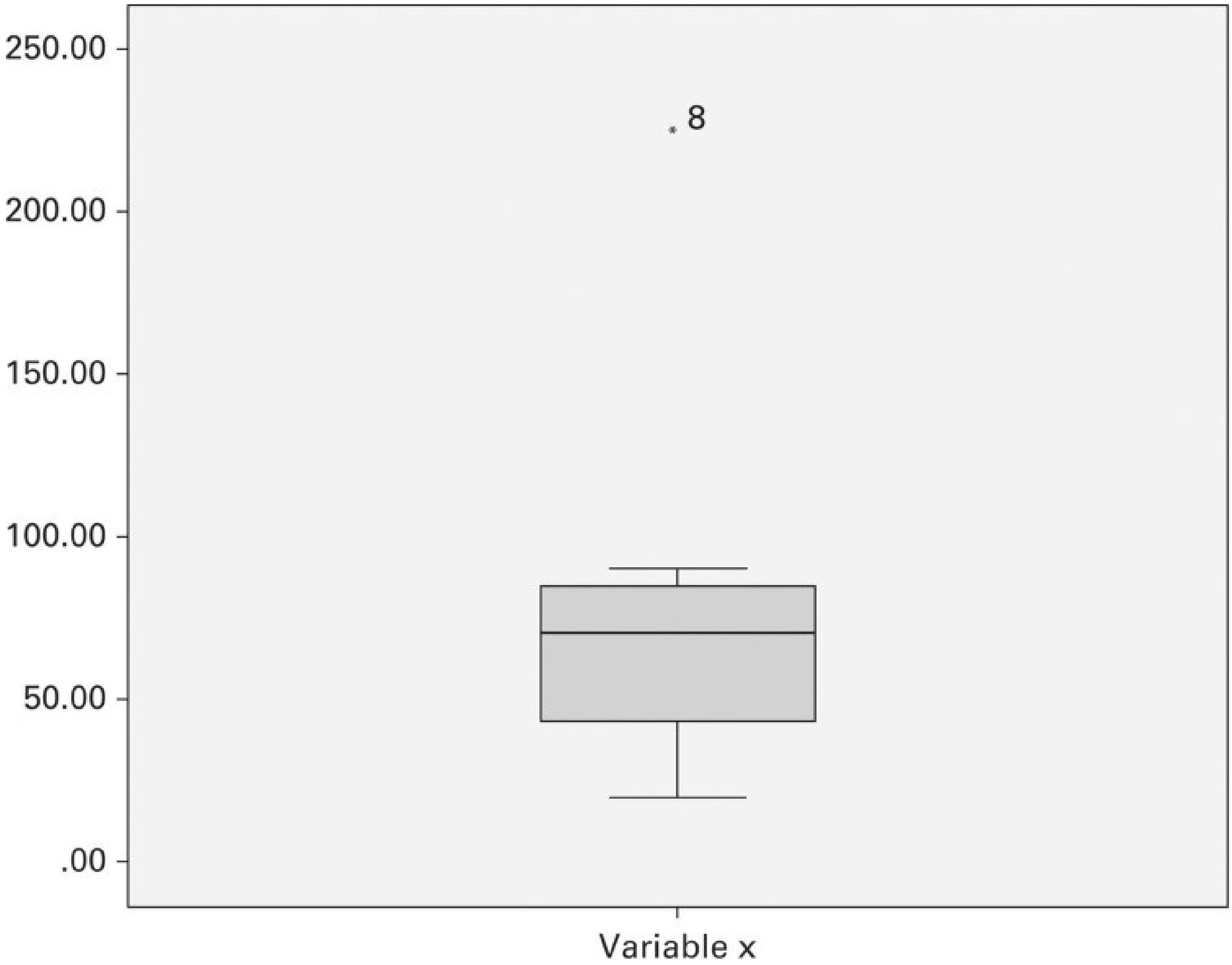
■ Nest variables (no empty rows)

Columns:

■ Nest variables (no empty columns)

To close the Output screen, select File → Close. At this point there is no need to save the output, so select No when asked if you would like to save it, and then return to the Data View screen. However, we will want to save this example dataset as “Example” because we will use it later to develop additional examples and SPSS demonstrations. To save the dataset, use File → Save As → File Name: Example → Save. Be sure to note the location of the saved file (for example, in a folder on a hard drive or on a portable memory device) so that you can find it again.

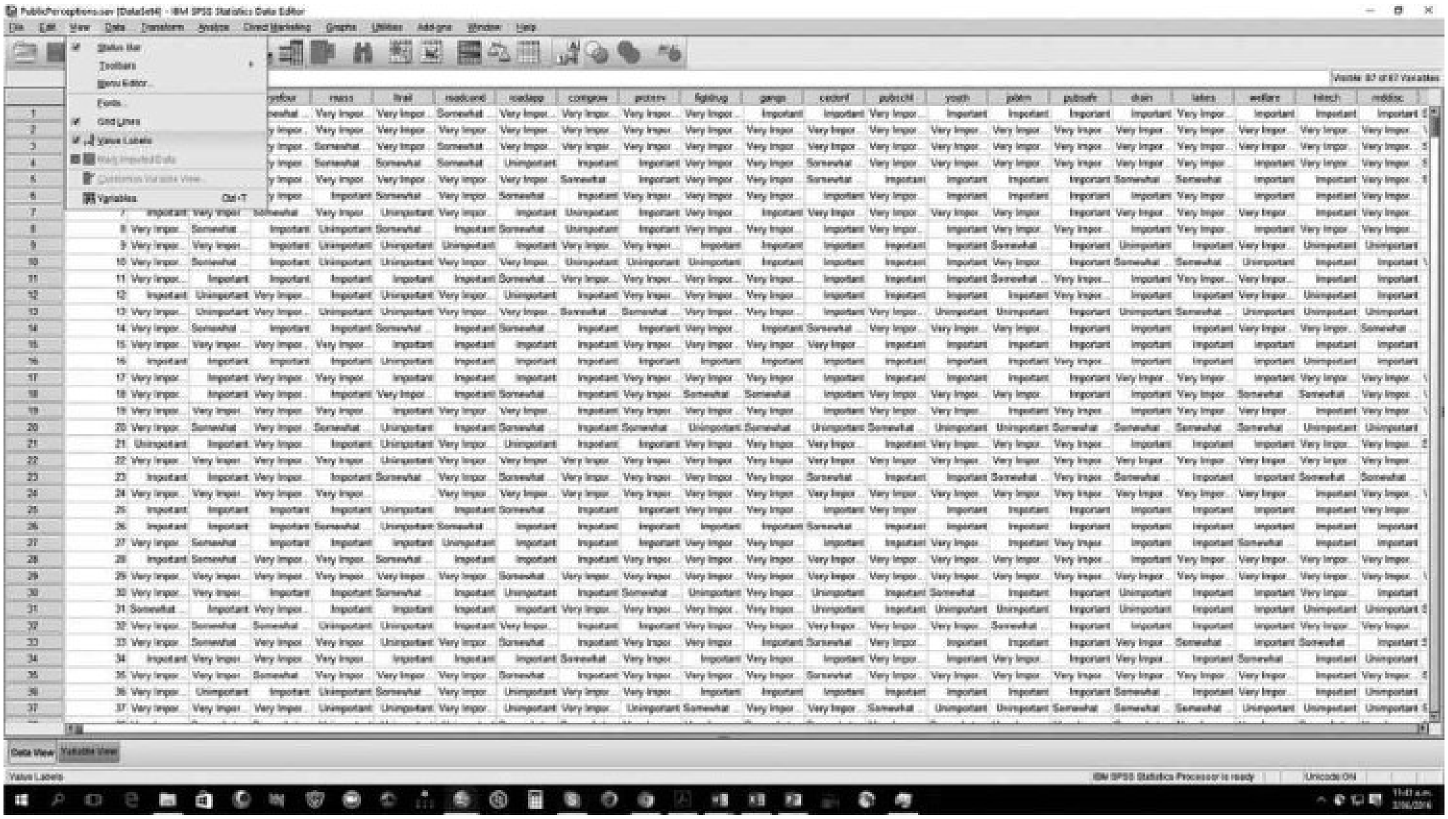
Screen W19.12 Boxplot Output Screen


250.00
200.00
150.00
100.00
50.00
.00
8
Variable x

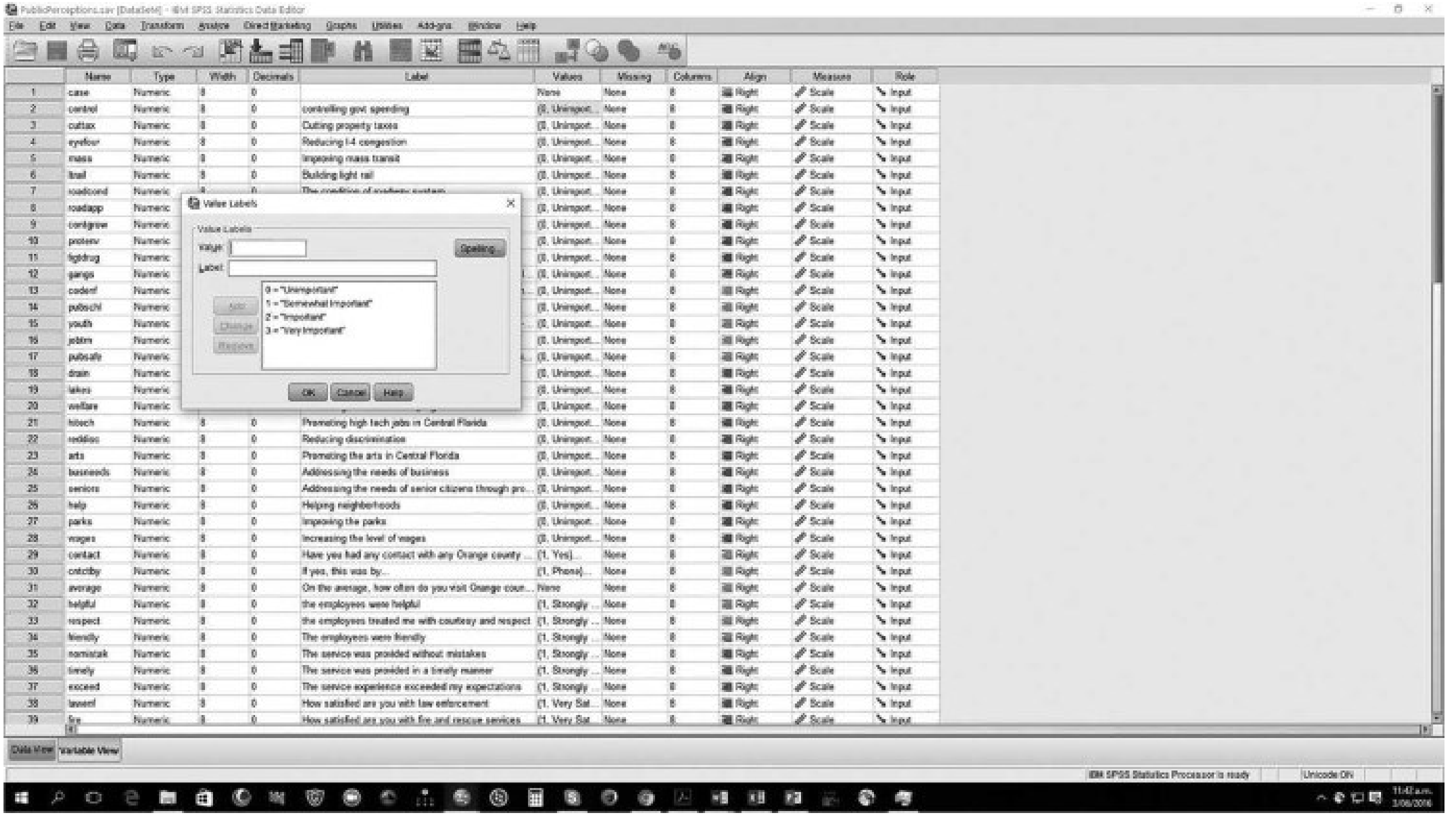
Variable Labels and Values

Open the Public Perceptions dataset, which can be found on this workbook's companion website (http://study.sagepub.com/bermaness4e), and be sure you are in the Data View screen. SPSS can show the data both as numbers and as value labels. If you see all labels, as in Screen W19.13, you might want to see the coded numbers instead. To do so, select View → Value Labels. When you see value labels, the Value Labels option has been selected, as shown in Screen W19.13. (Note the check mark next to Value Labels.) To view numbers, simply click on Value Labels to switch off this option (the check mark will disappear). Then the fields will show numbers.

Screen W19.13 Value Labels



Screen W19.14 Value Labels Dialogue Box



Value Labels

Value Labels

Value:

Label:

0 = "Unimportant"
1 = "Somewhat Important"
2 = "Important"
3 = "Very Important"

Add | Change | Remove | Spelling...

OK | Cancel | Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	case	Numeric	8	0		None	None	8	Right	Scale	Input
2	control	Numeric	8	0	controlling govt spending	{0, Unimport...	None	8	Right	Scale	Input
3	cuttax	Numeric	8	0	Cutting property taxes	{0, Unimport...	None	8	Right	Scale	Input
4	eyefour	Numeric	8	0	Reducing I-4 congestion	{0, Unimport...	None	8	Right	Scale	Input
5	mass	Numeric	8	0	Improving mass transit	{0, Unimport...	None	8	Right	Scale	Input
6	lrail	Numeric	8	0	Building light rail	{0, Unimport...	None	8	Right	Scale	Input
7	roadcond	Numeric	8	0	The condition of roadway system	{0, Unimport...	None	8	Right	Scale	Input
8	roadapp	Numeric				{0, Unimport...	None	8	Right	Scale	Input
9	contgrow	Numeric				{0, Unimport...	None	8	Right	Scale	Input
10	protenv	Numeric				{0, Unimport...	None	8	Right	Scale	Input
11	fgtdrug	Numeric				{0, Unimport...	None	8	Right	Scale	Input
12	gangs	Numeric				{0, Unimport...	None	8	Right	Scale	Input
13	codenf	Numeric				{0, Unimport...	None	8	Right	Scale	Input
14	pubschl	Numeric				{0, Unimport...	None	8	Right	Scale	Input
15	youth	Numeric				{0, Unimport...	None	8	Right	Scale	Input
16	jobtrn	Numeric				{0, Unimport...	None	8	Right	Scale	Input
17	pubsafe	Numeric				{0, Unimport...	None	8	Right	Scale	Input
18	drain	Numeric				{0, Unimport...	None	8	Right	Scale	Input
19	lakes	Numeric				{0, Unimport...	None	8	Right	Scale	Input
20	welfare	Numeric				{0, Unimport...	None	8	Right	Scale	Input
21	hitech	Numeric	8	0	Promoting high tech jobs in Central Florida	{0, Unimport...	None	8	Right	Scale	Input
22	reddisc	Numeric	8	0	Reducing discrimination	{0, Unimport...	None	8	Right	Scale	Input
23	arts	Numeric	8	0	Promoting the arts in Central Florida	{0, Unimport...	None	8	Right	Scale	Input
24	busneeds	Numeric	8	0	Addressing the needs of business	{0, Unimport...	None	8	Right	Scale	Input
25	seniors	Numeric	8	0	Addressing the needs of senior citizens through pro...	{0, Unimport...	None	8	Right	Scale	Input
26	help	Numeric	8	0	Helping neighborhoods	{0, Unimport...	None	8	Right	Scale	Input
27	parks	Numeric	8	0	Improving the parks	{0, Unimport...	None	8	Right	Scale	Input
28	wages	Numeric	8	0	Increasing the level of wages	{0, Unimport...	None	8	Right	Scale	Input
29	contact	Numeric	8	0	Have you had any contact with any Orange county ...	{1, Yes}...	None	8	Right	Scale	Input
30	cntctby	Numeric	8	0	If yes, this was by...	{1, Phone}...	None	8	Right	Scale	Input
31	average	Numeric	8	0	On the average, how often do you visit Orange coun...	None	None	8	Right	Scale	Input
32	helpful	Numeric	8	0	the employees were helpful	{1, Strongly ...	None	8	Right	Scale	Input
33	respect	Numeric	8	0	the employees treated me with courtesy and respect	{1, Strongly ...	None	8	Right	Scale	Input
34	friendly	Numeric	8	0	The employees were friendly	{1, Strongly ...	None	8	Right	Scale	Input
35	nomistak	Numeric	8	0	The service was provided without mistakes	{1, Strongly ...	None	8	Right	Scale	Input
36	timely	Numeric	8	0	The service was provided in a timely manner	{1, Strongly ...	None	8	Right	Scale	Input
37	exceed	Numeric	8	0	The service experience exceeded my expectations	{1, Strongly ...	None	8	Right	Scale	Input
38	lawenf	Numeric	8	0	How satisfied are you with law enforcement	{1, Very Sat...	None	8	Right	Scale	Input
39	fire	Numeric	8	0	How satisfied are you with fire and rescue services	{1, Very Sat...	None	8	Right	Scale	Input

You can also view the value labels in the Variable View screen. In the Variable View screen, select the cell in the Values column that corresponds, for example, with the variable "control." Click on that cell and double click on the square icon that appears in the cell, which will generate the dialogue box shown in Screen W19.14. This box defines the values that correspond with the labels "Very Important," "Important," "Somewhat Important," and "Unimportant." Often, you will want to assign values to a variable. For example, on a scale of 1 to 4, a 1 might indicate that a respondent is "very satisfied" regarding some survey question. Such labels later appear on your output as well, greatly improving presentation.

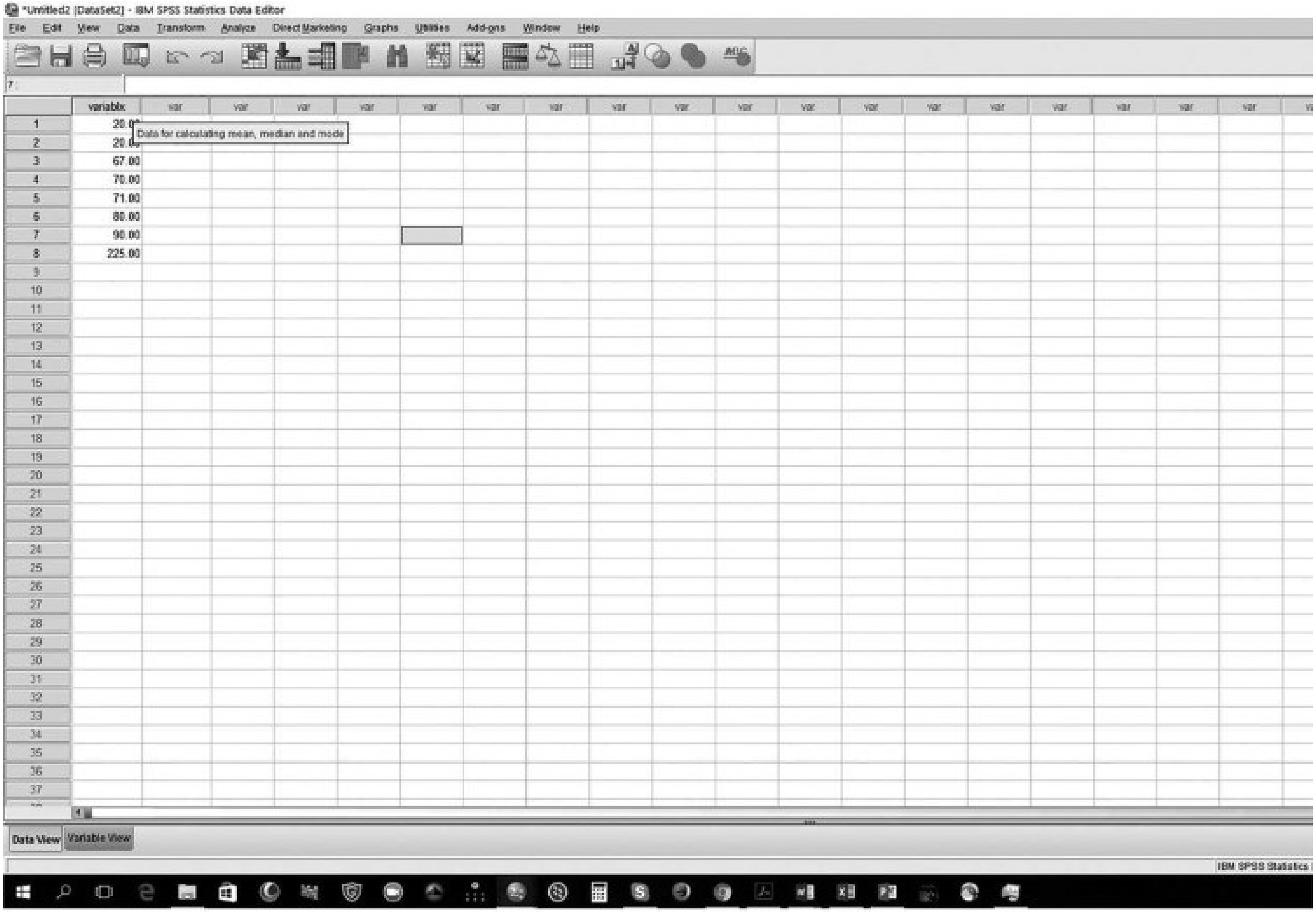
Now you should practice generating value labels on the Example dataset you saved earlier. Assigning labels is quite simple. Open the Example dataset. The screen shown in Screen W19.15 should be displayed. Select the Variable View screen. First, create a new variable with which you will practice. In row 2 of the Name column, type "dataskil," which is the name of your variable, shown in Screen W19.16. In the Label column, type "Importance of data analysis skills." This variable represents a hypothetical question posed to public sector managers concerning the importance of data analysis skills. (Double clicking on the Label data box allows label entry and editing of entries.)

Now define data values. Select the box in the Values column data cell for the variable "dataskil." Enter "1" in the Value box and then "Very Important" in the Value Label box. The dialogue box should now look as in Screen W19.16. Select Add. This causes the definition to be entered into SPSS programming. After you select Add, Screen W19.17 should appear. Repeat this process. Define the scale values "Important," "Somewhat Important," and Unimportant," which results in Screen W19.18. Select OK to complete the value definitions for this variable. All that remains is for the researcher to enter the data for this hypothetical question.

Defining Missing Values

A common problem is the coding and handling of missing values. Missing values are blank cells in the Data View screen when an observation has a missing value for a specific variable. Sometimes, however, analysts want to assign a specific value for missing values. Doing so helps differentiate between a value that an analyst has forgotten to enter and a value that really is missing, which occurs when a respondent does not answer a particular question. Suppose you wish to assign the value of 9 as our way of coding missing values. Select the Variable View screen, and highlight the cell in the second row of the Missing column. Screen W19.19 will be displayed. Next, select Discrete missing values, and enter "9" in the first block, and then select OK. This will cause every 9 for the given variable (in this case, dataskil) to be treated as missing. Proper definition of missing data is extremely important to ensure accurate data analysis.

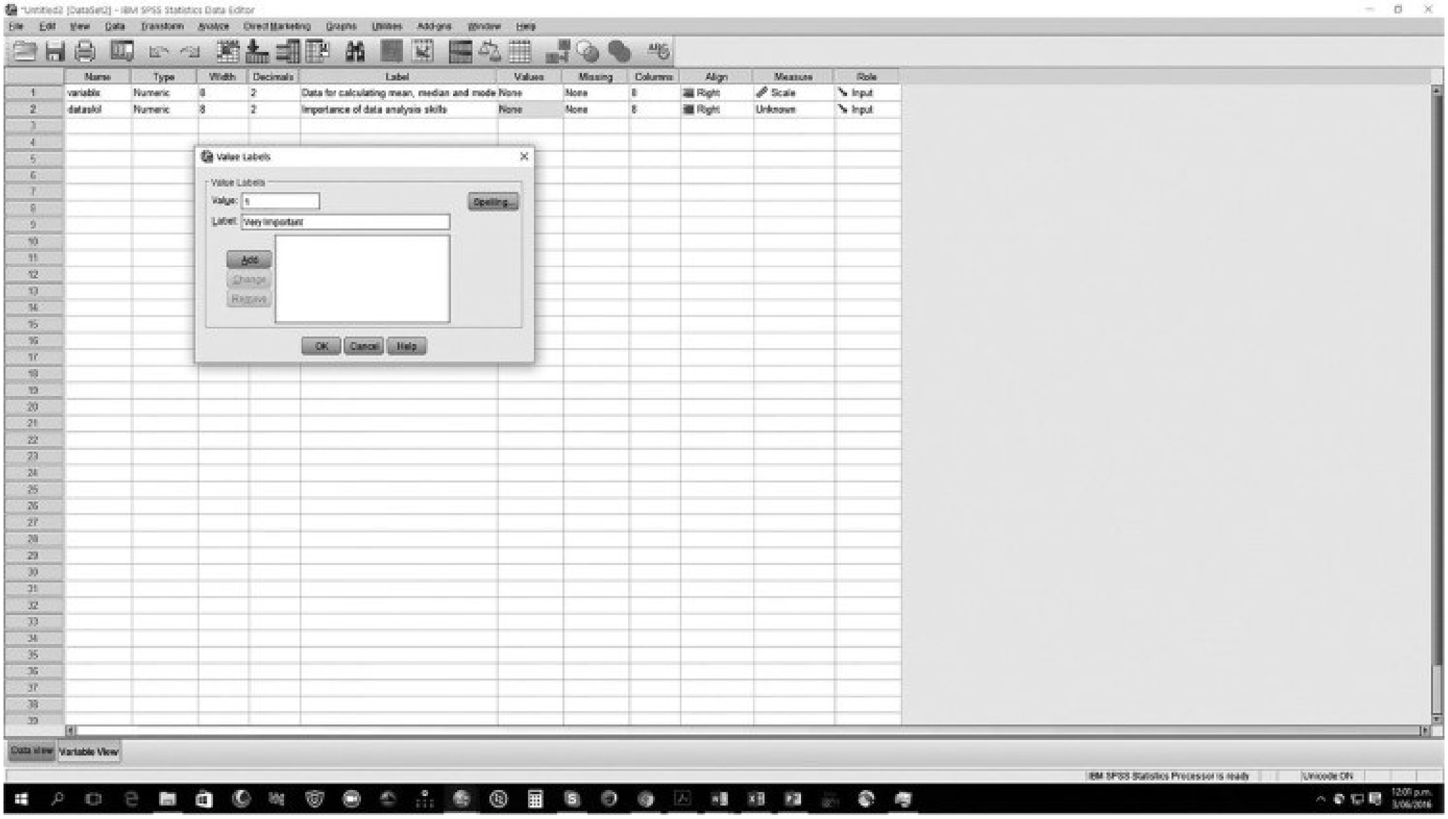
Screen W19.15 Generating Labels


*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor

	variablx
1	20.0
2	20.0
3	67.00
4	70.00
5	71.00
6	80.00
7	90.00
8	225.00

Data for calculating mean, median and mode

Screen W19.16 Defining Data Values


*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Value Labels
Value Labels
Spelling...
Add
Change
Remove
OK
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

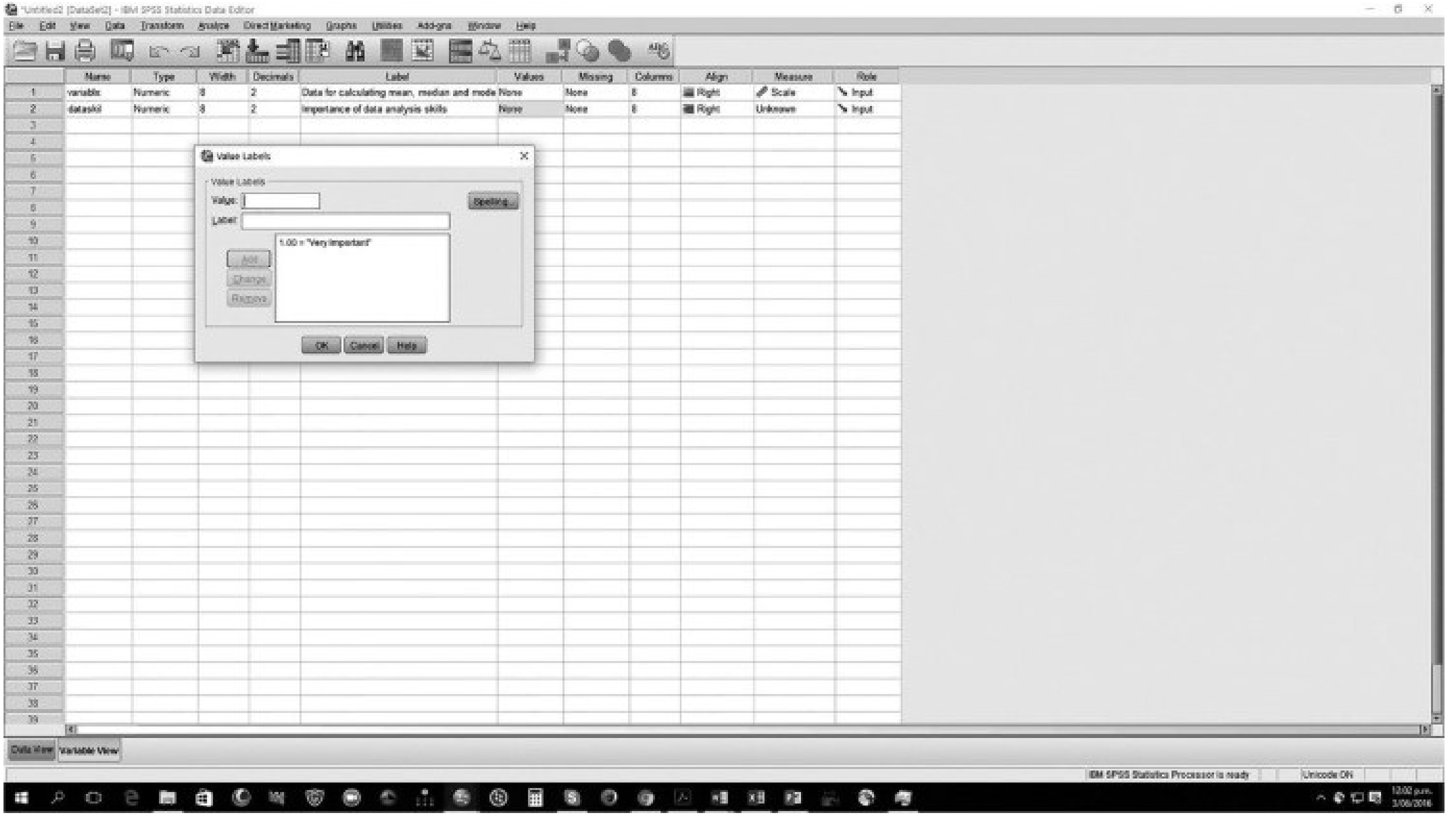
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2	Data for calculating mean, median and mode	None	None	8	Right	Scale	Input
2	dataskil	Numeric	8	2	Importance of data analysis skills	None	None	8	Right	Unknown	Input

Value: 1

Label: Very important

Screen W19.20 shows another highly useful feature of SPSS. Variable definitions (for example, missing values, data values, or labels) are readily copied from one variable to a range of other variables. Thus, you need to define missing values and data values (labels) only once. Indeed, after you create more new variables, newvar1 through newvar5, as shown in Screen W19.20, simply copy and paste the Values definition of the variable dataskil to these new cells. You do not need to define these values for each new variable. Similarly, you can now copy and paste the missing values definition from the variable dataskil to these new variables, too.

Screen W19.17 Adding Labels



*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2	Data for calculating mean, median and mode	None	None	8	Right	Scale	Input
2	dataskil	Numeric	8	2	Importance of data analysis skills	None	None	8	Right	Unknown	Input

Value Labels

Value Labels

Value:

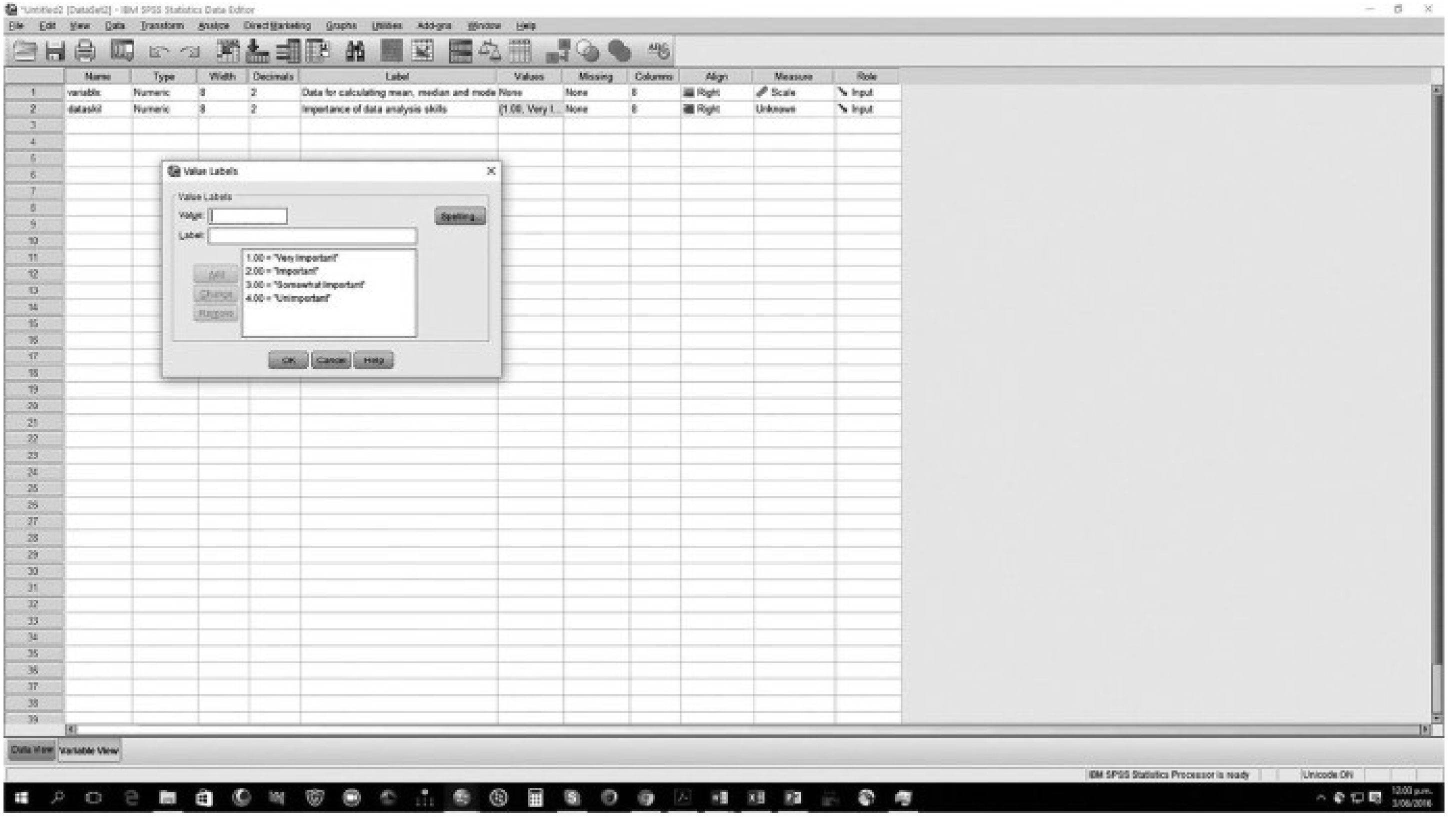
Label:

1.00 = "Very Important"

IBM SPSS Statistics Processor is ready

Unicode:ON

Screen W19.18 Value Label Definitions



*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2	Data for calculating mean, median and mode	None	None	8	Right	Scale	Input
2	dataskil	Numeric	8	2	Importance of data analysis skills	{1.00, Very I...	None	8	Right	Unknown	Input

Value Labels

Value Labels

Value:

Label:

1.00 = "Very Important"
2.00 = "Important"
3.00 = "Somewhat Important"
4.00 = "Unimportant"

IBM SPSS Statistics Processor is ready

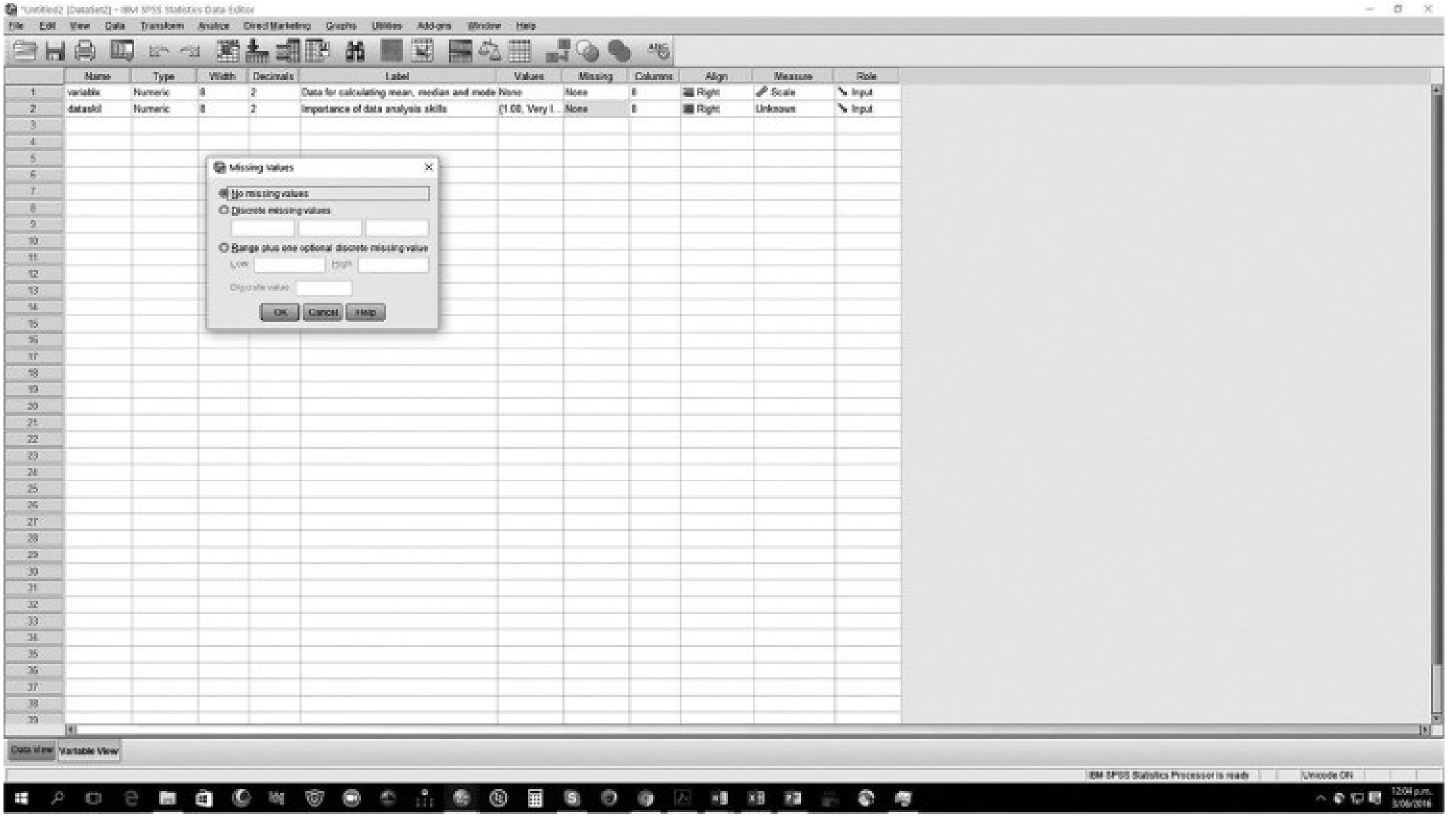
Unicode:ON

Selecting a Subset of Observations for Analysis

Often, analysts want to perform their analyses on a subset of observations. Using the Public Perceptions dataset, assume that you wish to assess perceptions of client satisfaction, but *only* among those respondents who have had contact with a county employee during the past year. Then, you would want to select for subsequent analysis only those observations that meet this condition.

Open the Public Perceptions dataset. The variable “contact” (row 29) measures whether the respondent has had contact with county employees within the past year. To filter out respondents not meeting this criterion, perform the following sequence of commands. From the toolbar, select Data → Select Cases → If condition is satisfied. Add “contact” to the right-hand box, and type “=1.” Screen W19.21 is displayed. Select Continue → OK. Note the left-hand column of the Data View screen. Observations not meeting our new requirement of having contact with county employees during the past year now show a diagonal line through their respective sequence number, indicating that they are not included in subsequent calculations (Screen W19.22). (To undo this restriction, simply choose Data → Select Cases → Select: All cases → OK.)

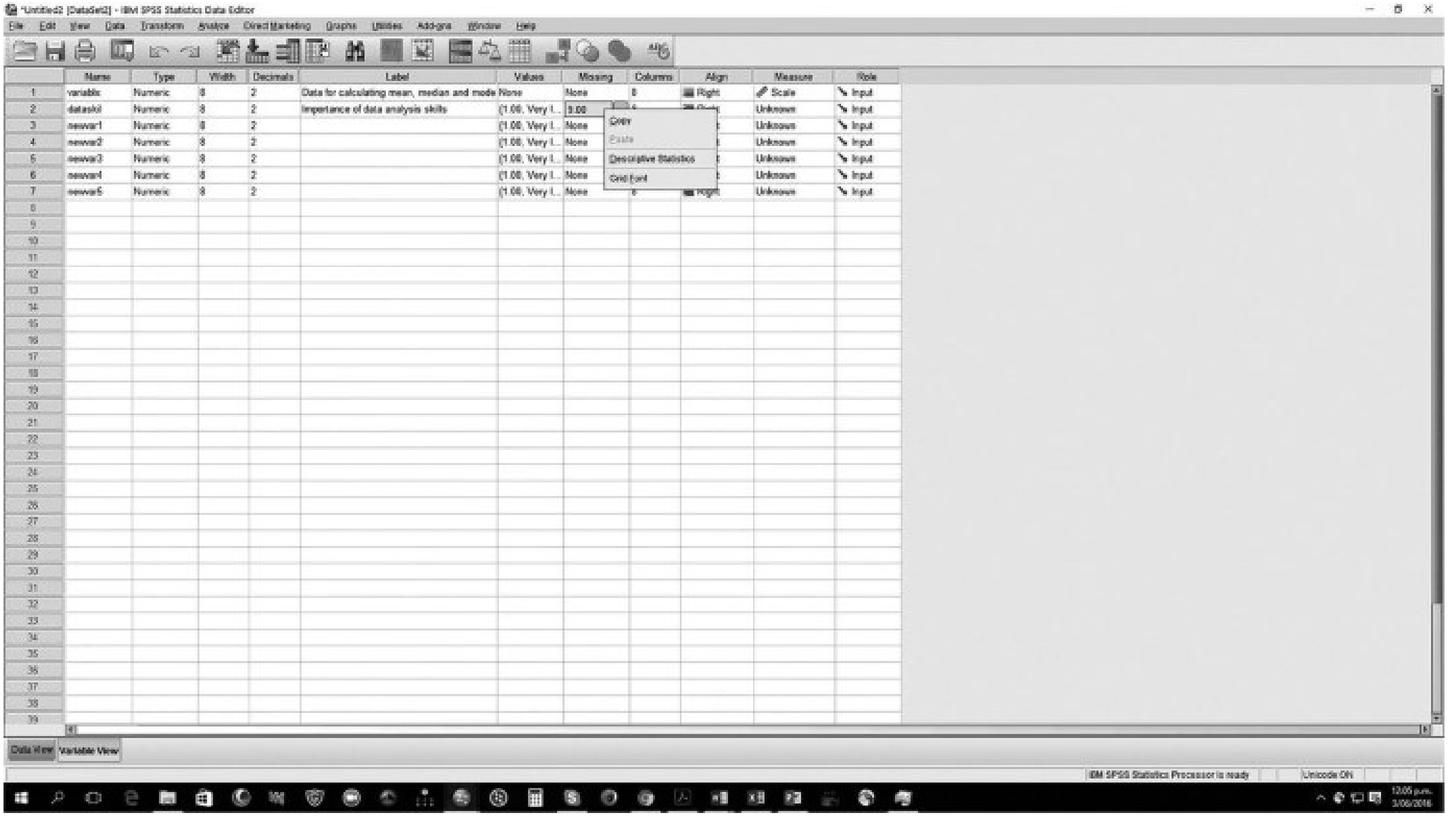
Screen W19.19 Missing Values Dialogue Box


*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Missing Values
OK
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
12:04 p.m.
3/06/2016

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2	Data for calculating mean, median and mode	None	None	8	Right	Scale	Input
2	dataskil	Numeric	8	2	Importance of data analysis skills	{1.00, Very I...	None	8	Right	Unknown	Input

● No missing values
○ Discrete missing values
○ Range plus one optional discrete missing value
Low: High:
Discrete value:

Screen W19.20 Copying Variable Definitions

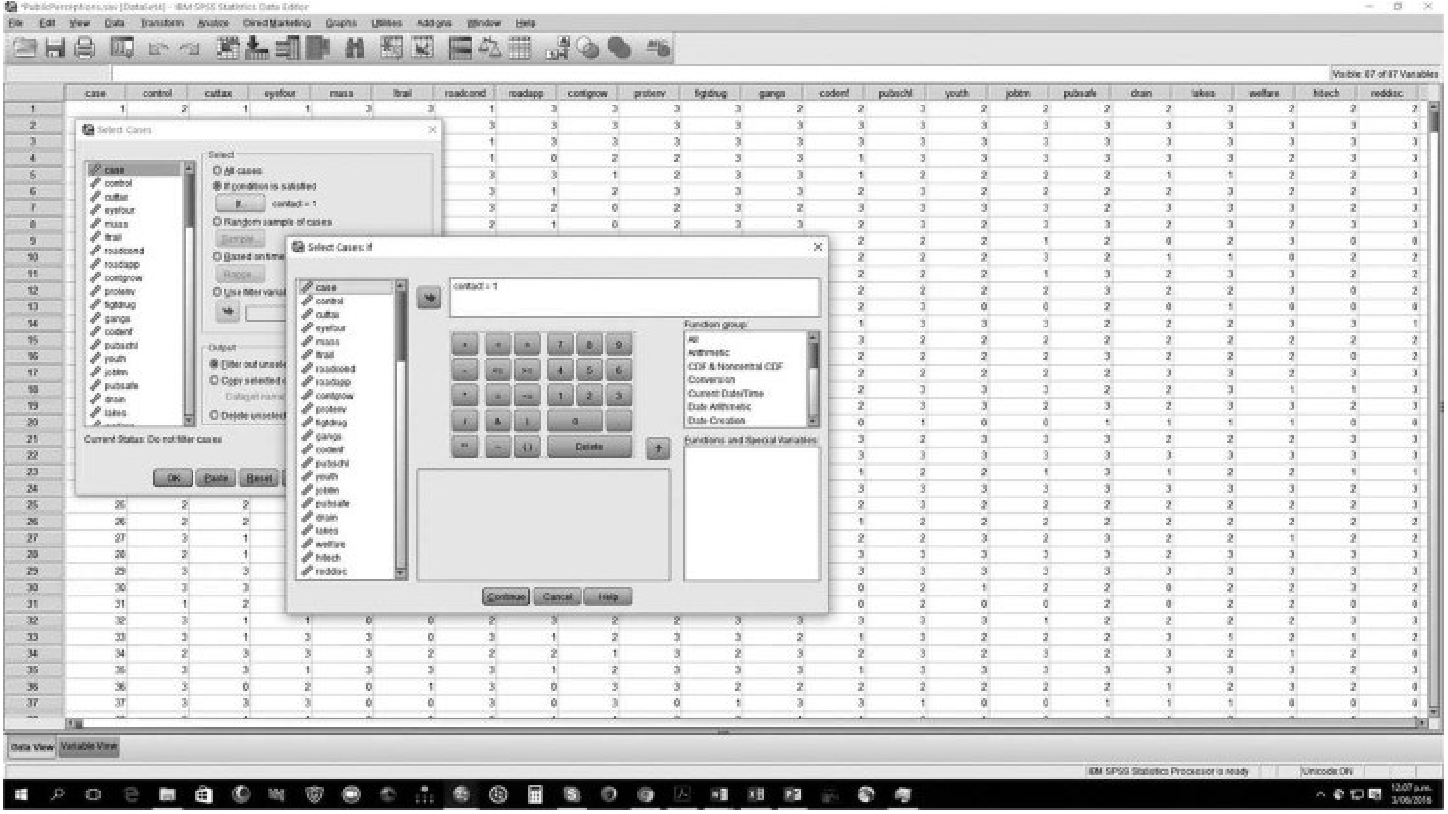


*Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	variable	Numeric	8	2	Data for calculating mean, median and mode	None	None	8	Right	Scale	Input
2	dataskil	Numeric	8	2	Importance of data analysis skills	{1.00, Very I...	3.00			Unknown	Input
3	newvar1	Numeric	8	2		{1.00, Very I...	None			Unknown	Input
4	newvar2	Numeric	8	2		{1.00, Very I...	None			Unknown	Input
5	newvar3	Numeric	8	2		{1.00, Very I...	None			Unknown	Input
6	newvar4	Numeric	8	2		{1.00, Very I...	None			Unknown	Input
7	newvar5	Numeric	8	2		{1.00, Very I...	None	8	Right	Unknown	Input

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Copy
Paste
Descriptive Statistics
Grid Font
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
12:05 p.m.
3/06/2016

Screen W19.21 Selecting Cases (Filtering Observations)



*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor

Select Cases

Select
○ All cases
● If condition is satisfied
If... contact = 1
○ Random sample of cases
○ Based on time
○ Use filter vari...

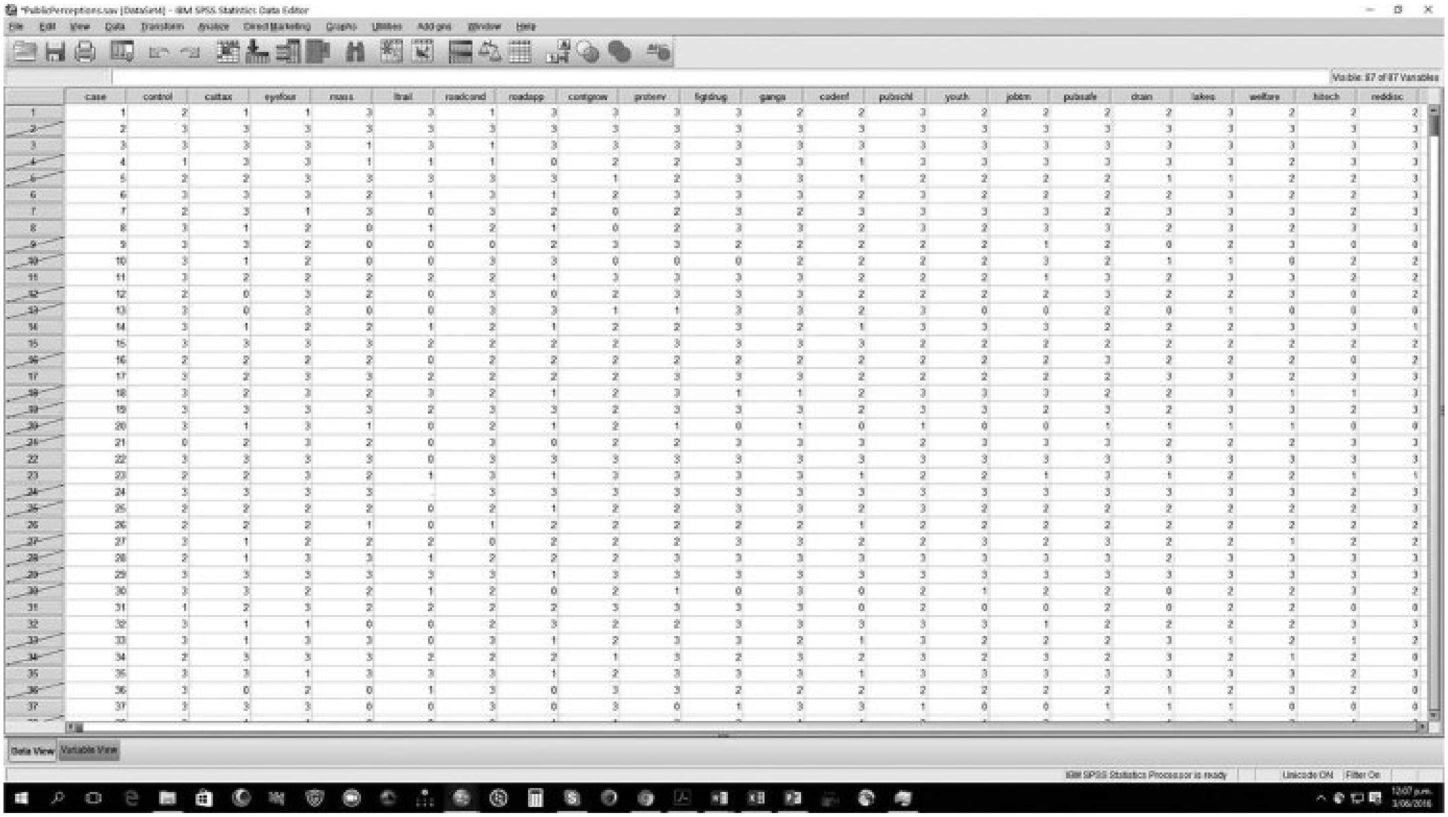
Output
● Filter out unsele...
○ Copy selected c...
○ Delete unselect...

Current Status: Do not filter cases

Select Cases: If

contact = 1

Screen W19.22 Selected Cases


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 87 of 87 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode ON
Filter On
12:07 p.m.
3/06/2016

	case	control	cuttax	eyefour	mass	ltrail	roadcond	roadapp	contgrow	protenv	fgtdrug	gangs	coderf	pubschl	youth	jobtrn	pubsafe	drain	lakes	welfare	hitech	reddisc
1	1	2	1	1	3	3	1	3	3	3	3	2	2	3	2	2	2	2	3	2	2	2
~~2~~	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	1	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
~~4~~	4	1	3	3	1	1	1	0	2	2	3	3	1	3	3	3	3	3	3	2	3	3
~~5~~	5	2	2	3	3	3	3	3	1	2	3	3	1	2	2	2	2	1	1	2	2	3
6	6	3	3	3	2	1	3	1	2	3	3	3	2	3	2	2	2	2	3	2	2	3
7	7	2	3	1	3	0	3	2	0	2	3	2	3	3	3	3	2	3	3	3	2	3
8	8	3	1	2	0	1	2	1	0	2	3	3	2	3	2	3	3	2	3	2	3	3
~~9~~	9	3	3	2	0	0	0	2	3	3	2	2	2	2	2	1	2	0	2	3	0	0
~~10~~	10	3	1	2	0	0	3	3	0	0	0	2	2	2	2	3	2	1	1	0	2	2
11	11	3	2	2	2	2	2	1	3	3	3	3	2	2	2	1	3	2	3	3	2	2
~~12~~	12	2	0	3	2	0	3	0	2	3	3	3	2	2	2	2	3	2	2	3	0	2
~~13~~	13	3	0	3	0	0	3	3	1	1	3	3	2	3	0	0	2	0	1	0	0	0
14	14	3	1	2	2	1	2	1	2	2	3	2	1	3	3	3	2	2	2	3	3	1
15	15	3	3	3	3	2	2	2	2	3	3	3	3	2	2	2	2	2	2	2	2	2
~~16~~	16	2	2	2	2	0	2	2	2	2	2	2	2	2	2	2	3	2	2	2	0	2
17	17	3	2	3	3	2	2	2	2	3	3	3	2	2	2	2	2	3	3	2	3	3
~~18~~	18	3	2	3	2	3	2	1	2	3	1	1	2	3	3	3	2	2	3	1	1	3
~~19~~	19	3	3	3	3	2	3	3	2	3	3	3	2	3	3	2	3	2	3	3	2	3
~~20~~	20	3	1	3	1	0	2	1	2	1	0	1	0	1	0	0	1	1	1	1	0	0
~~21~~	21	0	2	3	2	0	3	0	2	2	3	3	3	2	3	3	3	2	2	2	3	3
22	22	3	3	3	3	0	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
23	23	2	2	3	2	1	3	1	3	3	3	3	1	2	2	1	3	1	2	2	1	1
~~24~~	24	3	3	3	3	.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	3
~~25~~	25	2	2	2	2	0	2	1	2	2	3	3	2	3	2	2	2	2	2	2	2	3
26	26	2	2	2	1	0	1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2
~~27~~	27	3	1	2	2	2	0	2	2	2	3	3	2	2	3	2	3	2	2	1	2	2
~~28~~	28	2	1	3	3	1	2	2	2	3	3	3	3	3	3	3	3	2	3	3	3	3
~~29~~	29	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3
~~30~~	30	3	3	2	2	1	2	0	2	1	0	3	0	2	1	2	2	0	2	2	3	2
31	31	1	2	3	2	2	2	2	3	3	3	3	0	2	0	0	2	0	2	2	0	0
32	32	3	1	1	0	0	2	3	2	2	3	3	3	3	3	1	2	2	2	2	3	3
~~33~~	33	3	1	3	3	0	3	1	2	3	3	2	1	3	2	2	2	3	1	2	1	2
~~34~~	34	2	3	3	3	2	2	2	1	3	2	3	2	3	2	3	2	3	2	1	2	0
35	35	3	3	1	3	3	3	1	2	3	3	3	1	3	3	3	3	3	3	3	2	3
~~36~~	36	3	0	2	0	1	3	0	3	3	2	2	2	2	2	2	2	1	2	3	2	0
37	37	3	3	3	0	0	3	0	3	0	1	3	3	1	0	0	1	1	1	0	0	0

Index Variables I: Cronbach Alpha

Chapter 3 of the textbook discusses the use of index variables to measure concepts comprising several dimensions. In short, different variables are used to measure different dimensions, and these variables are subsequently aggregated into an index variable.

Remember, index variable construction is a two-step process. Before aggregating the data, you need to *justify* (that is, to persuasively argue) that the disparate variables indeed measure different dimensions of the same underlying concept (see Chapter 3 of the textbook). The justification should be based on theoretical grounds (do the variables make sense as dimensions of the same underlying concept?) and empirical grounds. One way to empirically justify the selection of variables is to examine the extent to which the disparate variables are correlated with each other. If the variables measure dimensions of the *same* concept, then the dimensions should be correlated with each other as well.

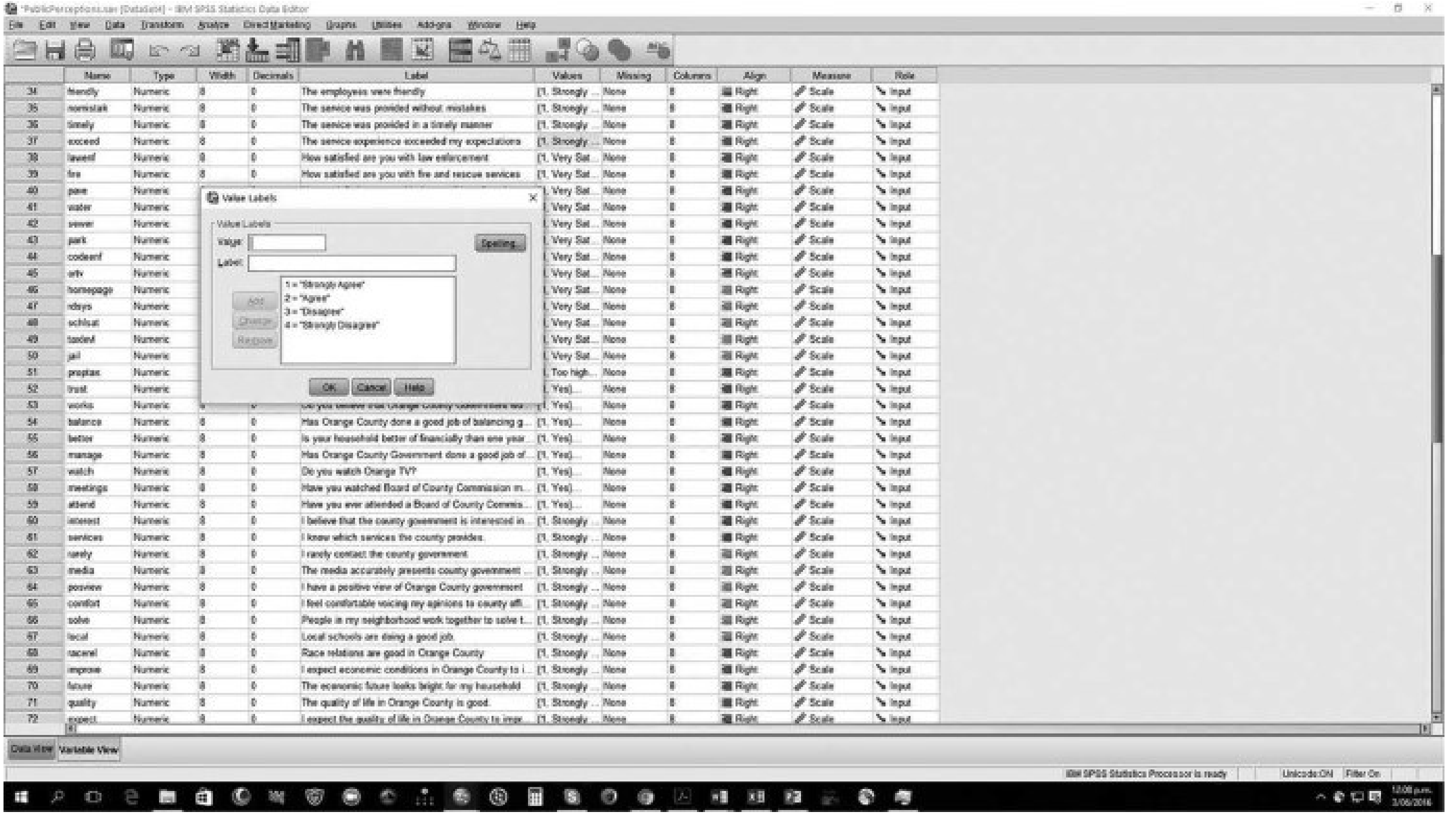
Cronbach alpha is the measure that assesses the empirical correlation (or internal reliability) of variables. As discussed in Chapter 3, values between 0.80 and 1.00 indicate high internal reliability, and values between 0.70 and 0.80 indicate moderate but acceptable internal reliability.

Open the Public Perceptions dataset. This dataset contains several variables (survey items) that can be used to measure customer satisfaction. Specifically, go to the Variable View screen. Scroll down to the first of these variables, helpful (on or about row 32). Double click on the Label cell; this allows you to scroll the entire label entry ("the employees were helpful"). The other variables—respect, friendly, nomistak, timely, and exceed—are immediately below that variable. These individual variables appear to be possible dimensions of an overall concept, for example, "customer satisfaction." Screen W19.23 also shows the value labels for these items (rows 32 through 37).

Next, examine the Cronbach alpha statistic to determine whether these items are adequately correlated with each other. But first, practice what you learned in the previous section; you wish to create the index variable only for respondents who have had contact with county officials in the past year. Select only those observations that meet this condition, that is, "contact = 1." This selection produces a screen similar to Screen W19.21.

The command sequence for producing the Cronbach alpha statistic is Analyze → Scale → Reliability Analysis. Add all six variables to the Items box. Then, select Statistics and, from the "Descriptives for" box, the items "Item," "Scale," "Scale if item deleted," as shown in Screen W19.24. Then, to perform the analysis, select Continue and OK. Screen W19.25 shows the results. Note that the value for alpha is .882. This value is well within the required range of 0.8 to 1.0, indicating a very high degree of *internal consistency* (or *internal reliability*). Also note that deleting any item (variable), as shown in the right-hand column, results in a lower value for alpha. Therefore, combining these separate dimension variables into a concept index variable, "customer satisfaction," is empirically justified. Finally, note that the number of observations is now limited to 353, reflecting both the constraint (contact = 1) and missing values for some of the six variables. (If the analysis had not been limited to only those respondents who have had contact, the result would have been *n* = 615 and alpha = .8855, resulting in the same conclusion.)

Screen W19.23 Value Labels for Public Perceptions Dataset


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON Filter On

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
34	friendly	Numeric	8	0	The employees were friendly	{1, Strongly ...	None	8	Right	Scale	Input
35	nomistak	Numeric	8	0	The service was provided without mistakes	{1, Strongly ...	None	8	Right	Scale	Input
36	timely	Numeric	8	0	The service was provided in a timely manner	{1, Strongly ...	None	8	Right	Scale	Input
37	exceed	Numeric	8	0	The service experience exceeded my expectations	{1, Strongly ...	None	8	Right	Scale	Input
38	lawenf	Numeric	8	0	How satisfied are you with law enforcement	{1, Very Sat...	None	8	Right	Scale	Input
39	fire	Numeric	8	0	How satisfied are you with fire and rescue services	{1, Very Sat...	None	8	Right	Scale	Input
40	pave	Numeric				Very Sat...	None	8	Right	Scale	Input
41	water	Numeric				Very Sat...	None	8	Right	Scale	Input
42	sewer	Numeric				Very Sat...	None	8	Right	Scale	Input
43	park	Numeric				Very Sat...	None	8	Right	Scale	Input
44	codeenf	Numeric				Very Sat...	None	8	Right	Scale	Input
45	artv	Numeric				Very Sat...	None	8	Right	Scale	Input
46	homepage	Numeric				Very Sat...	None	8	Right	Scale	Input
47	rdsys	Numeric				Very Sat...	None	8	Right	Scale	Input
48	schlsat	Numeric				Very Sat...	None	8	Right	Scale	Input
49	taxdevl	Numeric				Very Sat...	None	8	Right	Scale	Input
50	jail	Numeric				Very Sat...	None	8	Right	Scale	Input
51	proptax	Numeric				Too high...	None	8	Right	Scale	Input
52	trust	Numeric				Yes}...	None	8	Right	Scale	Input
53	works	Numeric	8	0	Do you believe that Orange County Government wo...	{1, Yes}...	None	8	Right	Scale	Input
54	balance	Numeric	8	0	Has Orange County done a good job of balancing g...	{1, Yes}...	None	8	Right	Scale	Input
55	better	Numeric	8	0	Is your household better of financially than one year ...	{1, Yes}...	None	8	Right	Scale	Input
56	manage	Numeric	8	0	Has Orange County Government done a good job of ...	{1, Yes}...	None	8	Right	Scale	Input
57	watch	Numeric	8	0	Do you watch Orange TV?	{1, Yes}...	None	8	Right	Scale	Input
58	meetings	Numeric	8	0	Have you watched Board of County Commission m...	{1, Yes}...	None	8	Right	Scale	Input
59	attend	Numeric	8	0	Have you ever attended a Board of County Commis...	{1, Yes}...	None	8	Right	Scale	Input
60	interest	Numeric	8	0	I believe that the county government is interested in ...	{1, Strongly ...	None	8	Right	Scale	Input
61	services	Numeric	8	0	I know which services the county provides.	{1, Strongly ...	None	8	Right	Scale	Input
62	rarely	Numeric	8	0	I rarely contact the county government	{1, Strongly ...	None	8	Right	Scale	Input
63	media	Numeric	8	0	The media accurately presents county government ...	{1, Strongly ...	None	8	Right	Scale	Input
64	posview	Numeric	8	0	I have a positive view of Orange County government	{1, Strongly ...	None	8	Right	Scale	Input
65	comfort	Numeric	8	0	I feel comfortable voicing my opinions to county offi...	{1, Strongly ...	None	8	Right	Scale	Input
66	solve	Numeric	8	0	People in my neighborhood work together to solve t...	{1, Strongly ...	None	8	Right	Scale	Input
67	local	Numeric	8	0	Local schools are doing a good job.	{1, Strongly ...	None	8	Right	Scale	Input
68	racerel	Numeric	8	0	Race relations are good in Orange County	{1, Strongly ...	None	8	Right	Scale	Input
69	improve	Numeric	8	0	I expect economic conditions in Orange County to i...	{1, Strongly ...	None	8	Right	Scale	Input
70	future	Numeric	8	0	The economic future looks bright for my household	{1, Strongly ...	None	8	Right	Scale	Input
71	quality	Numeric	8	0	The quality of life in Orange County is good.	{1, Strongly ...	None	8	Right	Scale	Input
72	expect	Numeric	8	0	I expect the quality of life in Orange County to impr...	{1, Strongly ...	None	8	Right	Scale	Input

Value Labels

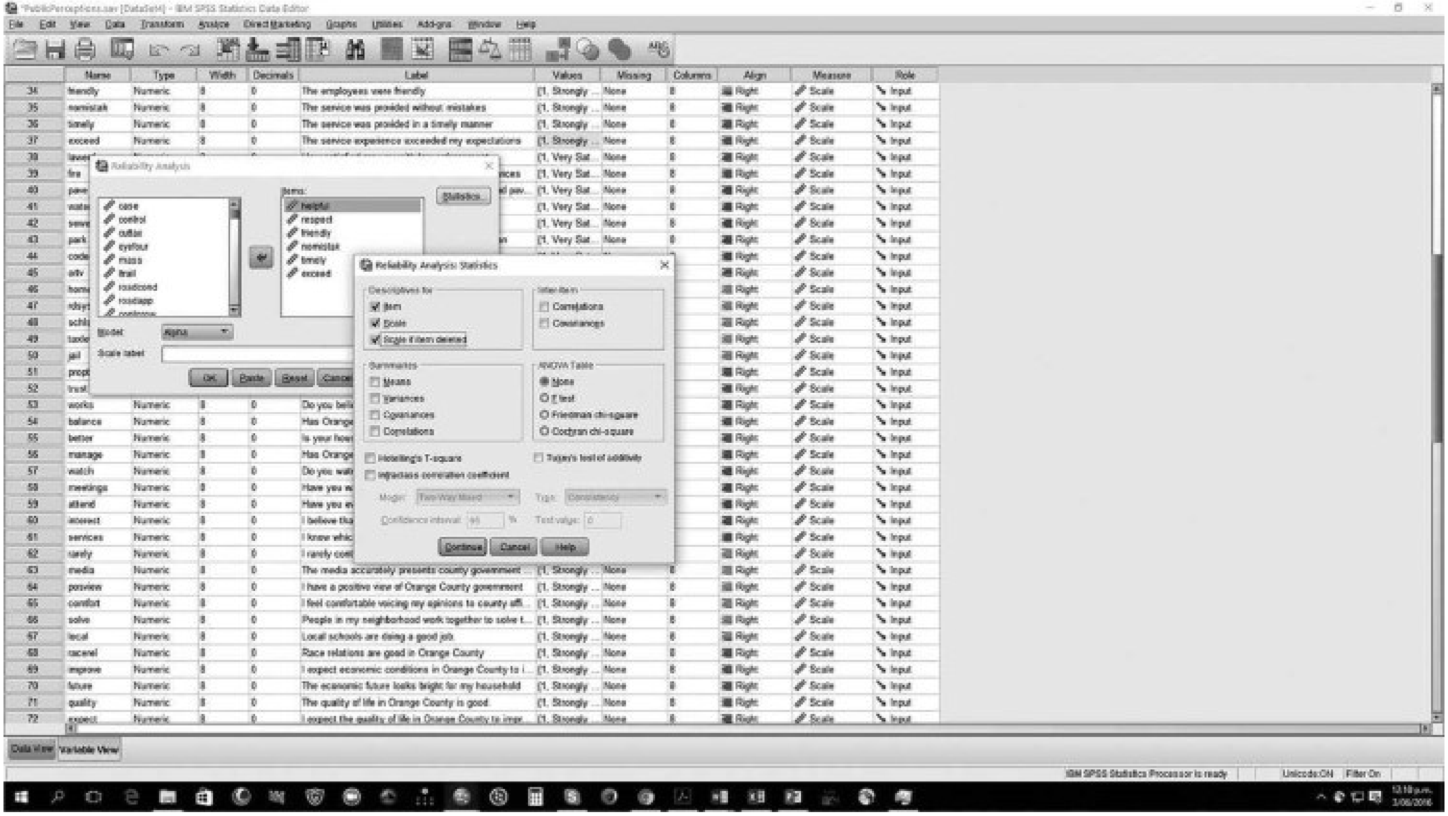
Value Labels

Value:

Label:

1 = "Strongly Agree"
2 = "Agree"
3 = "Disagree"
4 = "Strongly Disagree"

Screen W19.24 Producing Cronbach Alpha


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON Filter On

Reliability Analysis

case
control
outfair
eyefour
mass
trail
roadcond
roadapp
conforow

Items:
helpful
respect
friendly
nomistak
timely
exceed

Model: Alpha

Scale label:

Statistics...
OK
Paste
Reset
Cancel

Reliability Analysis: Statistics

Descriptives for
☑ Item
☑ Scale
☑ Scale if item deleted

Inter-Item
□ Correlations
□ Covariances

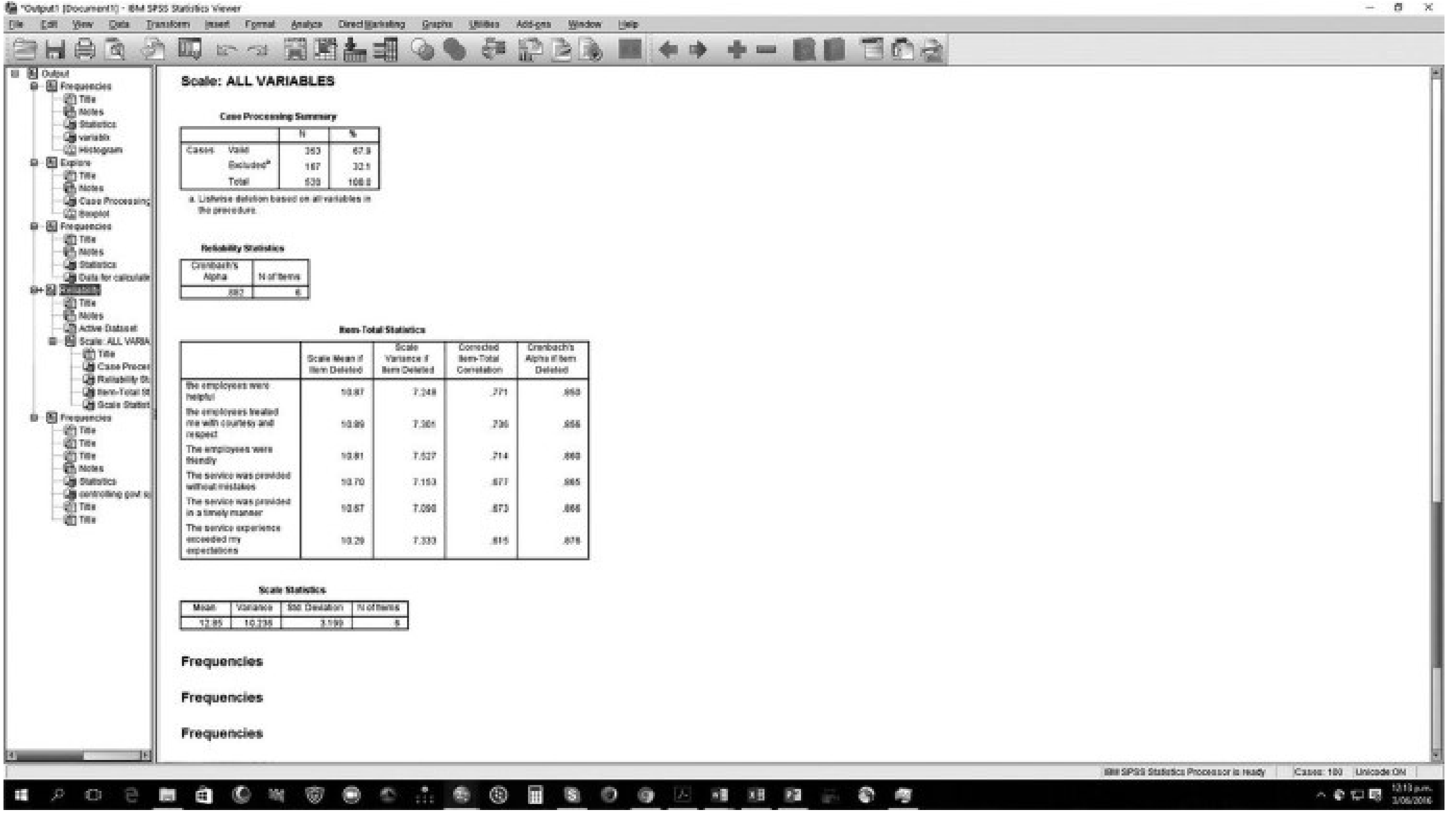
Summaries
□ Means
□ Variances
□ Covariances
□ Correlations

ANOVA Table
● None
○ F test
○ Friedman chi-square
○ Cochran chi-square

□ Hotelling's T-square
□ Intraclass correlation coefficient
□ Tukey's test of additivity

Model: Two-Way Mixed
Type: Consistency
Confidence interval: 95 %
Test value: 0

Screen W19.25 Cronbach Alpha Output



Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	353	67.9
	Excluded^a	167	32.1
	Total	520	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.882	6

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
the employees were helpful	10.87	7.248	.771	.858
the employees treated me with courtesy and respect	10.89	7.301	.736	.856
The employees were friendly	10.81	7.527	.714	.868
The service was provided without mistakes	10.70	7.153	.677	.865
The service was provided in a timely manner	10.67	7.090	.673	.866
The service experience exceeded my expectations	10.29	7.333	.615	.876

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
12.85	10.238	3.199	6

Frequencies

Frequencies

Frequencies

Index Variables II: Construction

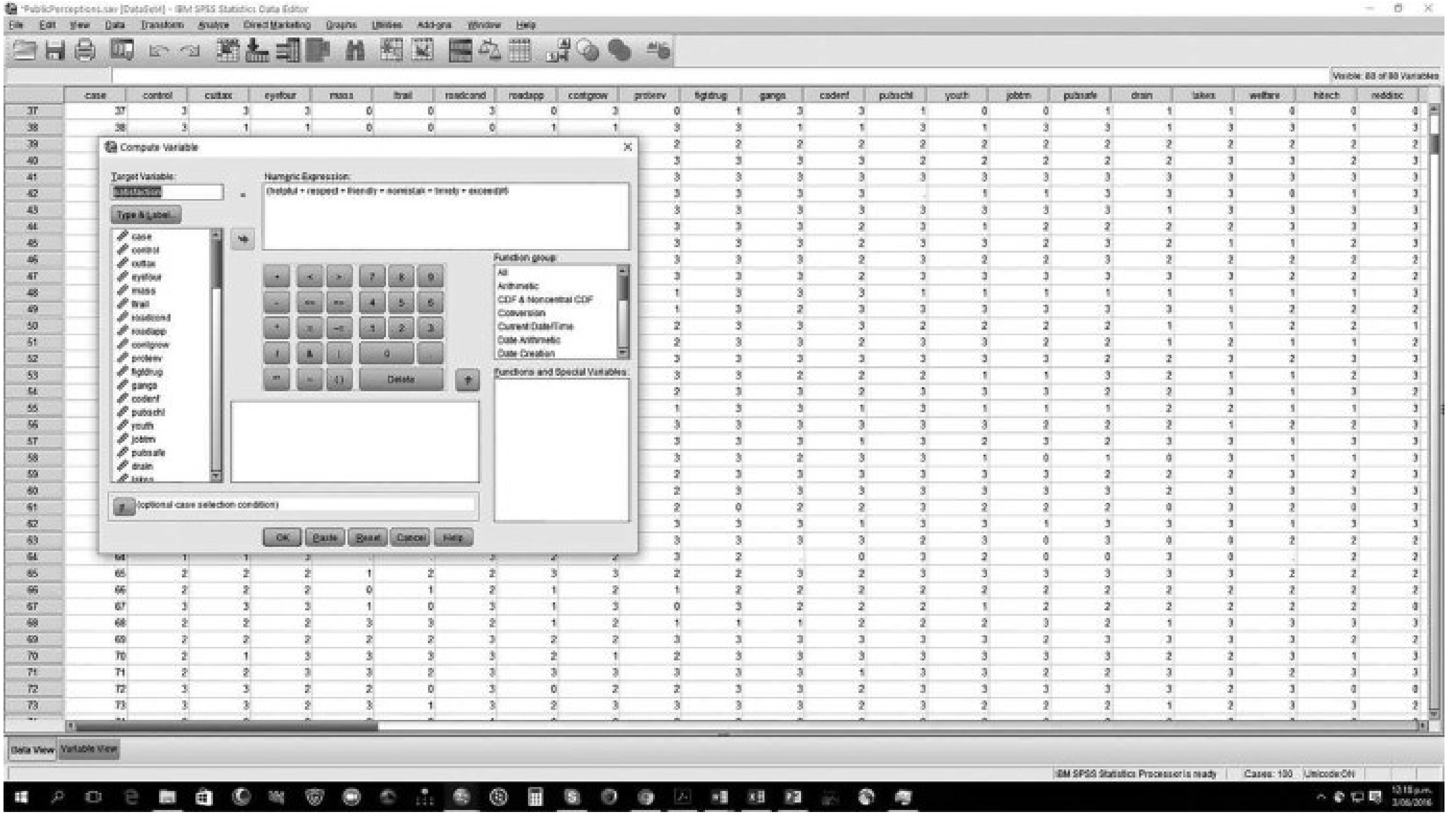
Now you will construct the index variable. From the toolbar, select Transform → Compute. The Compute Variable dialogue box will appear (Screen W19.26). In the Target Variable box, enter the name of your new index variable for customer satisfaction. In the Numeric Expression box, enter the six variables comprising your index variable, and divide by six. Be sure to note the brackets in Screen W19.26 to ensure the appropriate division. Then select OK. Examine the Data View screen; satisfaction is now the last variable. Examine the new variable using Analyze → Descriptive Statistics → Frequencies, which yields the results shown in Screen W19.27 (among those who have had contact, i.e., Data→ Select Cases →If contact = 1. If you analyze the data or all respondents, you will have n = 615 valid responses in your analysis, rather than n = 353 as shown). Also, you can produce summary statistics by selecting Frequencies → Statistics; choose Mean, Median, and Mode from the Central Tendency box, and then select Continue → OK.

Recoding Data

Depending on a variable’s data distribution, there may be so many response categories and resulting low frequency counts that recoding the variable not only is appropriate but also will allow for more meaningful data interpretation. Redefining data into new categories is a relatively simple process.

Open the Public Perceptions dataset. Perform a frequency distribution and histogram of the variable “ethnic.” The output shown in Screen W19.28 and Figure W19.2 is produced. (Be sure to select all cases for this analysis.) You will see that ethnic has several response categories with low counts. Asian/Pacific Islander (category 4), Native American (5), and Other (6) could be reasonably redefined as an aggregate variable “Other.” Doing so would yield a total of four categories within the variable ethnic. Specifically, the new variable, “ethrecod,” will have the following categories: 1 = White, 2 = Black/African-American, 3 = Hispanic, and 4 = Other.

Screen W19.26 Constructing the Index Variable for “Customer Satisfaction”

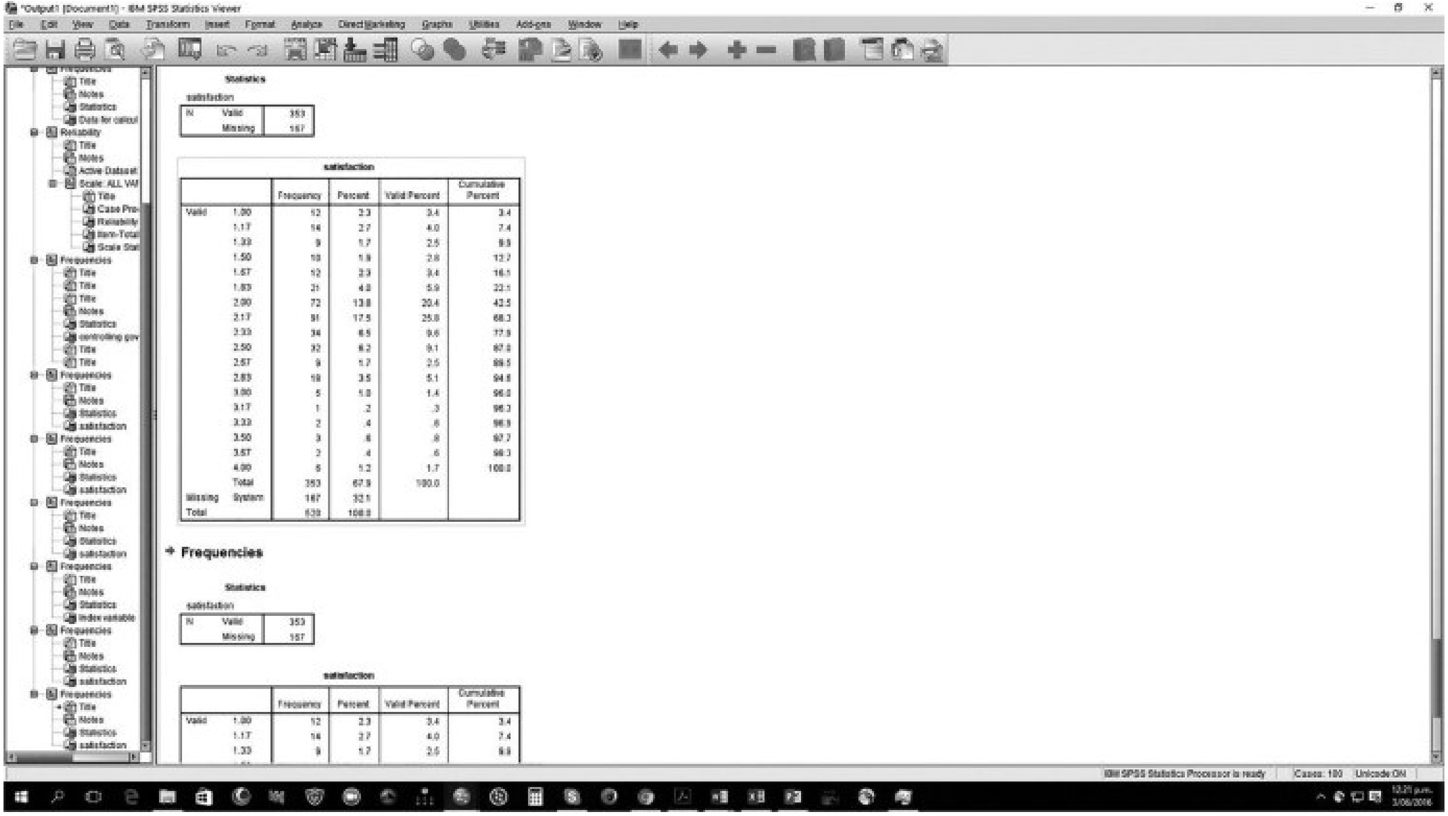

*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Compute Variable
Type & Label...
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Delete
If... (optional case selection condition)
OK Paste Reset Cancel Help
Data View Variable View

Target Variable: Satisfaction

Numeric Expression: (helpful + respect + friendly + nonmistak + timely + exceed)/6

To perform recoding, you will recode and create a new variable, ethrecod, in order to preserve the coding of the existing variable. Select Transform → Recode → Into Different Variables. Then select the existing variable, ethnic, which results in the dialogue box shown in Screen W19.29. Note that the dialogue box is now prompting the user for the name of the new variable (note the question mark in the dialogue box). Type “ethrecod” in the Name box, shown in Screen W19.30. Next, select Change (in the far right of the dialogue box) to complete this part of the variable recoding, which should produce Screen W19.31.

Screen W19.27 Output Screen for Satisfyd


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode:ON

Statistics

satisfaction

N	Valid	353
	Missing	167

satisfaction

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	12	2.3	3.4	3.4
	1.17	14	2.7	4.0	7.4
	1.33	9	1.7	2.5	9.9
	1.50	10	1.9	2.8	12.7
	1.67	12	2.3	3.4	16.1
	1.83	21	4.0	5.9	22.1
	2.00	72	13.8	20.4	42.5
	2.17	91	17.5	25.8	68.3
	2.33	34	6.5	9.6	77.9
	2.50	32	6.2	9.1	87.0
	2.67	9	1.7	2.5	89.5
	2.83	18	3.5	5.1	94.6
	3.00	5	1.0	1.4	96.0
	3.17	1	.2	.3	96.3
	3.33	2	.4	.6	96.9
	3.50	3	.6	.8	97.7
	3.67	2	.4	.6	98.3
	4.00	6	1.2	1.7	100.0
	Total	353	67.9	100.0	
Missing	System	167	32.1		
Total		520	100.0		

Frequencies

Statistics

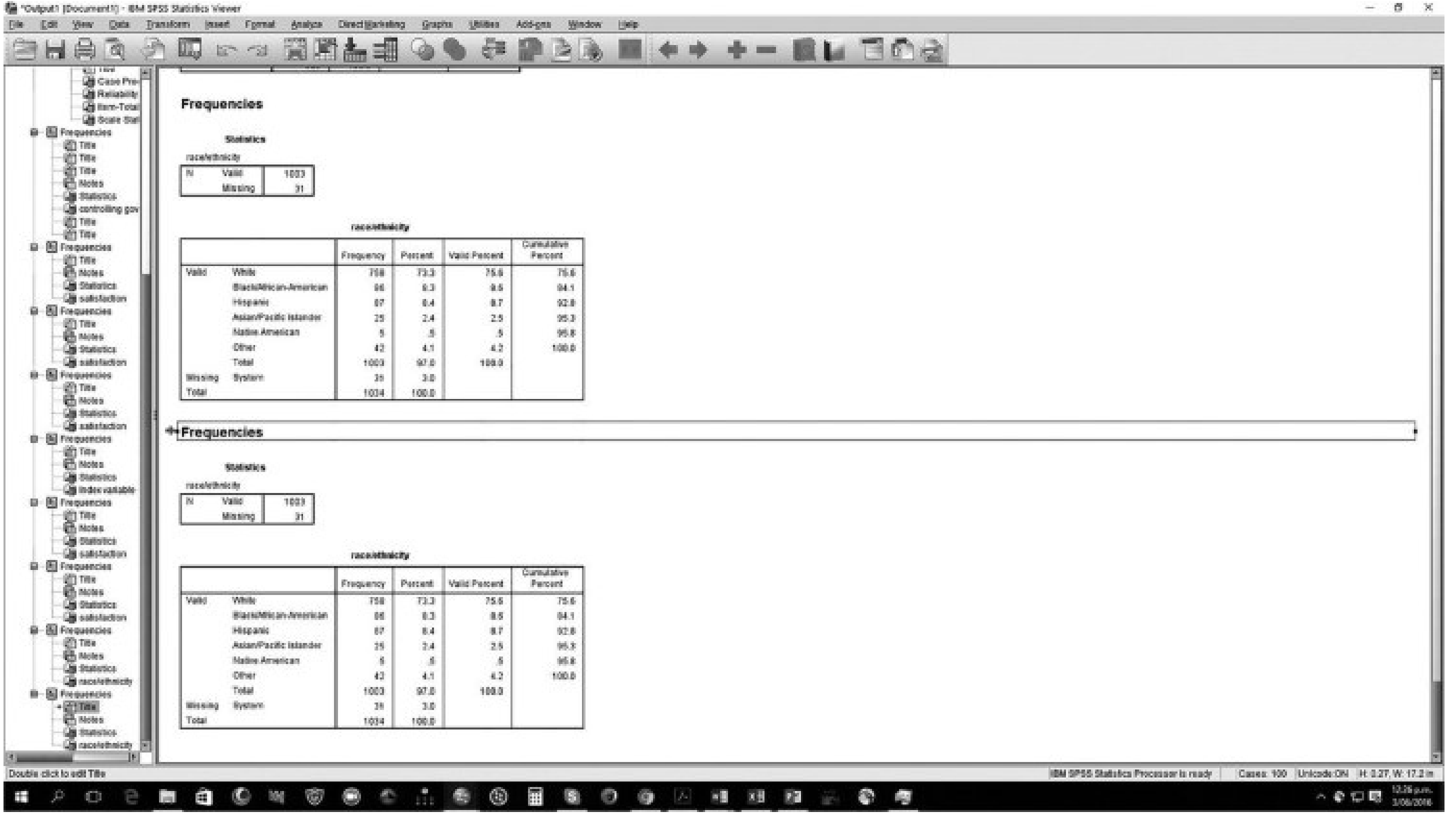
satisfaction

N	Valid	353
	Missing	167

satisfaction

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	12	2.3	3.4	3.4
	1.17	14	2.7	4.0	7.4
	1.33	9	1.7	2.5	9.9

Screen W19.28 Output Screen for Ethnic


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Double click to edit Title
IBM SPSS Statistics Processor is ready | Cases: 100 | Unicode:ON | H: 0.27, W: 17.2 in

Frequencies

Statistics

race/ethnicity

N	Valid	1003
	Missing	31

race/ethnicity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	758	73.3	75.6	75.6
	Black/African-American	86	8.3	8.6	84.1
	Hispanic	87	8.4	8.7	92.8
	Asian/Pacific Islander	25	2.4	2.5	95.3
	Native American	5	.5	.5	95.8
	Other	42	4.1	4.2	100.0
	Total	1003	97.0	100.0	
Missing	System	31	3.0		
Total		1034	100.0		

Frequencies

Statistics

race/ethnicity

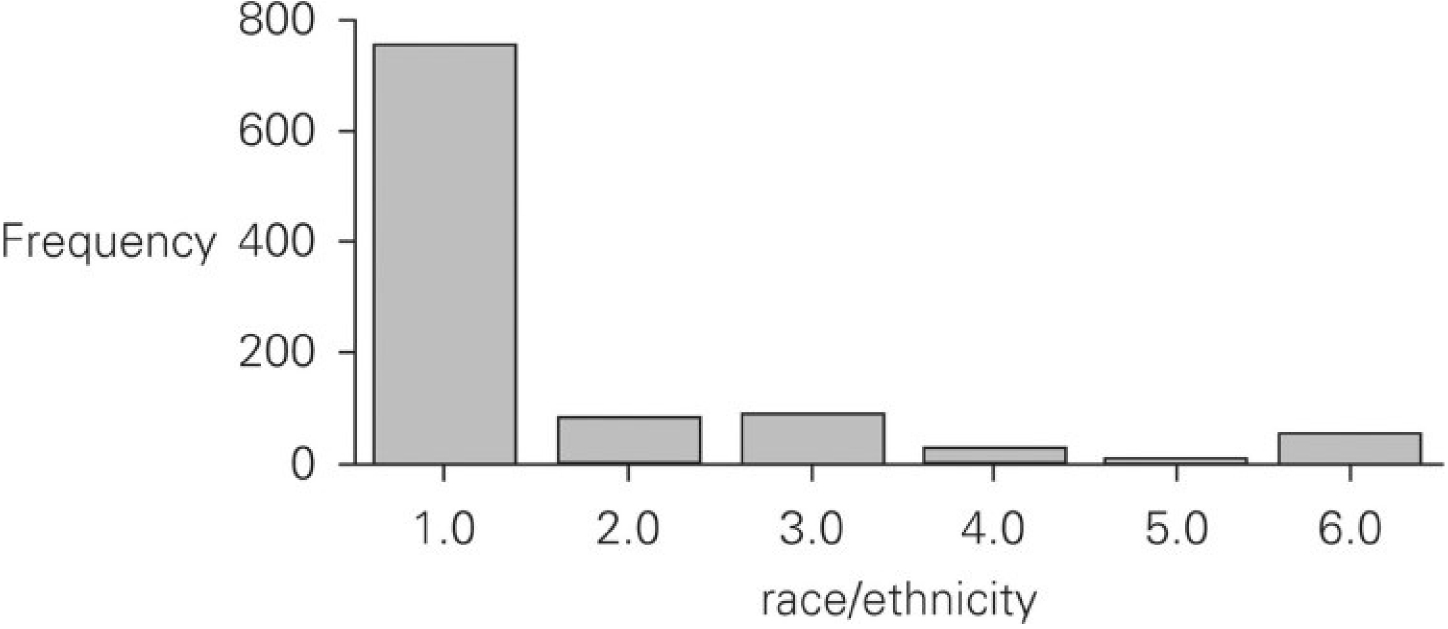
N	Valid	1003
	Missing	31

race/ethnicity

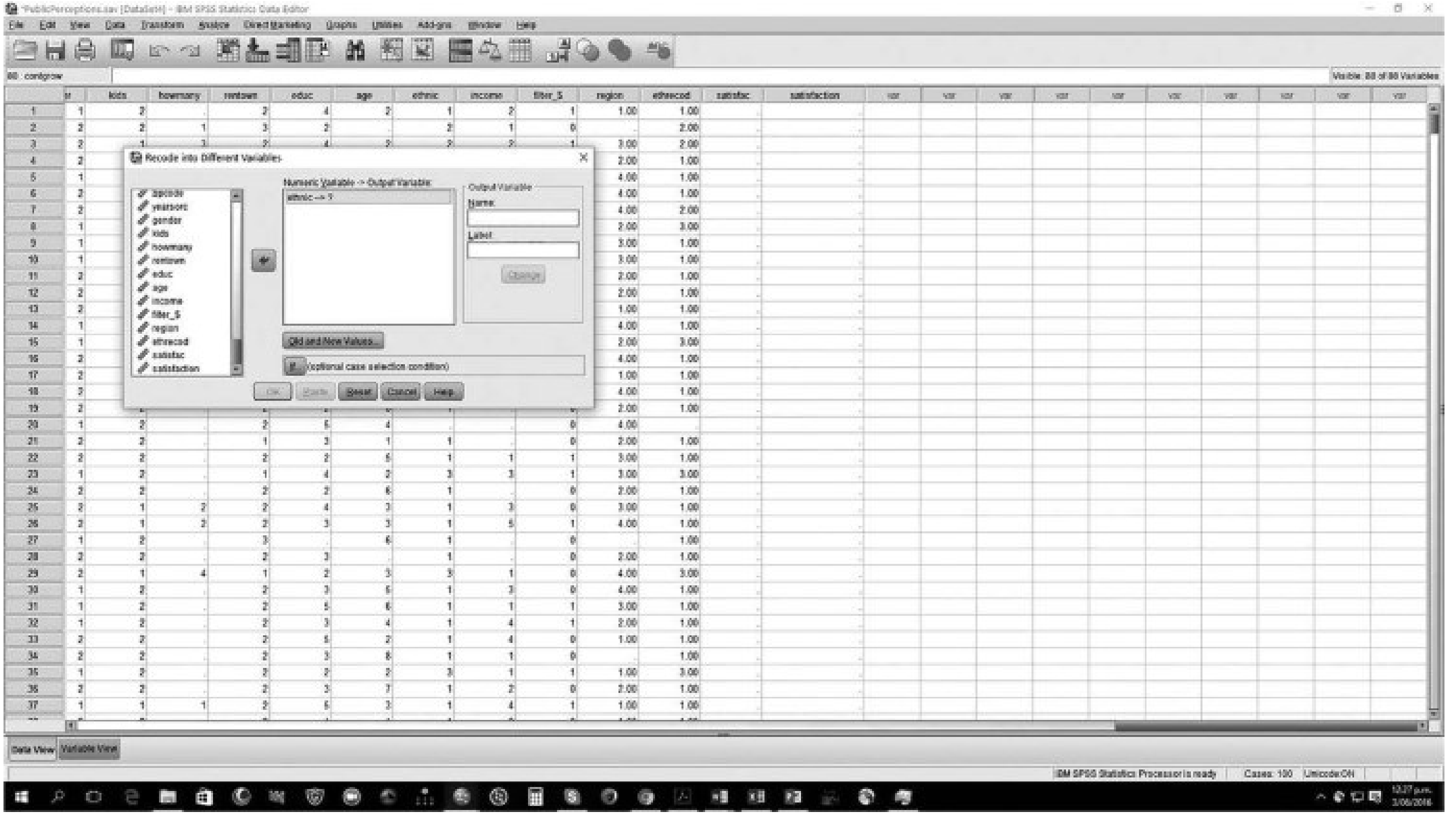
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	758	73.3	75.6	75.6
	Black/African-American	86	8.3	8.6	84.1
	Hispanic	87	8.4	8.7	92.8
	Asian/Pacific Islander	25	2.4	2.5	95.3
	Native American	5	.5	.5	95.8
	Other	42	4.1	4.2	100.0
	Total	1003	97.0	100.0	
Missing	System	31	3.0		
Total		1034	100.0		

Now that you have defined the new variable, you must define the new categories. Select Old and New Values, which brings up the dialogue box shown in Screen W19.32. The original values for response categories 1 should be retained. Enter “1” for both Old Value and New Value, and then Add. In turn, Screens W19.33 and W19.34 are produced.

Figure W19.2 Race/Ethnicity Output



Screen W19.29 Recoding Values for Ethnic


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Recode into Different Variables

Numeric Variable -> Output Variable:
ethnic --> ?

Output Variable
Name:
Label:
Change

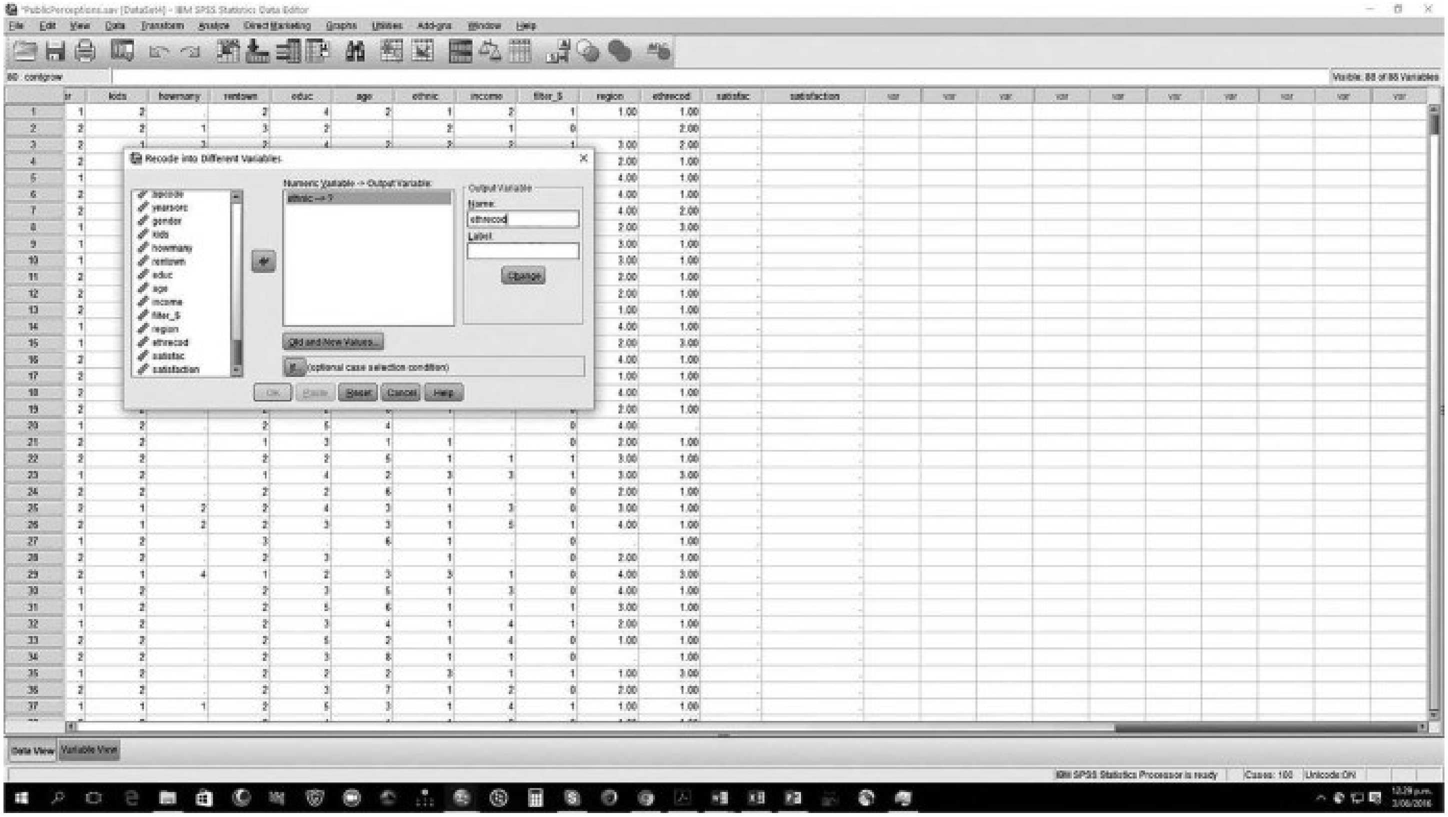
zipcode
yearsorc
gender
kids
howmany
rentown
educ
age
income
filter_$
region
ethrecod
satisfac
satisfaction

Old and New Values...

If... (optional case selection condition)

OK Paste Reset Cancel Help

Screen W19.30 Naming the Variable


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 88 of 88 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON

Recode into Different Variables

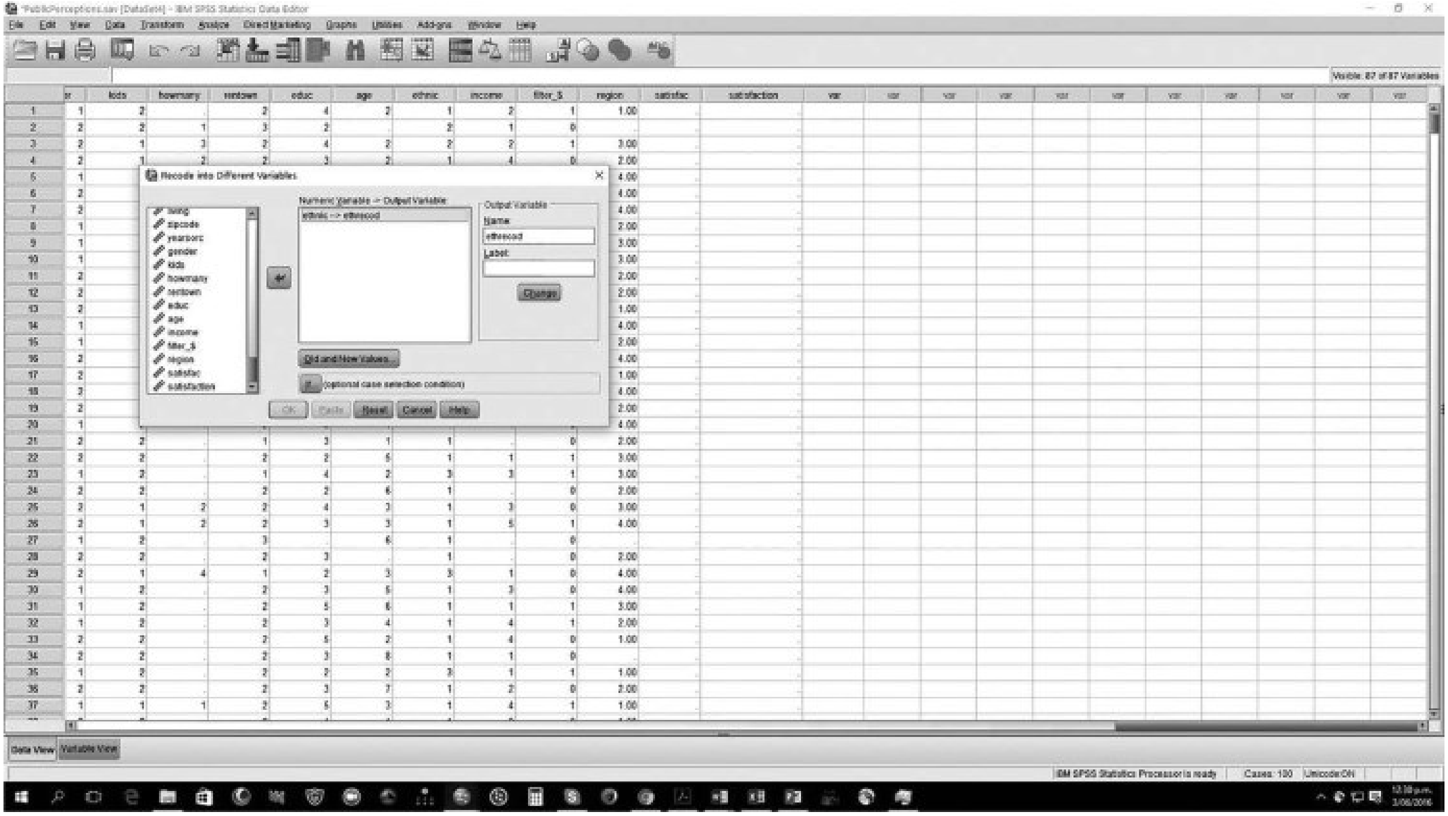
zipcode, yearsorc, gender, kids, howmany, rentown, educ, age, income, filter_$, region, ethrecod, satisfac, satisfaction

Numeric Variable -> Output Variable:
ethnic --> ?

Output Variable
Name: ethrecod
Label:

Change
Old and New Values...
If... (optional case selection condition)
OK
Paste
Reset
Cancel
Help

Screen W19.31 Recoding Dialogue Box


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 87 of 87 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON

Recode into Different Variables

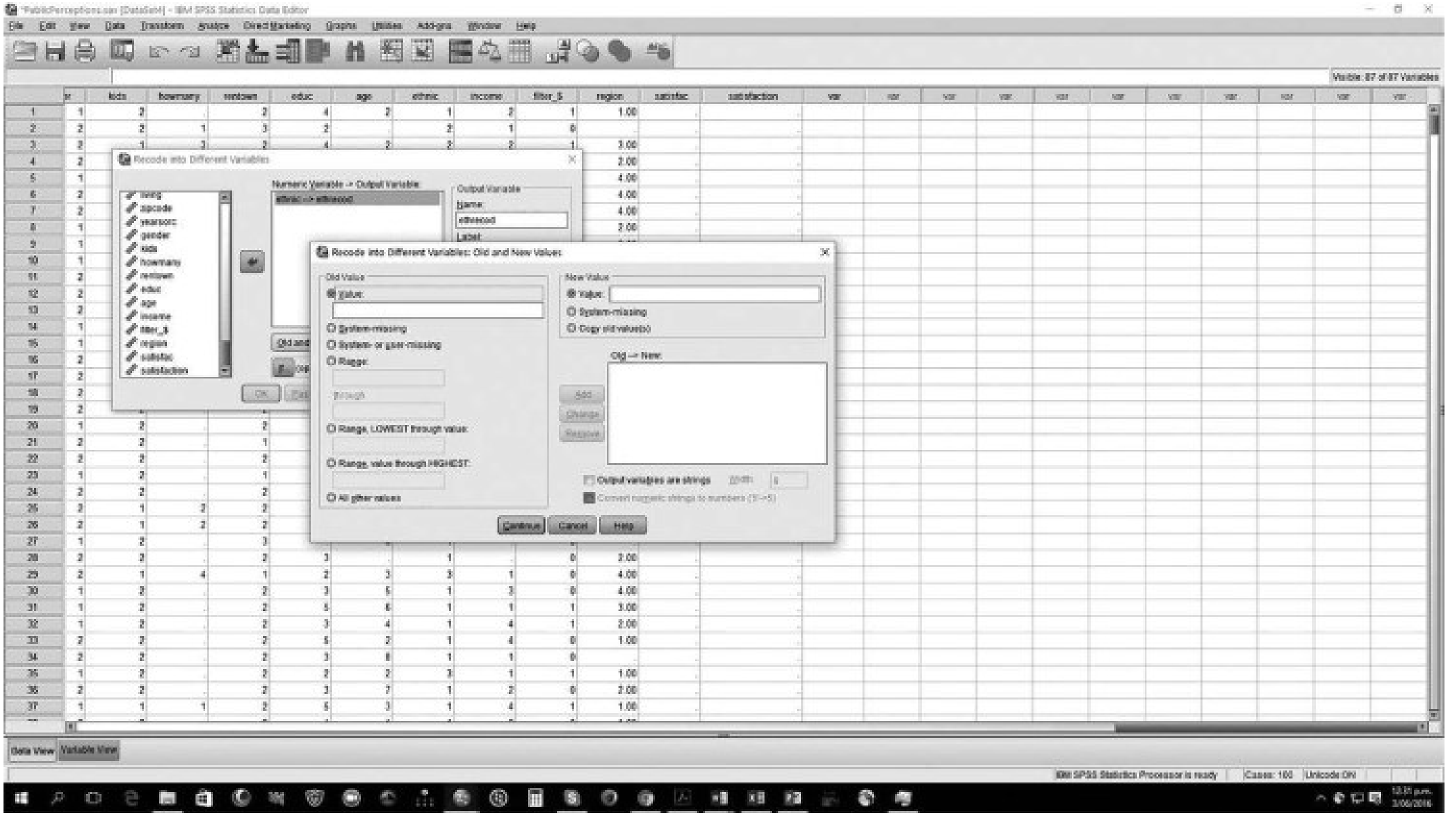
living, zipcode, yearsorc, gender, kids, howmany, rentown, educ, age, income, filter_$, region, satisfac, satisfaction

Numeric Variable -> Output Variable:
ethnic --> ethrecod

Output Variable
Name: ethrecod
Label:

Change
Old and New Values...
If... (optional case selection condition)
OK
Paste
Reset
Cancel
Help

Screen W19.32 Old and New Values

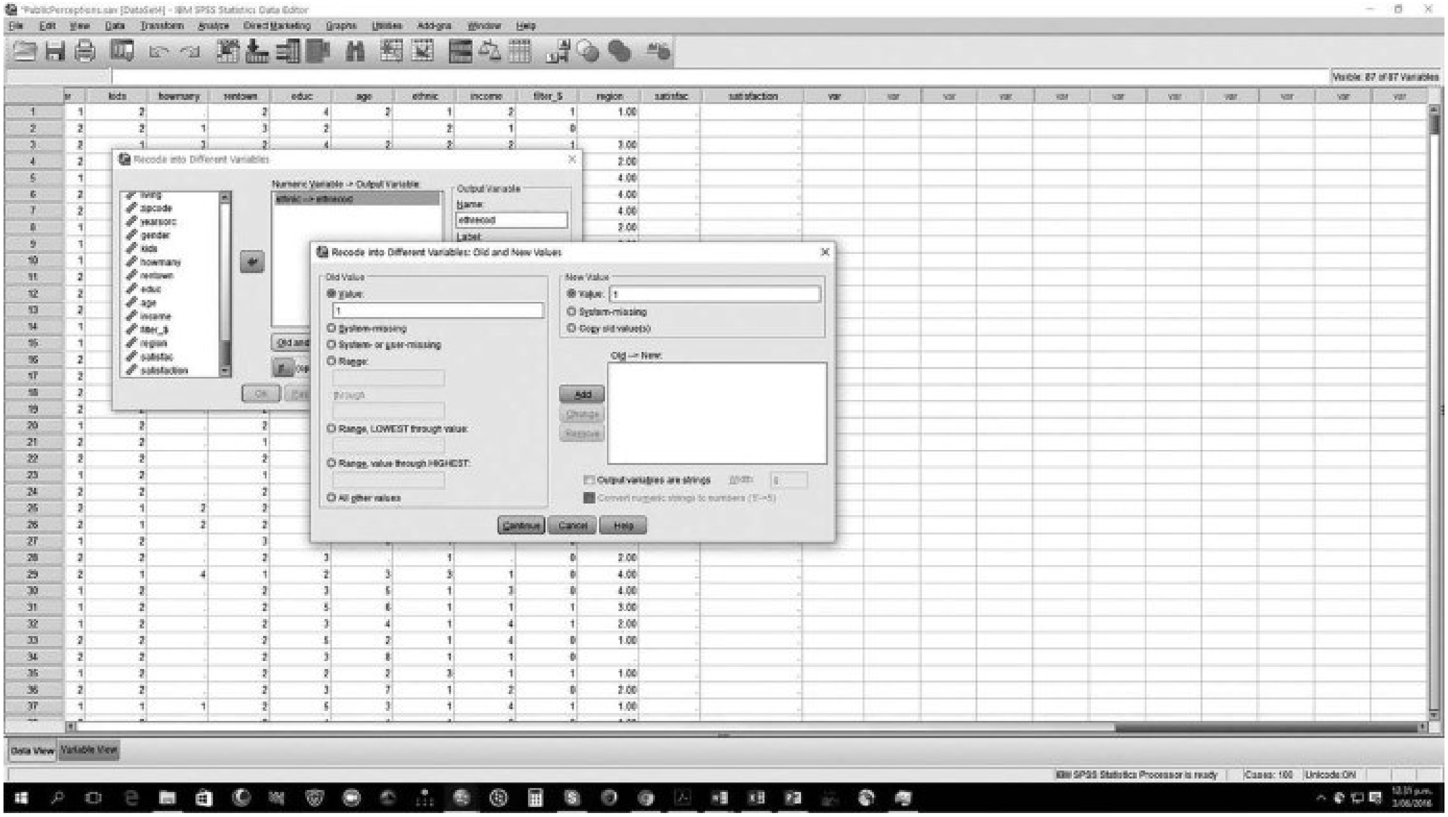

*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 87 of 87 Variables
Recode into Different Variables
Numeric Variable -> Output Variable:
Output Variable
Name:
Label:
Old and New Values...
OK
Recode into Different Variables: Old and New Values
Old Value
Value:
System-missing
System- or user-missing
Range:
through
Range, LOWEST through value:
Range, value through HIGHEST:
All other values
New Value
Value:
System-missing
Copy old value(s)
Old --> New:
Add
Change
Remove
Output variables are strings
Width:
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON

Numeric Variable -> Output Variable: ethnic --> ethnecod

Name: ethnecod

Continue this procedure for response categories 2 and 3. Now recode response categories 4, 5, and 6 into a single new category, 4. Select Range, and enter “4” and “6.” Note that the New Value cell is now illuminated; enter “4” → Add. To ensure that missing data from the original variable are retained and recognized as such by SPSS, select “System- or user-missing” under Old Value, and “System-missing” under New Value → Add. Screen W19.35 is displayed. As an important side note, observe the two cells “System- or user-missing” and “System-missing.” System-missing cells are empty, whereas user-missing cells have user-defined missing values, as discussed earlier. It is a safe practice to specify System- or user-missing when the intent is to refer to all missing values. In this case, the Public Perceptions dataset does not contain user-defined missing values (verify this in the Variable View screen, Missing column); thus, using System-missing and System- or user-missing produces the same result.

Screen W19.33 Entering New Value


Recode into Different Variables

Numeric Variable -> Output Variable: ethnic --> ethrecod

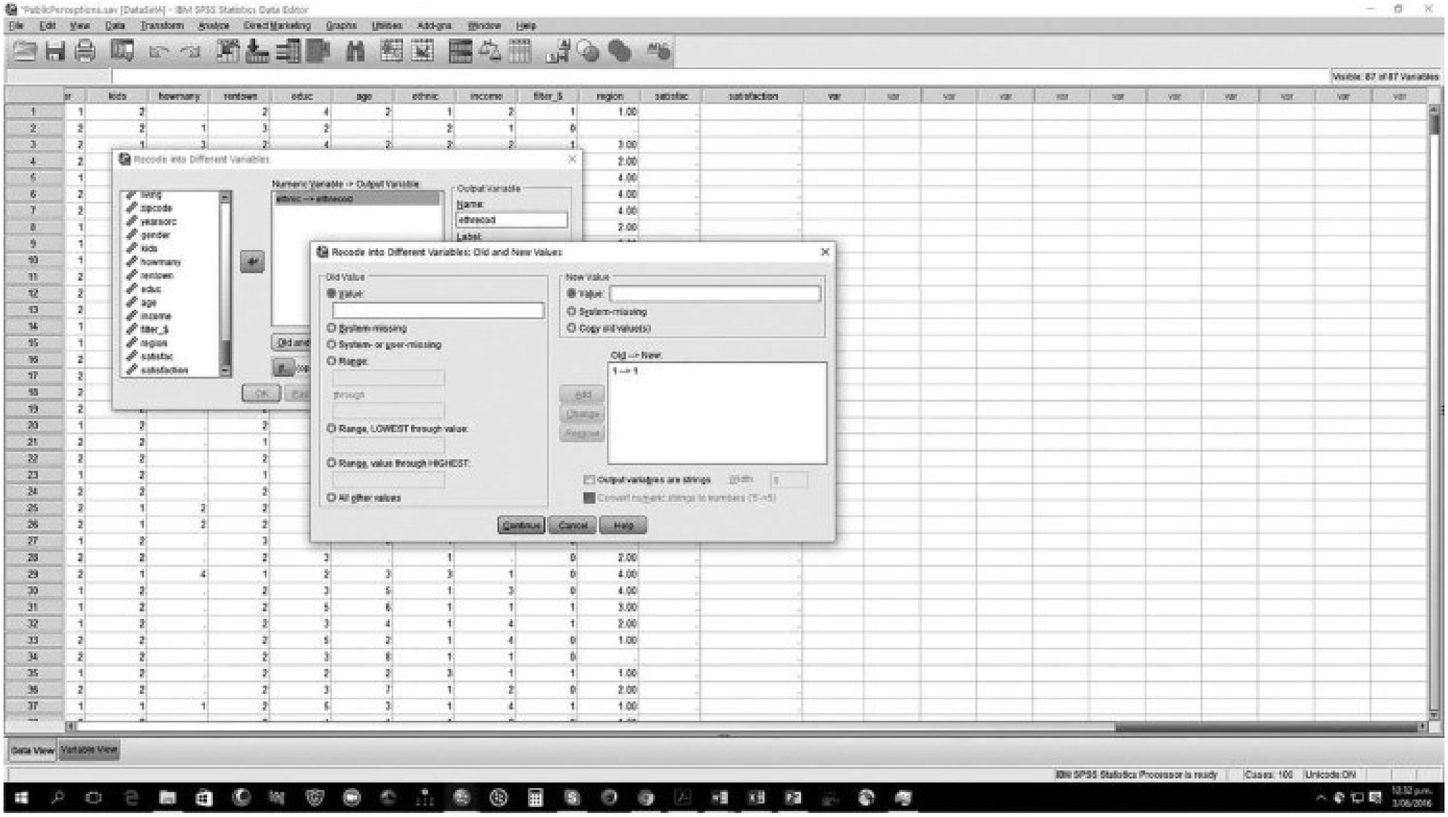
Output Variable Name: ethrecod

Recode into Different Variables: Old and New Values

Old Value: ● Value: 1

New Value: ● Value: 1

Screen W19.34 Old to New Values


Recode into Different Variables

Numeric Variable -> Output Variable: ethnic --> ethrecod

Output Variable Name: ethrecod

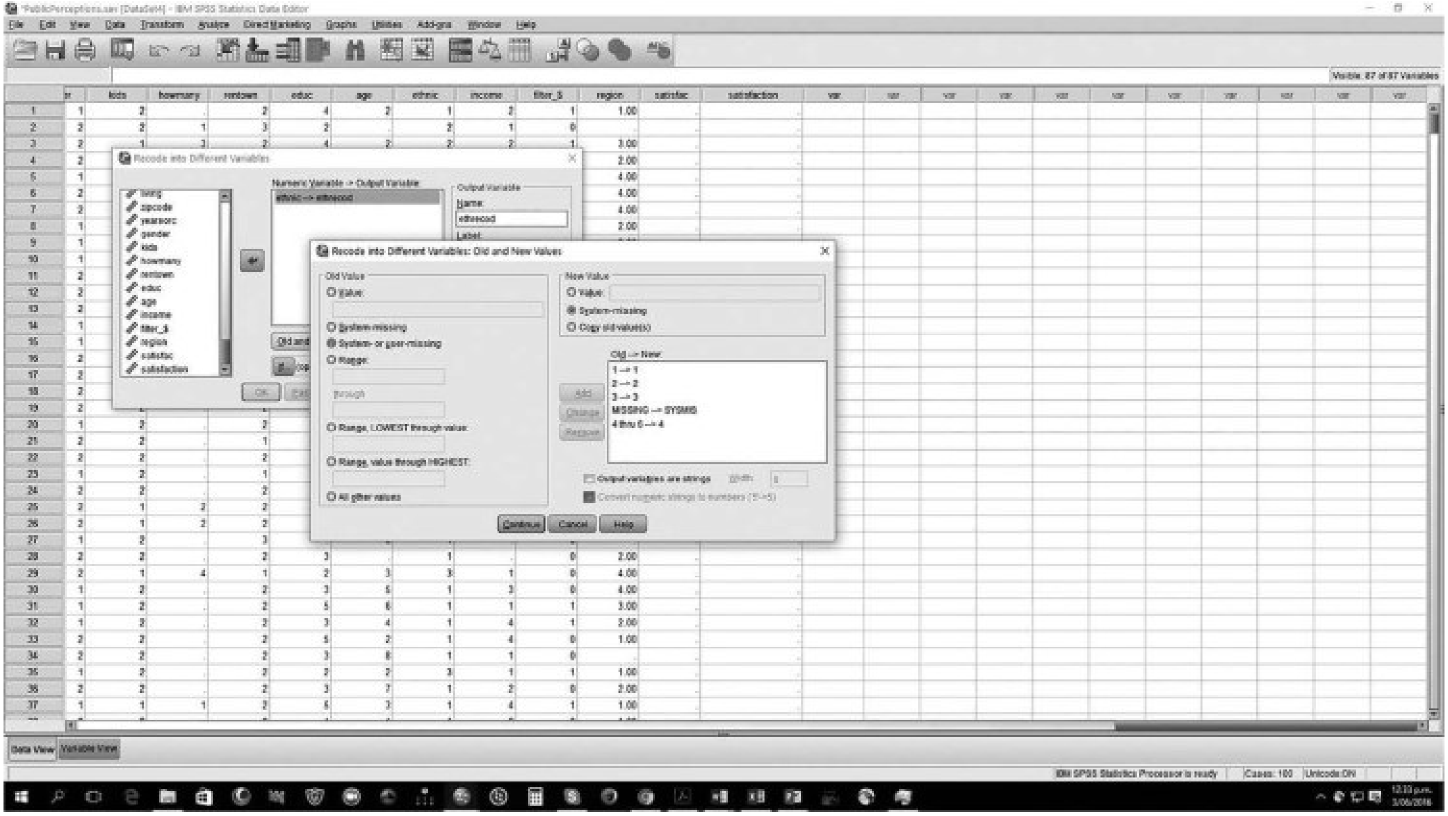
Recode into Different Variables: Old and New Values

Old --> New: 1 --> 1

Select Continue and OK to complete the recoding. To verify that the recoding was successful, perform a Frequency Distribution (recall, Analyze → Descriptive Statistics → Frequencies). The output shown in Screen W19.36 and Figure W19.3 are displayed. The total number of valid responses is still 1,003; there are still 31 missing system values; and the total number of responses is still 1,034. Also, the number of responses comprising the new value of 4 is 72, the sum of Old Values (prior to recoding) of categories 4, 5, and 6. To finish construction of the new variable, ethrecod, remember to define Values and Label on the Variable View screen.

The same procedure is used to define a response value *within* a variable as missing. For example, from the Public Perceptions dataset, select the variable "trust." This variable has three response categories, "Yes," "No," and "Can't Say." To recode "Can't Say" as a missing value, select Transform → Recode → Into Same Variable. When defining response category 3, enter "3" under Old Value and then select "System-missing" for the New Value. The dialogue box shown in Screen W19.37 should be displayed. Select Continue to complete the recoding. Examine the variable trust. Previous cells containing a "Can't Say" response are now coded as missing, as indicated by a period.

Screen W19.35 Missing Data Screen


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 87 of 87 Variables
Data View Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON

Recode into Different Variables

Numeric Variable -> Output Variable: ethnic --> ethrecod

Output Variable Name: ethrecod

Recode into Different Variables: Old and New Values

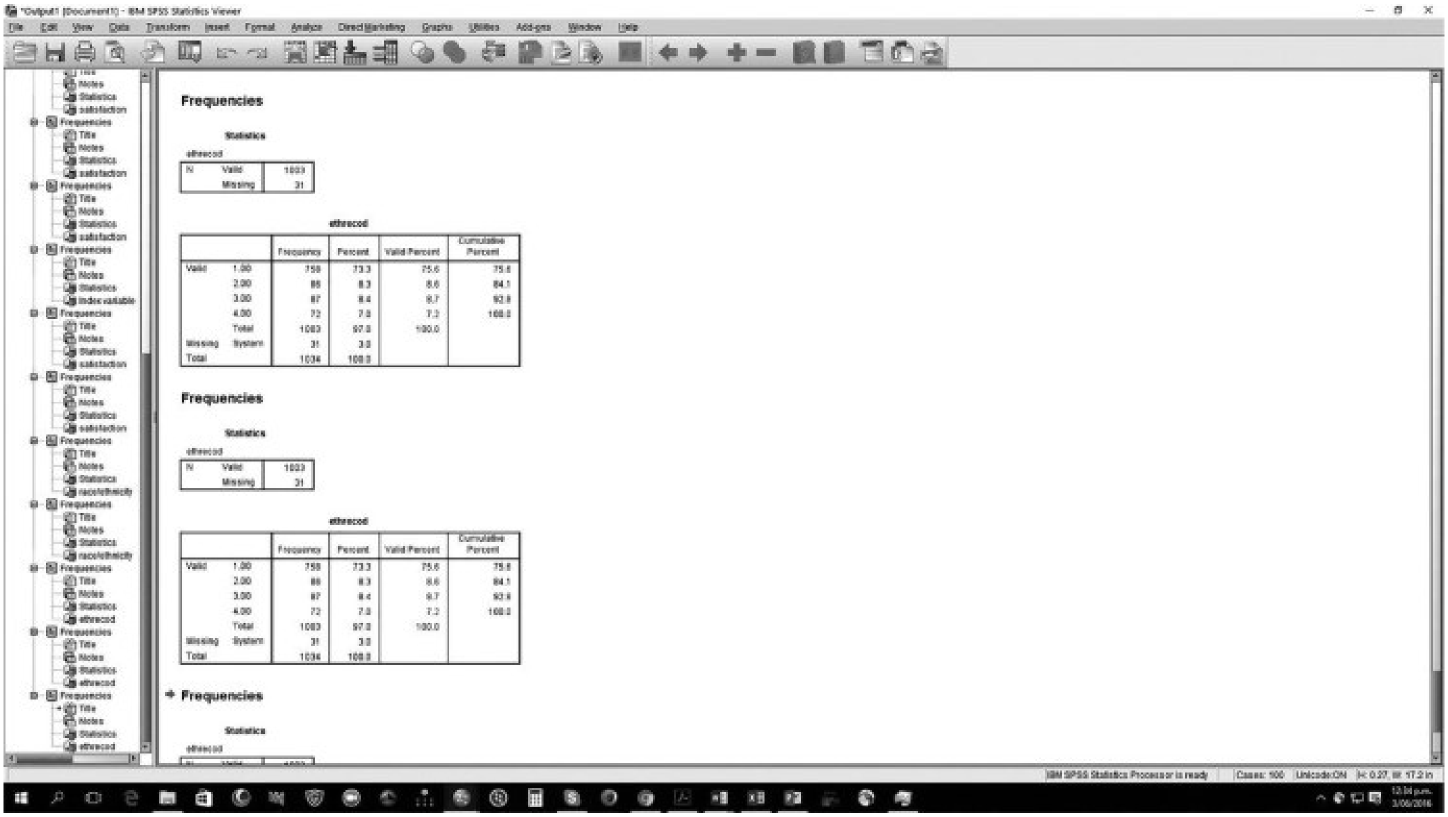
Old Value
○ Value:
○ System-missing
● System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values

New Value
○ Value:
● System-missing
○ Copy old value(s)

Old --> New:
1 --> 1
2 --> 2
3 --> 3
MISSING --> SYSMIS
4 thru 6 --> 4

Add
Change
Remove
Output variables are strings
Width:
Convert numeric strings to numbers ('5'->5)
Continue
Cancel
Help
OK

Screen W19.36 Output for Recoded Data


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Frequencies
Title
Notes
Statistics
satisfaction
index variable
race/ethnicity
ethrecod
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON
H: 0.27, W: 17.2 in

Frequencies

Statistics

ethrecod

N	Valid	1003
	Missing	31

ethrecod

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	759	73.3	75.6	75.6
	2.00	86	8.3	8.6	84.1
	3.00	87	8.4	8.7	92.8
	4.00	72	7.0	7.2	100.0
	Total	1003	97.0	100.0	
Missing	System	31	3.0		
Total		1034	100.0		

Frequencies

Statistics

ethrecod

N	Valid	1003
	Missing	31

ethrecod

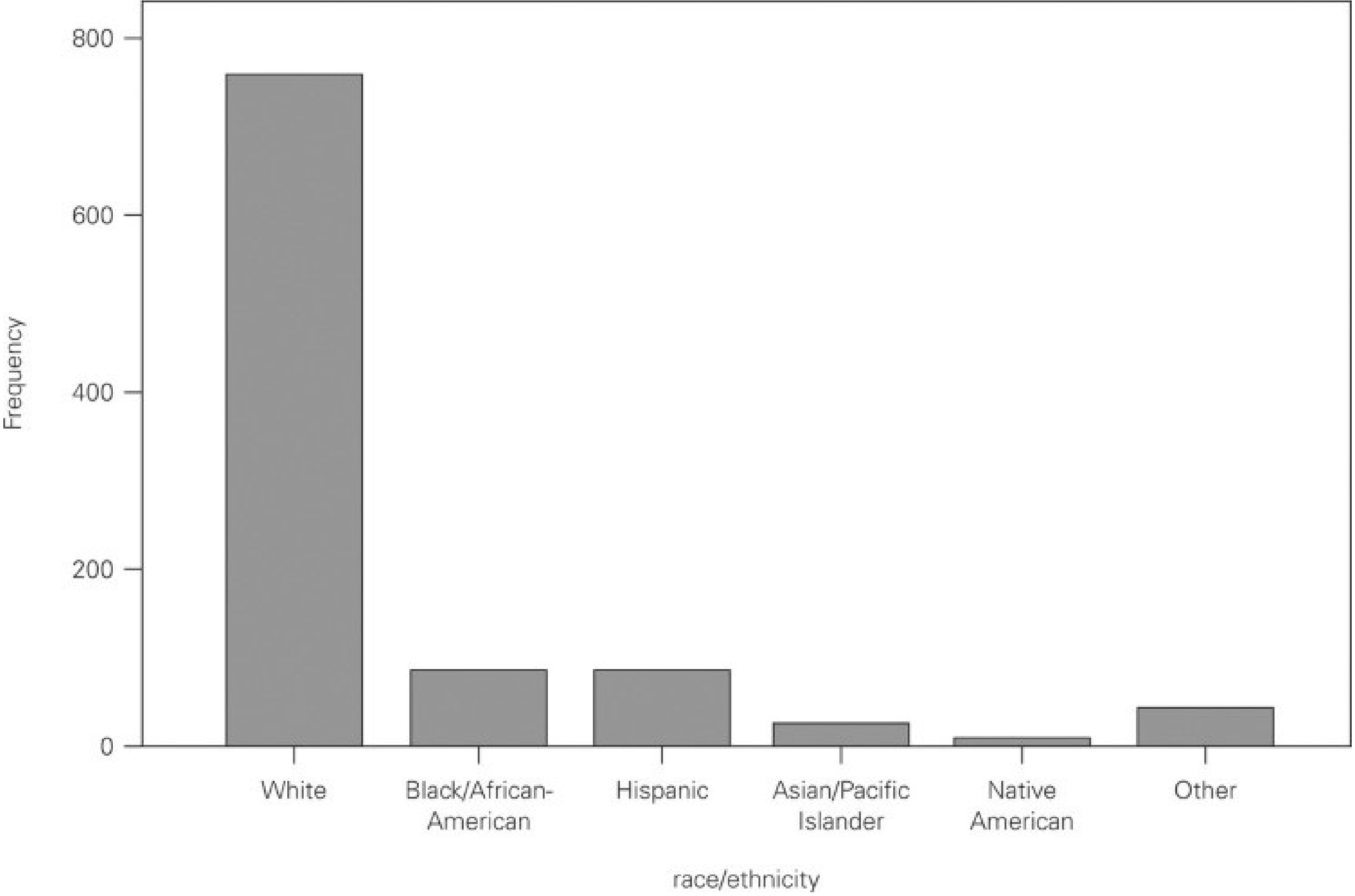
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	759	73.3	75.6	75.6
	2.00	86	8.3	8.6	84.1
	3.00	87	8.4	8.7	92.8
	4.00	72	7.0	7.2	100.0
	Total	1003	97.0	100.0	
Missing	System	31	3.0		
Total		1034	100.0		

Frequencies

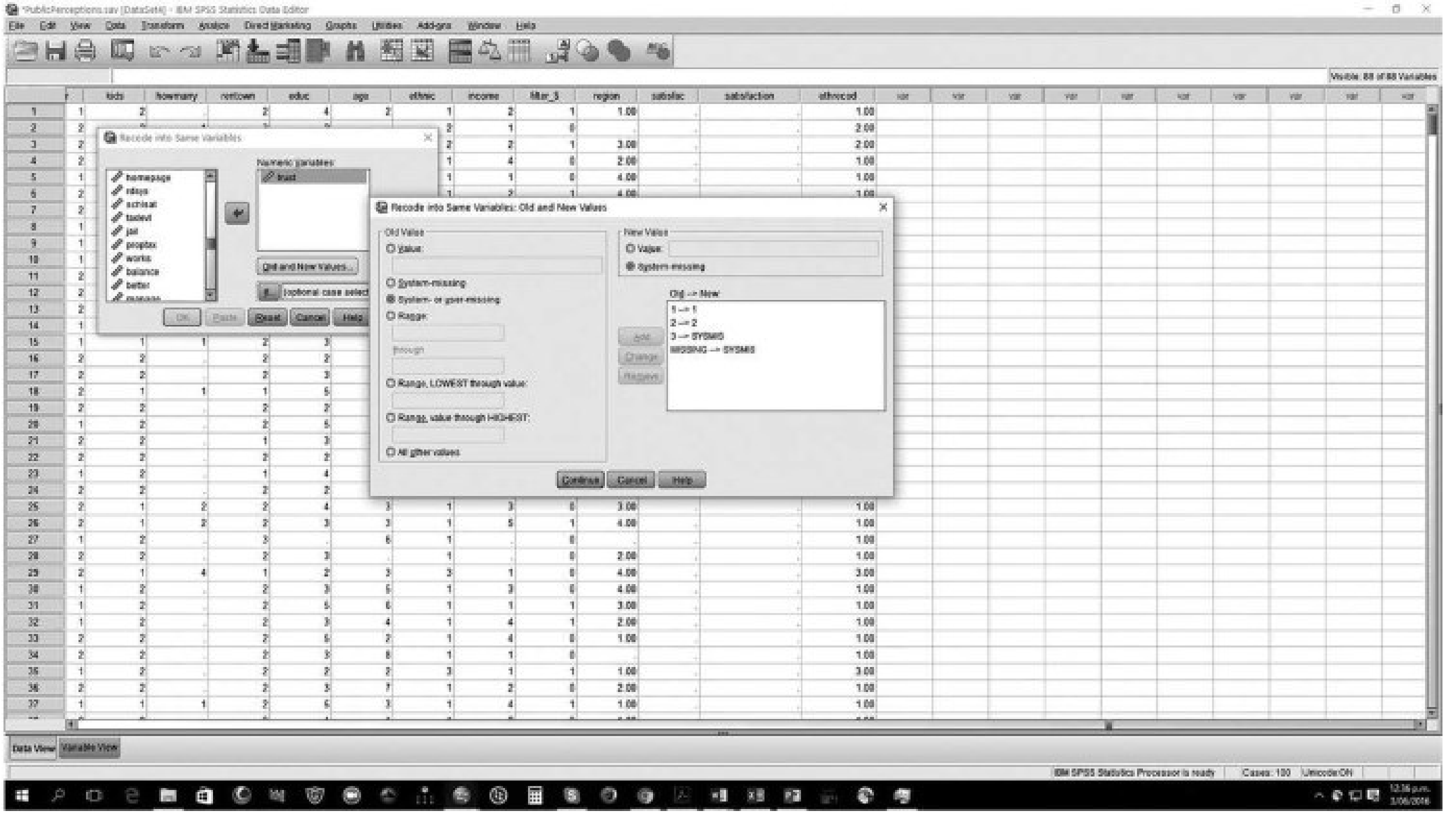
Statistics

ethrecod

Figure W19.3 Recoded Data Bar Chart



Screen W19.37 Recoding Response as a Missing Category



*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 88 of 88 Variables
kids howmany rentown educ age ethnic income filter_$ region satisfac satisfaction ethrecod var
Recode into Same Variables
homepage
rdays
schisat
tadevt
jail
propax
works
balance
better
Numeric Variables:
trust
Old and New Values...
If... (optional case selection condition)
OK Paste Reset Cancel Help
Recode into Same Variables: Old and New Values
Old Value
○ Value:
○ System-missing
● System- or user-missing
○ Range:
through
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values
New Value
○ Value:
● System-missing
Old --> New:
Add
Change
Remove
Continue Cancel Help
Data View Variable View
IBM SPSS Statistics Processor is ready
Cases: 100
Unicode:ON
12:36 p.m.
3/06/2016

Old --> New:

1 --> 1
2 --> 2
3 --> SYSMIS
MISSING --> SYSMIS

Another use for recoding might be to combine response categories from scales, possibly when cell counts are less than sufficient for evaluation. For example, a seven-point Likert scale might have the following response categories: “Strongly Agree,” “Agree,” “Somewhat Agree,” “Neutral,” “Somewhat Disagree,” “Disagree,” and “Strongly Disagree.” This scale could be easily recoded to five response categories by recoding “Strongly Agree/Agree” to “Agree,” and “Strongly Disagree/Disagree” to “Disagree.”

Hypothesis Testing with Chi-Square

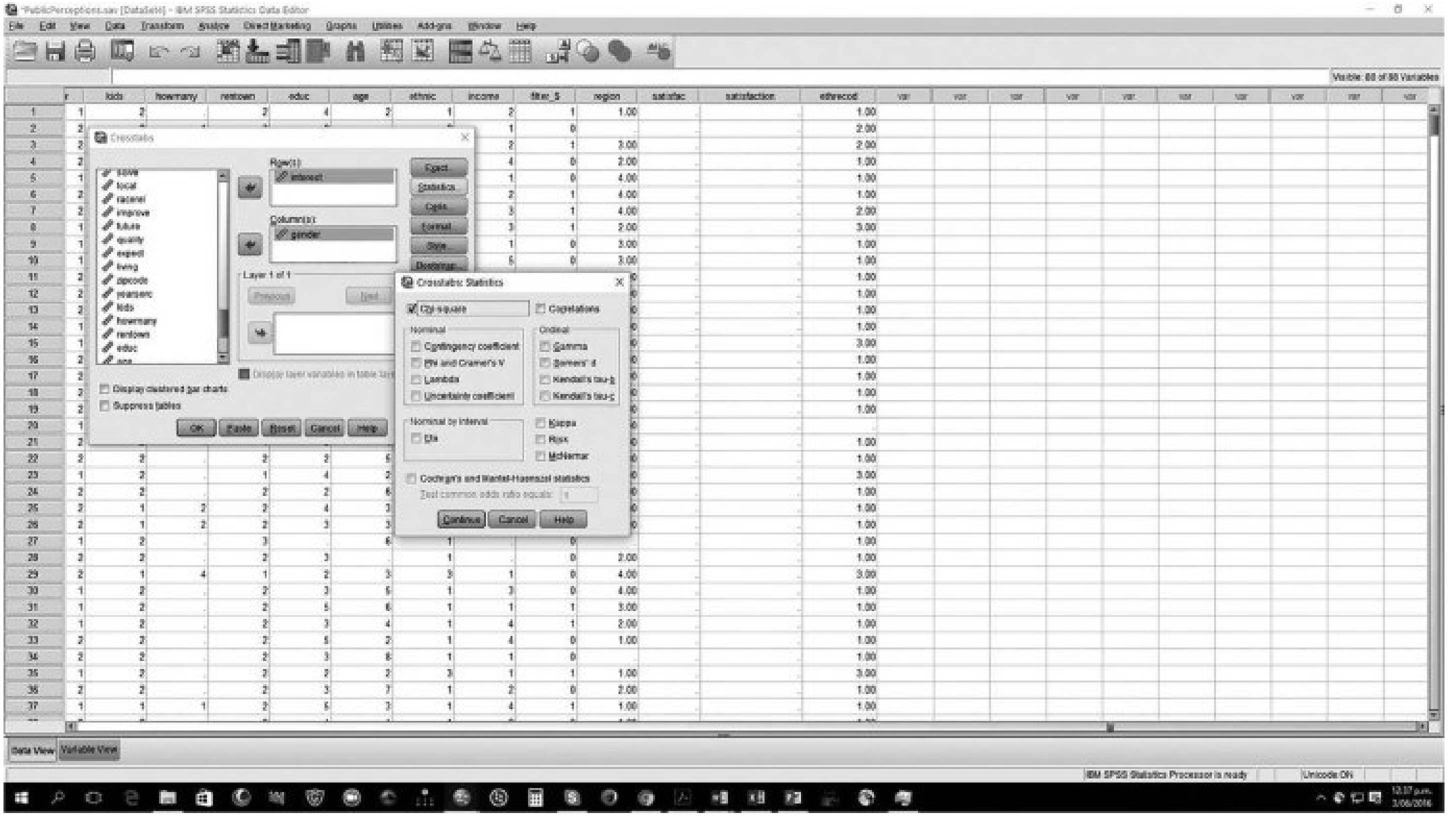
In Chapter 8 of the textbook, you learned that a *contingency table* expresses the relationship between two categorical variables. One variable is shown in rows and the other in columns. Each row shows the frequency of observations with specific values for both variables. Typically, column totals are present. Statistics such as chi-square express the relationship between variables quantitatively. SPSS easily performs contingency table analysis.

Open the Public Perceptions dataset. In the Variable View screen, highlight the variable "interest." The variable label identifies that this variable assesses the extent to which respondents feel county government is interested in what they have to say about issues affecting them. An analyst might ask whether a significant difference exists between the opinions held by male and female respondents on this question.

Chapter 11 of the textbook discusses the concept of the null hypothesis and its application in hypothesis testing. For this exercise, the null hypothesis (H_0) is "Gender has no effect on perception of government interest in citizen opinions." The alternative hypothesis (H_A) is "Gender does influence perceptions of government interest in citizen opinions." You now want to examine whether sufficient statistical evidence exists to reject the null hypothesis and, hence, establish that a relationship does exist. To test this hypothesis, select Analyze → Descriptive Statistics → Crosstabs. Then select "gender" in the column and "interest" in rows. Also, select Statistics → Chi-square → Continue. The dialogue box shown in Screen W19.38 will be displayed. Note how the selection of "gender" in the column is consistent with Chapter 11 in the textbook.

Next, select Cells → Observed → Column → Total → Continue → OK. The results in Figure W19.4 will be produced. First, note that 443 males and 551 females make up the total sample. Second, note the percentages produced within columns. These indicate how males and females responded to this question. For example, 34.5 percent of males and 31.0 percent of females disagree that county government is interested in what they have to say about issues affecting them.

Screen W19.38 Chi-Square Dialogue Box


*PublicPerceptions.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Crosstabs
Exact... Statistics... Cells... Format... Style... Bootstrap...
Layer 1 of 1
Previous Next
Display layer variables in table layers
Display clustered bar charts
Suppress tables
OK Paste Reset Cancel Help
Crosstabs: Statistics
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1
Continue Cancel Help
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Row(s): interest

Column(s): gender

☑ Chi-square

Figure W19.4 Chi-Square Output

I believe that the county government is interested in what I have to say about issues that affect me. *gender Crosstabulation

			gender		
			male	female	Total
I believe that the county government is interested in what I have to say about issues that affect me.	Strongly Agree	Count	11	19	30
		% within gender	2.5%	3.4%	3.0%
	Agree	Count	236	326	562
		% within gender	53.3%	59.2%	56.5%
	Disagree	Count	153	171	324
		% within gender	34.5%	31.0%	32.6%
	Strongly Disagree	Count	43	35	78
		% within gender	9.7%	6.4%	7.8%
Total		Count	443	551	994
		% within gender	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6.711[a]	3	.082
Likelihood Ratio	6.699	3	.082
Linear-by-Linear Association	6.590	1	.010
N of Valid Cases	944		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.37

Review the remaining response categories. Is there a significant difference between male and female respondents? To answer that question, you should rely on the chi-square calculation, which produces the results also shown in Figure W19.4. The chi-square value 6.711 is significant at the 0.082 level. Since this value exceeds the 0.05 standard for statistical significance, you cannot reject the null hypothesis, which means that there is insufficient statistical evidence to conclude that a relationship exists between gender and interest shown by county officials in issues that affect respondents.

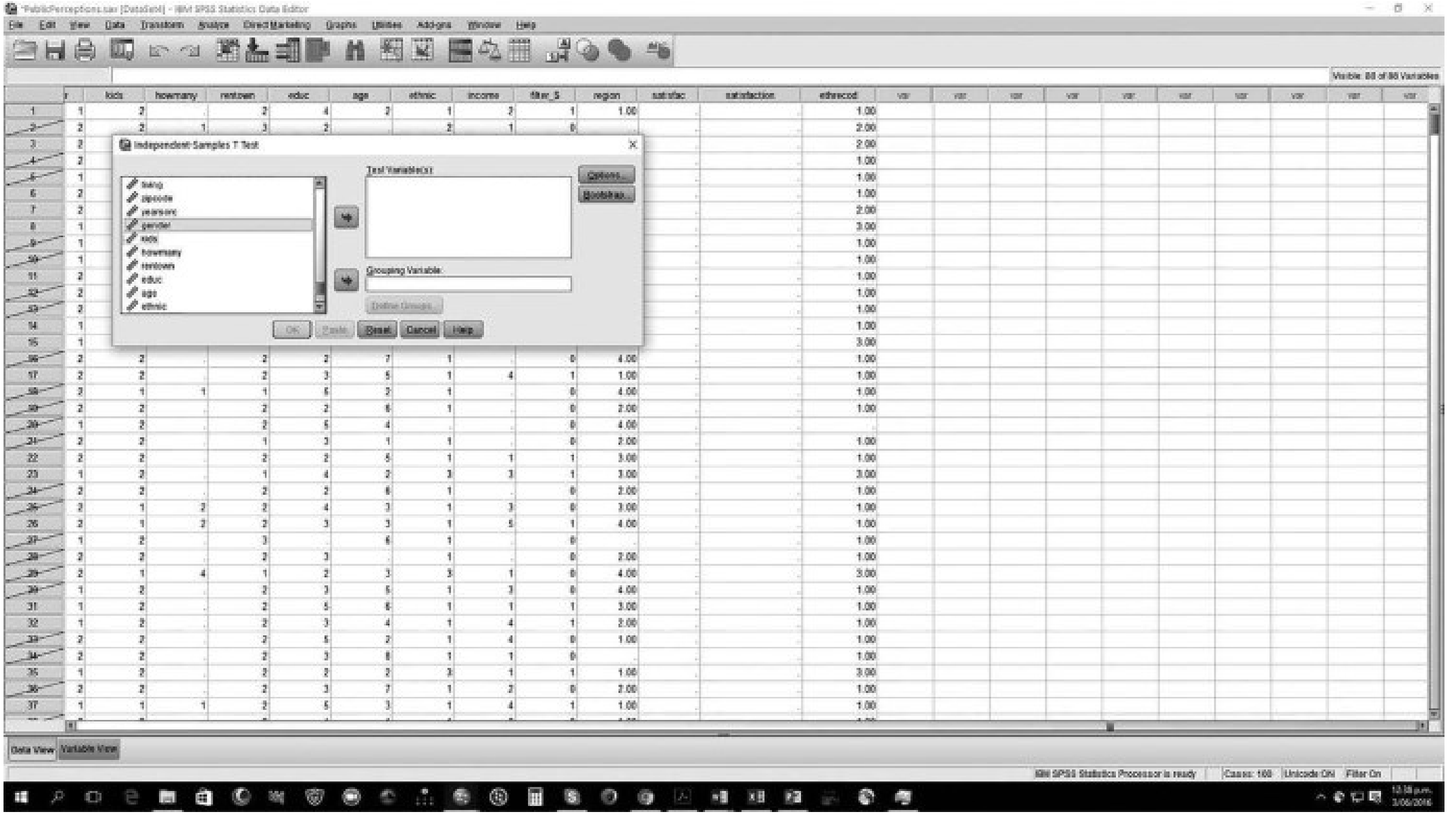
T-Tests

Chapter 12 of the textbook discusses how t-tests are used to determine whether two groups have different means of a continuous variable. For example, do men and women differ in their opinion of the quality of service they receive from county employees? If there is a perceived difference, perhaps county employees deal differently with male and female residents. Again, you need to limit the analysis to those respondents who have had contact with county employees (that is, if contact = 1; see “Selecting a Subset of Observations for Analysis,” earlier in this chapter).

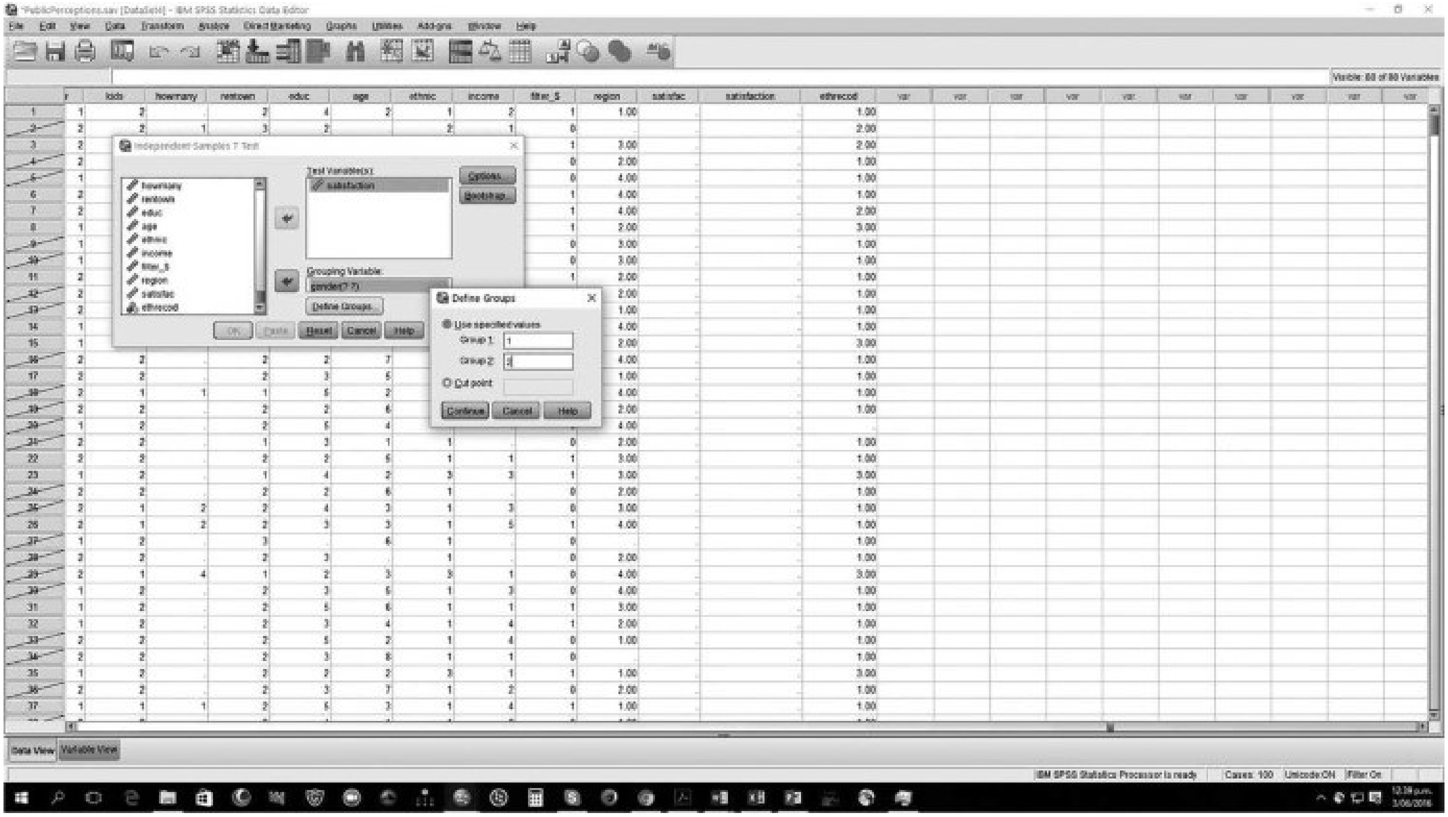
The null hypothesis (H_0) in this case is “Men and women do not have different opinions of the quality of service they receive from county employees.” The alternate hypothesis (H_A) is “Men and women do have different opinions of the quality of service they receive from county employees.”

Open the Public Perceptions dataset. Select Analyze → Compare Means → Independent Samples T-Test. Screen W19.39 will be displayed. Enter “satisfyd” in the Test Variable box and “gender” in the Grouping Variable box. SPSS requires definition of the dichotomous values of “gender” that constitute the grouping variable. Enter “1” and “2” for Group 1 and Group 2, respectively. (When the grouping variable is continuous, a cutoff point—called “Cut point” in Screen W19.40—is specified, which creates the two groups.)

Screen W19.39 Output Screen for T-Test



Screen W19.40 Independent Samples T-Test Dialogue Box



Test Variable(s): satisfaction

Grouping Variable: gender(? ?)

● Use specified values

Group 1: 1

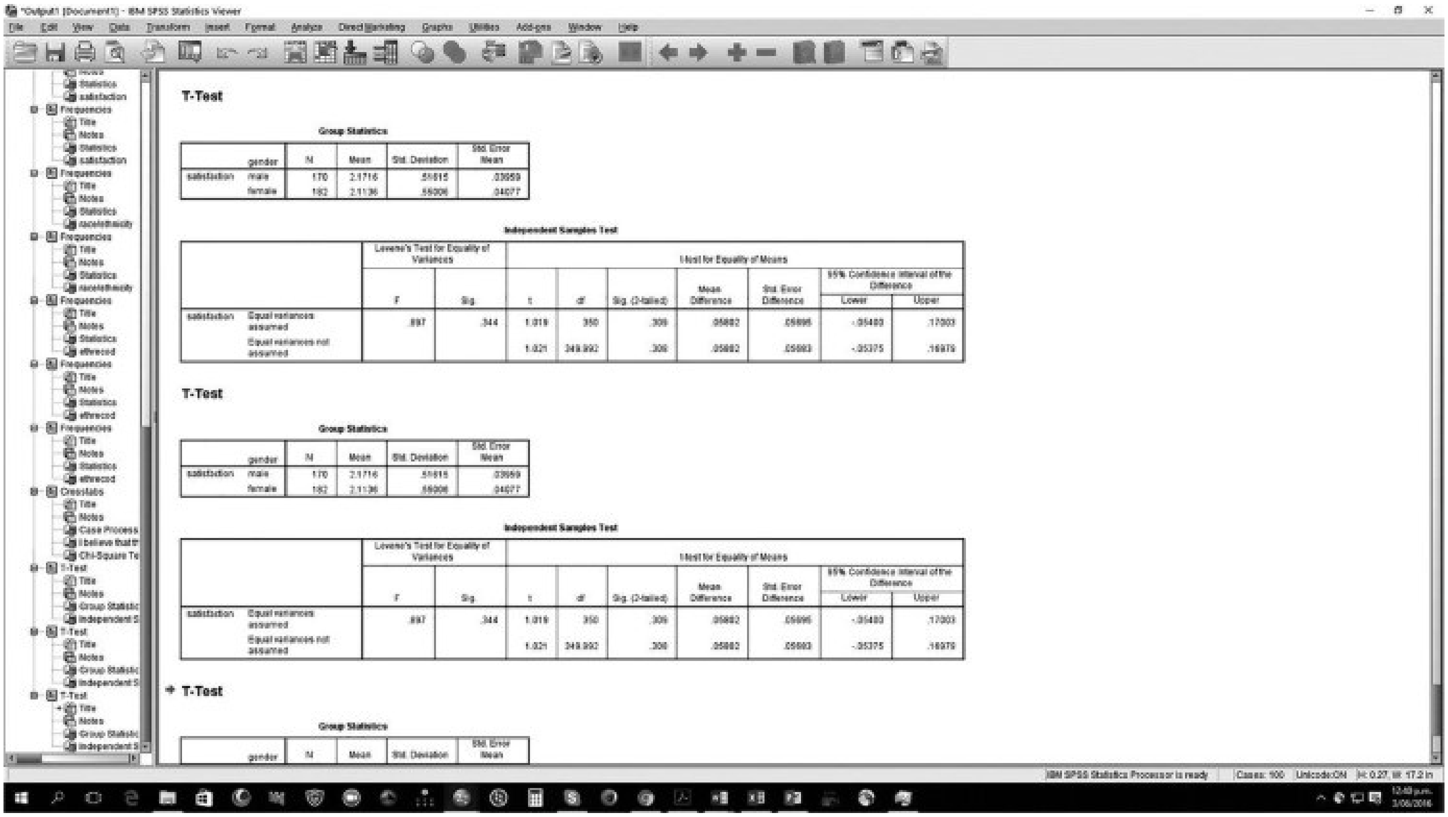
Group 2: 2

○ Cut point

Select Continue → OK. The output shown in Screen W19.41 is produced. The results show that the mean level of satisfaction is 2.17 among men and 2.11 among women. The statistical question is whether this difference in the survey sample is large enough to suggest that a difference exists in the population of all county residents as well.

As discussed in Chapter 12 of the textbook, using a t-test is a two-step process. First, we test whether the variances of the two groups are equal. The null hypothesis is that variances are equal. According to the output shown in Screen W19.41, you cannot reject this null hypothesis (p = .344). Second, you will test whether the difference of means (2.17 versus 2.11) is statistically significant. The result of this t-test is 1.019, which is significant at p = .309. This value exceeds the standard of 5 percent; thus, you conclude that in the population of all county residents, no statistically significant difference exists between men and women with regard to this item. (You also could have examined certain assumptions of the t-test, such as whether the variable is normally distributed and whether any extreme values are present that might affect the analysis.)

Screen W19.41 Output Screen for T-Test



T-Test

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
satisfaction	male	170	2.1716	.51815	.03959
	female	182	2.1136	.55006	.04077

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
satisfaction	Equal variances assumed	.897	.344	1.019	350	.309	.05802	.05696	-.05400	.17003
	Equal variances not assumed			1.021	349.992	.308	.05802	.05683	-.05375	.16979

T-Test

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
satisfaction	male	170	2.1716	.51815	.03959
	female	182	2.1136	.55006	.04077

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
satisfaction	Equal variances assumed	.897	.344	1.019	350	.309	.05802	.05696	-.05400	.17003
	Equal variances not assumed			1.021	349.992	.308	.05802	.05683	-.05375	.16979

T-Test

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean

Conclusion

By now, you have seen just how valuable a tool statistical software can be to public managers and policy analysts. In particular, SPSS's user-friendly package can make statistics easier and save valuable time by "doing the math." Instead of struggling through cumbersome equations, analysts can quickly assess their data and present their findings. In short, analysts now readily recognize the importance of statistical software packages such as SPSS in helping them to quickly add value to decision-making processes.

This guide touched on a number of practical statistical applications, but SPSS has many other features, which you can explore on your own. For example, SPSS can be used to merge datasets, which is described on the companion website. SPSS can also produce a wide range of other statistics. After you have familiarized yourself with other statistics described in the textbook, be sure to consult the SPSS manual that comes with your software to learn about other software features and functions.

Chapter 20 Dataset Documentation

Public Perceptions

General Description

The Public Perceptions dataset includes data from 1,034 telephone interviews among residents of Orange County, Florida. Orange County is the central metropolitan county of Central Florida, encompassing many major tourist attractions. Random digit dialing was used to select interviewees. This is a general citizen survey, encompassing assessments of general county conditions, satisfaction with county services, and demographic items.

As students, some of you may wonder why you should care about these results from Florida. The answer is simple: they could very well have been from your city or county. Such general surveys help managers and policy makers know, in valid ways, where citizens stand on a broad range of issues. Moreover, many citizen surveys explore related questions, and you may be involved in a similar effort or one that is targeted at assessing perceptions about a specific program or service.

Methods

A full description of the survey methods used to develop this dataset is provided in Box 5.2 of the textbook.

Detailed Summary

The survey instrument has seven sections encompassing 96 separate items and is reproduced below. The first section asks the importance of various issues to residents, from controlling government spending to building a light rail transportation system. The rationale for asking these questions is that their ranking may inform county officials in their strategic planning and allocation decisions. The second section asks about perceptions of service experience, generally. Customer service is a key concern to jurisdictions, especially those located close to Disney World, which offers world-class customer service. The third section assesses overall satisfaction with selected county services. The fourth section asks about property taxes. The fifth section addresses a few wide-ranging questions on some matters of current interest to the county. Of particular interest was how often residents watched the local county cable TV station, called Orange TV. The sixth section addresses some general conditions, such as trust in government, expectations of living in Orange County in future years, and perceptions of race relations. Finally, the seventh section provides demographic data regarding age, gender, income, race, and other items. These items are relevant to the analysis of the preceding items. For example, are overall levels of satisfaction with Orange County services associated with perceptions of public safety?

Note on Variables

Each survey item is represented as a variable in the dataset. The variables appear in the same order as in the survey instrument. The variable names are limited to eight characters, and the labels are consistent with the survey items. Variable values are indicated on the survey below, in the first line of each section or as appropriate. Missing values have been coded as periods (.) in the dataset. The sections use a broad range of response scales. The scale on the first section is modified from a seven-point Likert scale. Because all of these items were thought to be of some importance to respondents, a decision was made to group the categories “Very Unimportant,” “Unimportant,” and “Somewhat Unimportant” in one category labeled “Unimportant.”

The survey also includes one continuous variable, the index variable “satisfac.” This variable is the average of the six items shown in Survey Question II, which assesses citizen contact with county employees. This variable is provided only for the 353 respondents who have had contact.

Students can find additional information in the datasets provided on the companion website (http://study.sagepub.com/bermaness4e).

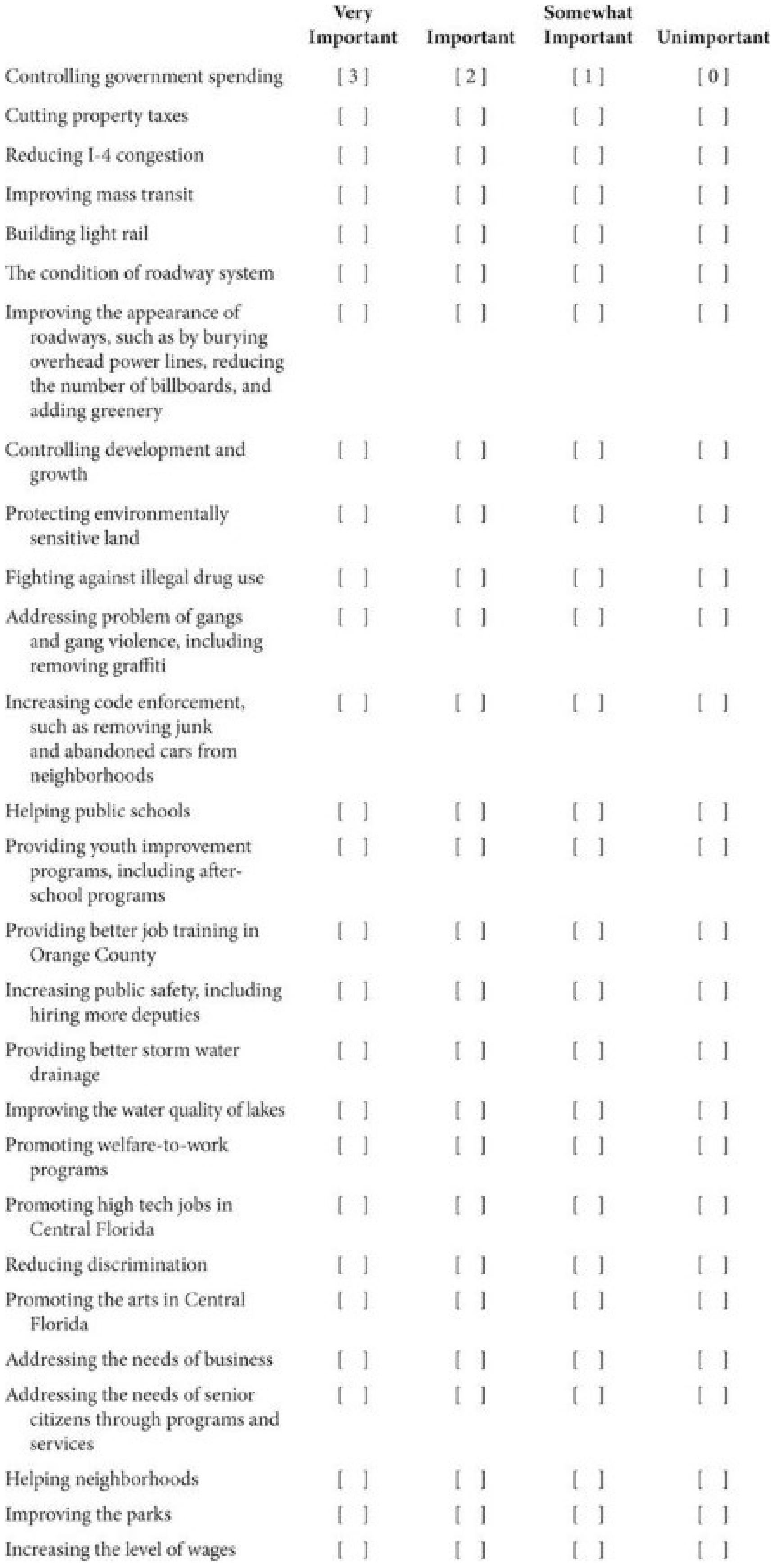
Survey Instrument: Orange County Citizen Survey

Hi, my name is __________, and I'm calling from the Survey Research Laboratory at the University of Central Florida. This is a legitimate survey; I'm not selling anything. We want to know how you feel about several issues facing Orange County. Of course, your participation is completely voluntary, but we hope you will participate. The entire interview should take less than ten minutes. The validity of our results depends on your willingness to help, so we hope that you will participate. Of course, you may discontinue the interview at any time or refuse to answer any questions that make you uncomfortable. We will only report group tendencies in this survey. Your individual answers will be held in strict confidence. Do you have any questions you want to ask before we begin? If you have any questions after the survey, you should call _________________, at UCF. His number is 407-xxx-xxxx.

In order for our survey to be valid, we must interview only persons over the age of eighteen living in Orange County. Would that be you?

If "NO," ask to speak to someone who is eligible, and start over or terminate the call.

1. How important are the following issues for you? Please state whether you consider each issue Very Important, Important, Somewhat Important, or Unimportant:



	Very Important	Important	Somewhat Important	Unimportant
Controlling government spending	[3]	[2]	[1]	[0]
Cutting property taxes	[ ]	[ ]	[ ]	[ ]
Reducing I-4 congestion	[ ]	[ ]	[ ]	[ ]
Improving mass transit	[ ]	[ ]	[ ]	[ ]
Building light rail	[ ]	[ ]	[ ]	[ ]
The condition of roadway system	[ ]	[ ]	[ ]	[ ]
Improving the appearance of roadways, such as by burying overhead power lines, reducing the number of billboards, and adding greenery	[ ]	[ ]	[ ]	[ ]
Controlling development and growth	[ ]	[ ]	[ ]	[ ]
Protecting environmentally sensitive land	[ ]	[ ]	[ ]	[ ]
Fighting against illegal drug use	[ ]	[ ]	[ ]	[ ]
Addressing problem of gangs and gang violence, including removing graffiti	[ ]	[ ]	[ ]	[ ]
Increasing code enforcement, such as removing junk and abandoned cars from neighborhoods	[ ]	[ ]	[ ]	[ ]
Helping public schools	[ ]	[ ]	[ ]	[ ]
Providing youth improvement programs, including after-school programs	[ ]	[ ]	[ ]	[ ]
Providing better job training in Orange County	[ ]	[ ]	[ ]	[ ]
Increasing public safety, including hiring more deputies	[ ]	[ ]	[ ]	[ ]
Providing better storm water drainage	[ ]	[ ]	[ ]	[ ]
Improving the water quality of lakes	[ ]	[ ]	[ ]	[ ]
Promoting welfare-to-work programs	[ ]	[ ]	[ ]	[ ]
Promoting high tech jobs in Central Florida	[ ]	[ ]	[ ]	[ ]
Reducing discrimination	[ ]	[ ]	[ ]	[ ]
Promoting the arts in Central Florida	[ ]	[ ]	[ ]	[ ]
Addressing the needs of business	[ ]	[ ]	[ ]	[ ]
Addressing the needs of senior citizens through programs and services	[ ]	[ ]	[ ]	[ ]
Helping neighborhoods	[ ]	[ ]	[ ]	[ ]
Improving the parks	[ ]	[ ]	[ ]	[ ]
Increasing the level of wages	[ ]	[ ]	[ ]	[ ]

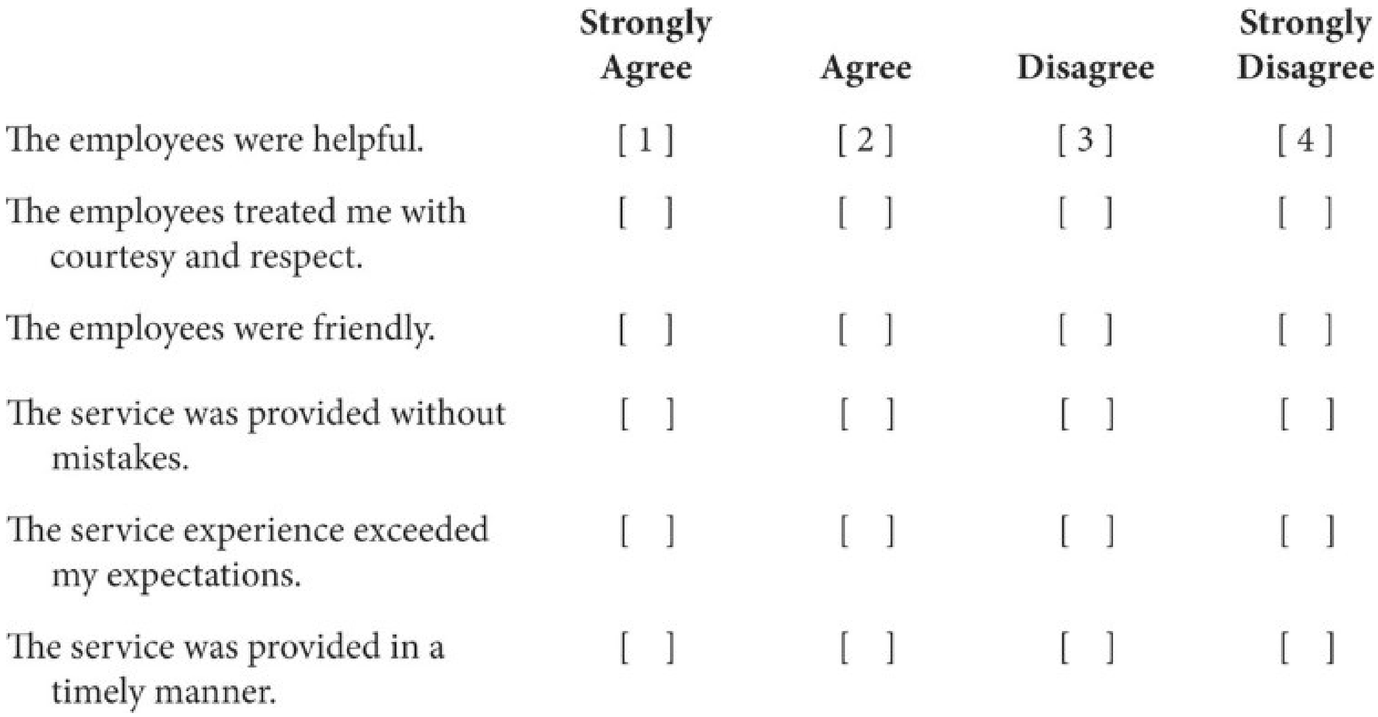
2. Have you had contact with any Orange County employees during the past twelve months (not a family member or friend)?

Yes [1]   No [2]

If yes, was this by phone [1], face-to-face [2], or both [3]? *(Check one.)*

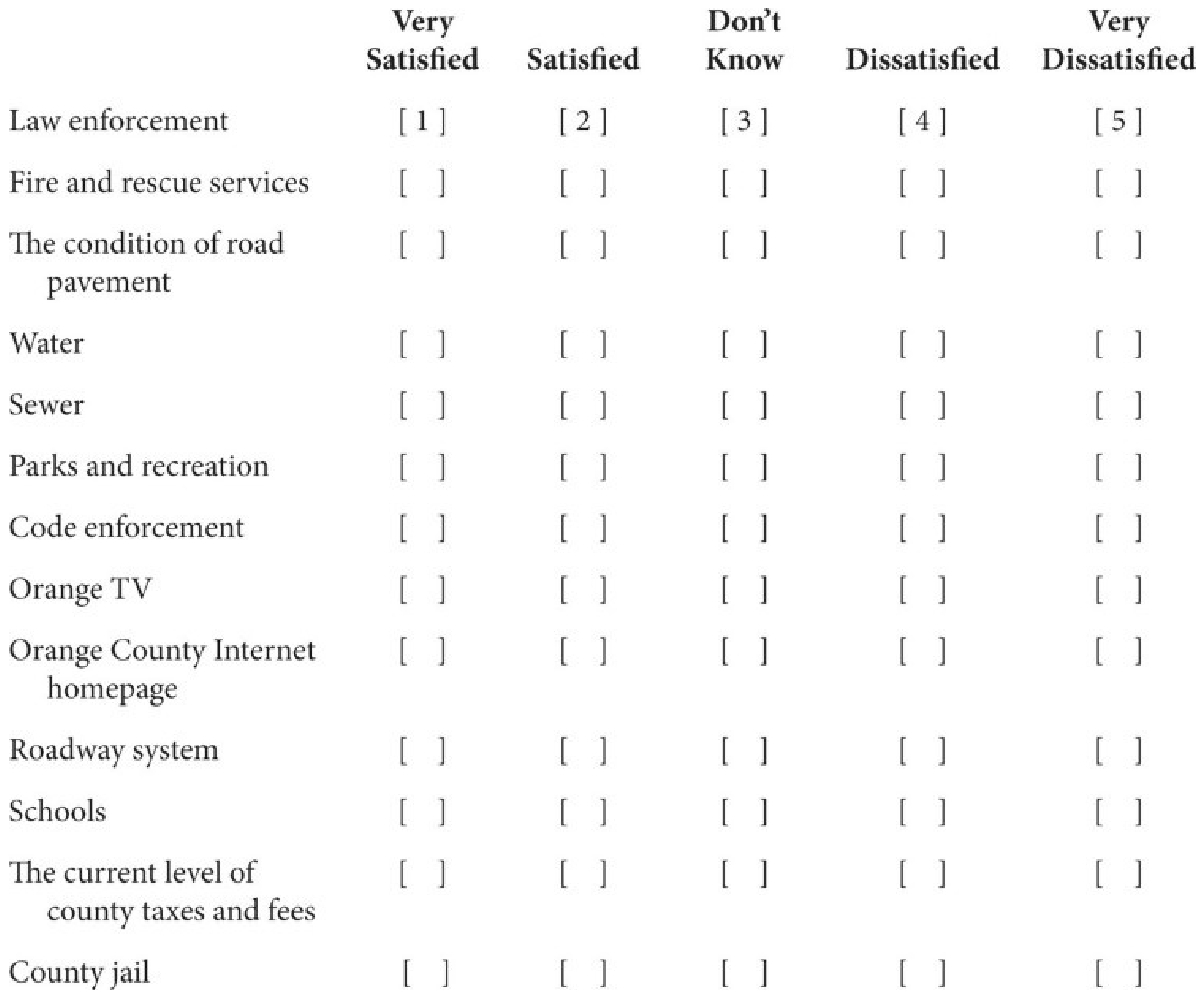
On the average, how often do you visit county offices? ____________ (number)

Thinking about those experiences, or your most recent contact with a County employee, please tell me whether you Strongly Agree, Agree, Disagree, or Strongly Disagree with the following statements:



	Strongly Agree	**Agree**	**Disagree**	**Strongly Disagree**
The employees were helpful.	[1]	[2]	[3]	[4]
The employees treated me with courtesy and respect.	[ ]	[ ]	[ ]	[ ]
The employees were friendly.	[ ]	[ ]	[ ]	[ ]
The service was provided without mistakes.	[ ]	[ ]	[ ]	[ ]
The service experience exceeded my expectations.	[ ]	[ ]	[ ]	[ ]
The service was provided in a timely manner.	[ ]	[ ]	[ ]	[ ]

3. How satisfied are you with the following services in Orange County? Would you say you are Very Satisfied, Satisfied, Dissatisfied, or Very Dissatisfied?



	Very Satisfied	Satisfied	Don’t Know	Dissatisfied	Very Dissatisfied
Law enforcement	[1]	[2]	[3]	[4]	[5]
Fire and rescue services	[ ]	[ ]	[ ]	[ ]	[ ]
The condition of road pavement	[ ]	[ ]	[ ]	[ ]	[ ]
Water	[ ]	[ ]	[ ]	[ ]	[ ]
Sewer	[ ]	[ ]	[ ]	[ ]	[ ]
Parks and recreation	[ ]	[ ]	[ ]	[ ]	[ ]
Code enforcement	[ ]	[ ]	[ ]	[ ]	[ ]
Orange TV	[ ]	[ ]	[ ]	[ ]	[ ]
Orange County Internet homepage	[ ]	[ ]	[ ]	[ ]	[ ]
Roadway system	[ ]	[ ]	[ ]	[ ]	[ ]
Schools	[ ]	[ ]	[ ]	[ ]	[ ]
The current level of county taxes and fees	[ ]	[ ]	[ ]	[ ]	[ ]
County jail	[ ]	[ ]	[ ]	[ ]	[ ]

4. Do you think county property taxes are: Too high [1]
   Too low [2]
   Just about right [3]
   Don’t know [4]
5. The following questions can be answered with a simple yes or no.

   Do you trust Orange County Government to do what is right most of the time?

   Yes [1] No [2] Can’t Say [3]

   Do you believe that Orange County Government works efficiently?

   Yes [ ] No [ ] Can’t Say [ ]

   Has Orange County done a good job of balancing growth against environmental concerns?

   Yes [ ] No [ ] Can’t Say [ ]

   Is your household better off financially than one year ago?

Yes [ ] No [ ] Can’t Say [ ]

Has Orange County Government done a good job of managing growth?

Yes [ ] No [ ] Can’t Say [ ]

Do you watch Orange TV?

Yes [ ] No [ ] Can’t Say [ ]

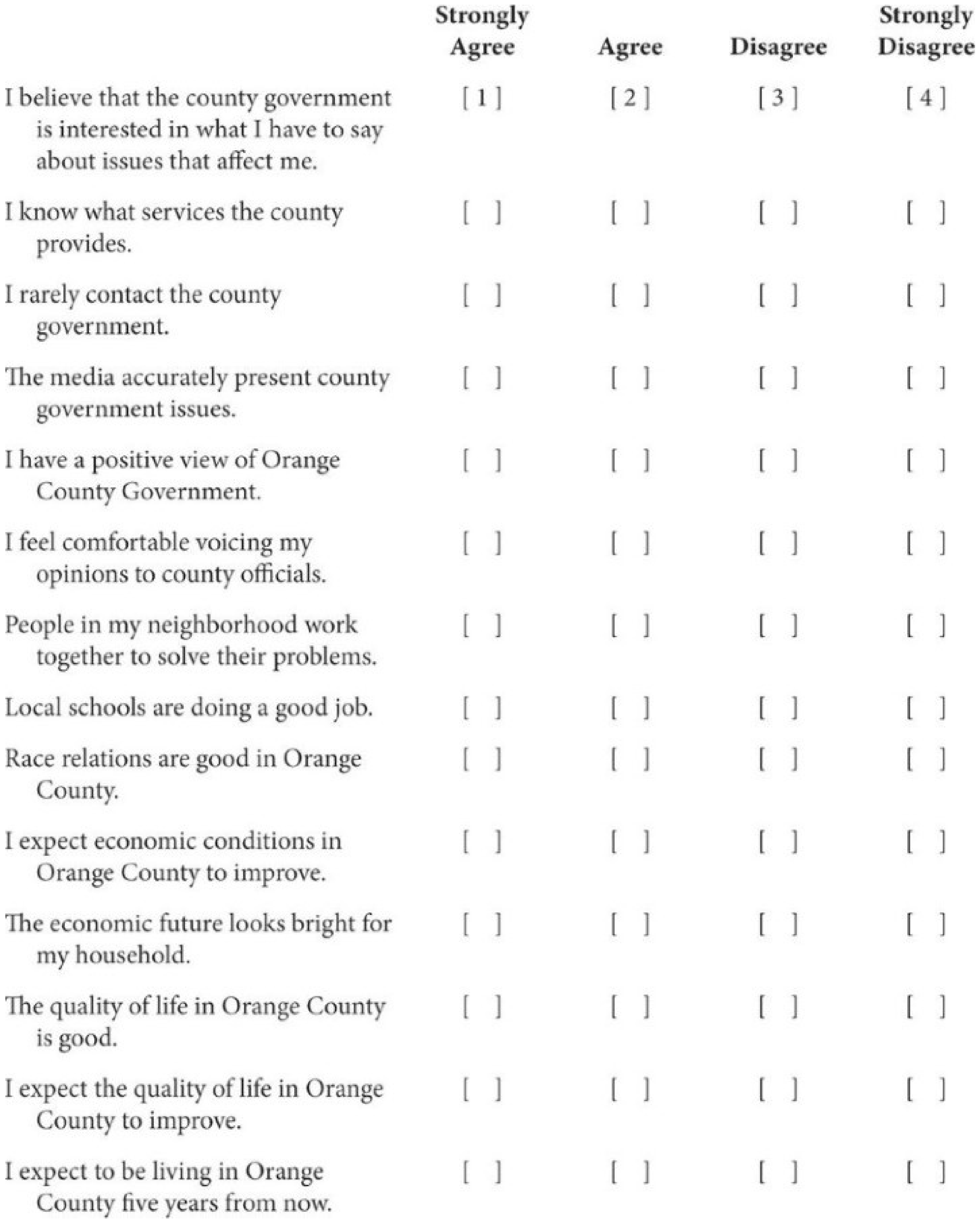
Have you watched Board of County Commission meetings on Orange TV during the last twelve months?

Yes [ ] No [ ] Can’t Say [ ]

Have you ever attended a Board of County Commission meeting?

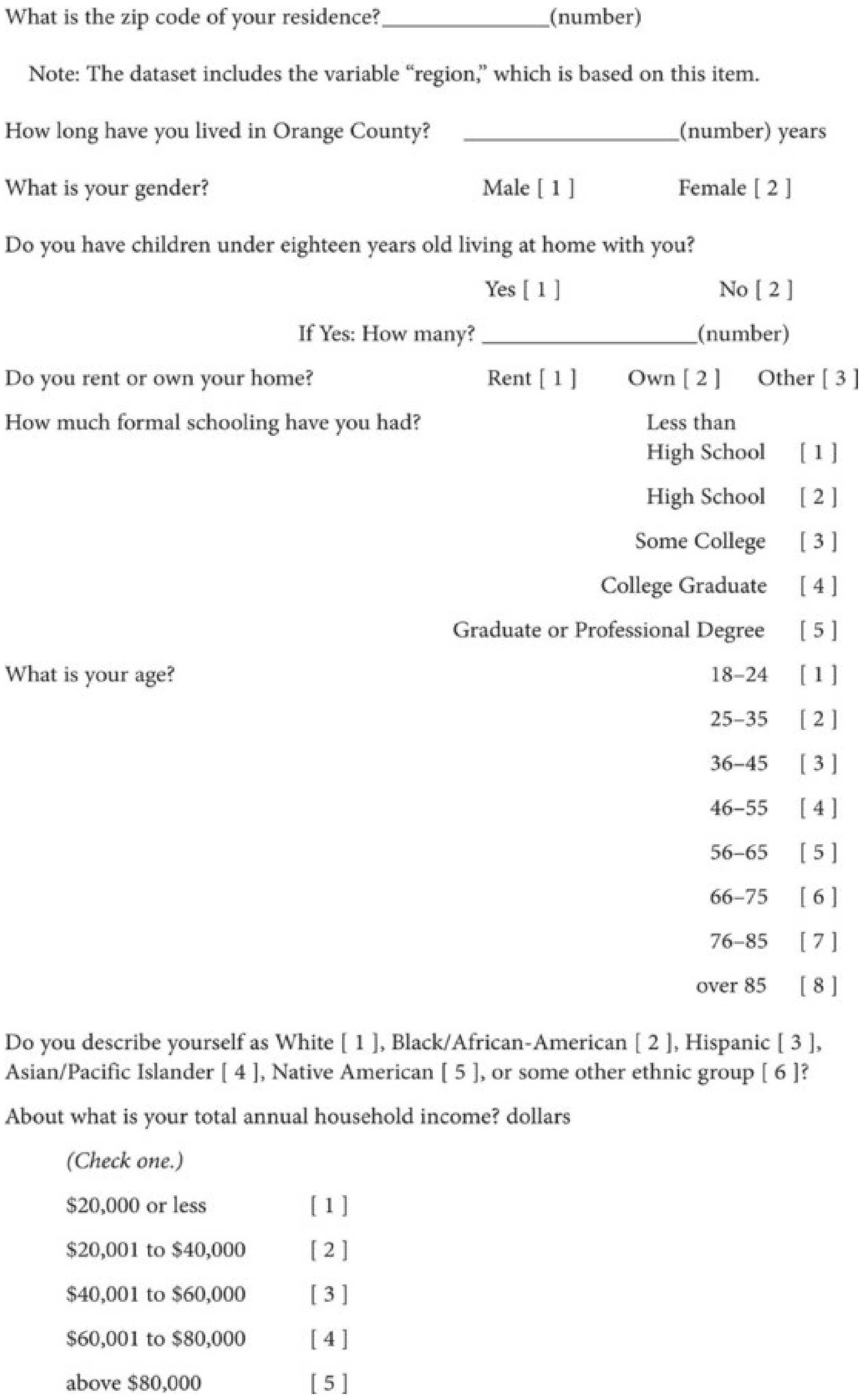
Yes [ ] No [ ] Can’t Say [ ]

6. I will read you some statements. Please tell me whether you Strongly Agree, Agree, Disagree, or Strongly Disagree with the following statements.



	Strongly Agree	Agree	Disagree	Strongly Disagree
I believe that the county government is interested in what I have to say about issues that affect me.	[1]	[2]	[3]	[4]
I know what services the county provides.	[ ]	[ ]	[ ]	[ ]
I rarely contact the county government.	[ ]	[ ]	[ ]	[ ]
The media accurately present county government issues.	[ ]	[ ]	[ ]	[ ]
I have a positive view of Orange County Government.	[ ]	[ ]	[ ]	[ ]
I feel comfortable voicing my opinions to county officials.	[ ]	[ ]	[ ]	[ ]
People in my neighborhood work together to solve their problems.	[ ]	[ ]	[ ]	[ ]
Local schools are doing a good job.	[ ]	[ ]	[ ]	[ ]
Race relations are good in Orange County.	[ ]	[ ]	[ ]	[ ]
I expect economic conditions in Orange County to improve.	[ ]	[ ]	[ ]	[ ]
The economic future looks bright for my household.	[ ]	[ ]	[ ]	[ ]
The quality of life in Orange County is good.	[ ]	[ ]	[ ]	[ ]
I expect the quality of life in Orange County to improve.	[ ]	[ ]	[ ]	[ ]
I expect to be living in Orange County five years from now.	[ ]	[ ]	[ ]	[ ]

7. Your answers to the following questions will help us better analyze the results of this survey. Let me remind you that you may skip any question you choose not to answer.



What is the zip code of your residence?______________(number)

Note: The dataset includes the variable "region," which is based on this item.

How long have you lived in Orange County? __________________(number) years

What is your gender? Male [1] Female [2]

Do you have children under eighteen years old living at home with you?

Yes [1] No [2]

If Yes: How many? __________________(number)

Do you rent or own your home? Rent [1] Own [2] Other [3]

How much formal schooling have you had?

Less than High School	[1]
High School	[2]
Some College	[3]
College Graduate	[4]
Graduate or Professional Degree	[5]

What is your age?

18–24	[1]
25–35	[2]
36–45	[3]
46–55	[4]
56–65	[5]
66–75	[6]
76–85	[7]
over 85	[8]

Do you describe yourself as White [1], Black/African-American [2], Hispanic [3], Asian/Pacific Islander [4], Native American [5], or some other ethnic group [6]?

About what is your total annual household income? dollars

(Check one.)

$20,000 or less	[1]
$20,001 to $40,000	[2]
$40,001 to $60,000	[3]
$60,001 to $80,000	[4]
above $80,000	[5]

That's all the questions I have. Thank you very much for your help. If you have any questions or comments regarding the survey, you should contact ____________ at UCF. Would you like his telephone number again or his email address? By the way, you may get a call from one of my supervisors to check on my performance.

Employee Attitudes

General Description

The Employee Attitudes dataset includes data from 977 employees of Seminole County Government (Florida). This *general employee survey* can easily be tailored to other research situations. Seminole County is one of five counties in Central Florida. One of the smaller counties in terms of area, it includes a mix of very affluent neighborhoods as well as some rural areas and poor neighborhoods. Human Resources Department staff conducted the survey and visited each county department. Staffs in each department were required to attend a meeting at which the anonymous and voluntary survey was administered; only a few employees chose not to participate in the survey. This is a general employee survey—employees are referred to in this county as “members”—and it includes employee assessments of working conditions, career development, benefits and compensation, supervisory management, customer relations, job skills and training, and satisfaction with Human Resources services. To ensure the anonymity of respondents, surveys were returned in sealed and unmarked envelopes, which were opened by an outside, independent consultant who did the analysis and prepared the final report.

Methods

Considerable care was taken to ensure that the questions were unbiased and consistent with questions typically found in employee surveys. In August 1999, Human Resources Department staff implemented the survey across departments and divisions by assembling all Seminole County Government employees, along with managers, for the purpose of completing the survey. Efforts were made to reach all members, and those unavailable on that day were advised of other opportunities to participate. To ensure the anonymity of members, participants were instructed not to write their name on their survey and to insert the survey into an unmarked white envelope. All envelopes were then delivered to the outside researcher, who opened and analyzed the responses.

A total of 977 surveys were received, or 84.4 percent of the 1,158 total full-time positions with Seminole County. Table W20.1 shows the distribution of these positions by department. A methodological issue is that Seminole County Government chose to state clearly that answering various standard demographic items was "voluntary," which reduced the number of completed responses for these items. Table W20.2 shows the number of respondents who answered various demographics items. A question arises as to whether the lower response rates for demographic variables may result in any significant *response bias*. To explore this possibility, the aggregate responses of all respondents were compared with the responses of those who answered specific demographic questions. Typically, mean responses should differ by no more than ±0.03, and no difference should be greater than ±0.06. To illustrate the analysis, Table W20.3 shows the results of the following three questions, which were selected for their importance or relevance to this concern.

> Item 1: "Seminole County is a good place to work for compared to other organizations I know or have worked for."
>
> Item 2: "The quality of service provided to citizens is the same regardless of their race, gender, or background."
>
> Item 3: "My supervisor deals fairly with everyone."

Looking over these data, we can conclude that analyses using the demographic variables are meaningful.

Finally, analysis of the data shows that some departments are quite small and that Seminole County Government has relatively few employees in each of the standard race classifications. For example, only 68 members, or 8.0 percent, of respondents identified themselves as African-American. Likewise, only 37, or 4.4 percent, identified themselves as being of Hispanic origin. The final report does not provide separate reporting for departments with either fewer than 20 respondents or less than a 50 percent response rate. Analyses of department by race contrast Caucasian and non-Caucasian employee perceptions in order to preserve the anonymity of minority respondents in each department.

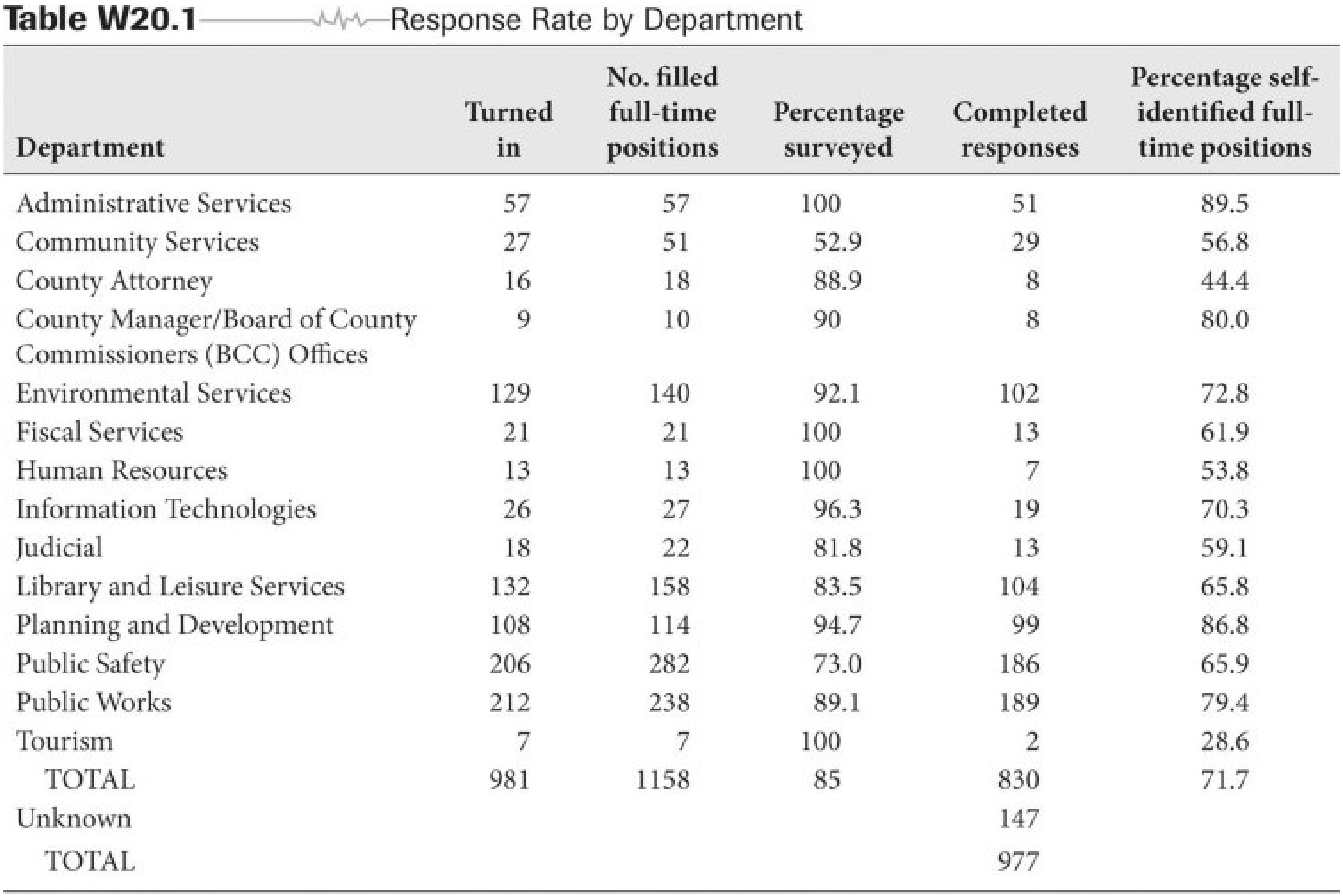


Table W20.1 Response Rate by Department

Department	Turned in	No. filled full-time positions	Percentage surveyed	Completed responses	Percentage self-identified full-time positions
Administrative Services	57	57	100	51	89.5
Community Services	27	51	52.9	29	56.8
County Attorney	16	18	88.9	8	44.4
County Manager/Board of County Commissioners (BCC) Offices	9	10	90	8	80.0
Environmental Services	129	140	92.1	102	72.8
Fiscal Services	21	21	100	13	61.9
Human Resources	13	13	100	7	53.8
Information Technologies	26	27	96.3	19	70.3
Judicial	18	22	81.8	13	59.1
Library and Leisure Services	132	158	83.5	104	65.8
Planning and Development	108	114	94.7	99	86.8
Public Safety	206	282	73.0	186	65.9
Public Works	212	238	89.1	189	79.4
Tourism	7	7	100	2	28.6
TOTAL	981	1158	85	830	71.7
Unknown				147	
TOTAL				977	

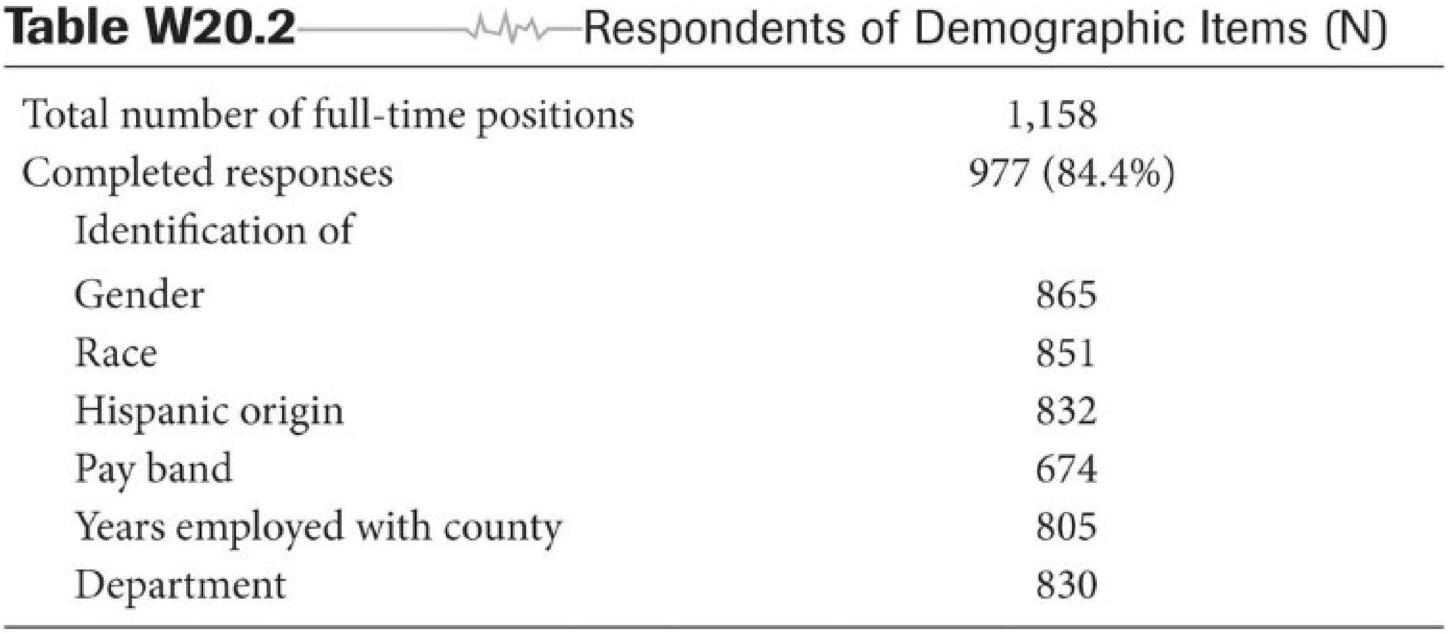


Table W20.2 Respondents of Demographic Items (N)

Total number of full-time positions	1,158
Completed responses	977 (84.4%)
Identification of	
Gender	865
Race	851
Hispanic origin	832
Pay band	674
Years employed with county	805
Department	830

Table W20.3 Mean Answers to Items 1, 2, and 3

	Item 1 (mean)	Item 2 (mean)	Item 3 (mean)
All responses completed	3.64	4.31	3.51
Gender	3.66	4.33	3.52
Race	3.64	4.35	3.54
Hispanic origin	3.65	4.36	3.54
Pay band	3.67	4.32	3.47
Years employed	3.66	4.35	3.51
Department	3.64	4.32	3.52

Detailed Summary

The questions on this survey are consistent with most general employee surveys, which cover many of the same aspects discussed below. The survey instrument has 10 sections designed to assess different aspects of the workplace. The first section measures overall satisfaction with the county as a place to work. The second section assesses a broad range of general working conditions, such as the presence of safety hazards and the manner in which problems are discussed. The third section looks at employee relations with supervisors and management. The fourth section examines how customers are treated. The fifth section focuses on career development. The sixth section assesses cooperation and coordination among departments. The seventh section features questions on the adequacy of job skills and training. The eighth section delves into the quality of Human Resources services and interaction. The ninth section consists of two questions about benefits and compensation. The tenth section asks a range of standard demographic questions.

Note on Variables

The variables appear in the same order as on the survey instrument. The variable names are limited to eight characters, and the labels are consistent with the survey items. Variable values are indicated in the survey below. Missing values are coded as periods (.) in the dataset. Most items are on a five-point Likert scale.

Students can find additional information in the datasets provided on the companion website (http://study.sagepub.com/bermaness4e).

Survey Instrument: Seminole County Government Employee Survey

The following questions are designed to assess and improve employee relations. Please evaluate the following statements by checking the appropriate box (mark an “X,” please).

SA = Strongly Agree   DK = Don't Know   SD = Strongly Disagree
A = Agree   D = Disagree

	SA	A	DK	D	SD
1. General Conditions					
Overall, I am satisfied with my job at Seminole County	5	4	3	2	1
Seminole County is a good place to work compared to other organizations I know about or have worked for	□	□	□	□	□
Each individual is treated with dignity	□	□	□	□	□
The morale of Seminole County Government employees is high	□	□	□	□	□
In general, my department is better to work for than it was two years ago	□	□	□	□	□
In general, Seminole County Government is a better place to work for than it was two years ago	□	□	□	□	□
Most days I feel good about coming to work	□	□	□	□	□
I would proudly recommend Seminole County Government as a good place to work	□	□	□	□	□
The work that I do is important	□	□	□	□	□
Our organization welcomes change	□	□	□	□	□
The County Newsletter is an effective communication tool	□	□	□	□	□
It is important for the organization to commit to learning about cultural diversity	□	□	□	□	□
My job is challenging and interesting	□	□	□	□	□
2. Working Conditions					
Problems are discussed openly, candidly and constructively	□	□	□	□	□
In general, everyone seems to carry his or her fair share of the workload	□	□	□	□	□
I have a lot of freedom to decide how to do my work	□	□	□	□	□
I seldom feel stressed because of work	□	□	□	□	□
I seldom work late to get my scheduled assignments done	□	□	□	□	□
My physical working conditions are reasonable for my type of work	□	□	□	□	□
The maintenance on the equipment I use is adequate	□	□	□	□	□
Safety hazards in my work area are quickly corrected	□	□	□	□	□
Most of my co-workers are receptive to trying new ways of conducting business to improve productivity	□	□	□	□	□
My supervisor is skilled and experienced	□	□	□	□	□
3. Supervisory/Management Relations					
I feel free to go to a "higher level" than my immediate supervisor to discuss any problems that are bothering me	□	□	□	□	□
Senior managers of the organization come through my work area often	□	□	□	□	□
I have confidence and trust in my supervisor	□	□	□	□	□
My supervisor deals fairly with everyone – does not play favorites	□	□	□	□	□
My division routinely holds meetings to keep people informed	□	□	□	□	□
My immediate supervisor helps my work group do its best	□	□	□	□	□
Feedback on performance is timely, accurate and constructive	□	□	□	□	□
My supervisor and I establish performance goals for the upcoming year as part of my performance evaluation	□	□	□	□	□
I understand what is expected of me in my job	□	□	□	□	□
4. Customer Relations					
The quality of service provided to citizens is the same regardless of their race, gender or background	□	□	□	□	□
I feel very confident to make decisions to take action to satisfy our customers	□	□	□	□	□
The opinions of our customers are important to my work group	□	□	□	□	□
Most of my co-workers are receptive to trying new ways of conducting business to improve service	□	□	□	□	□
5. Career Management and Rewards					
Job opportunities are posted and accessible to me	□	□	□	□	□
When things go well at my job, my contributions are recognized	□	□	□	□	□
The people that get promoted are among the best qualified for the job	□	□	□	□	□
I actively participate and provide input into my performance appraisal/review	□	□	□	□	□
Our organization has a good performance appraisal system in place	□	□	□	□	□
I am familiar with the Merit for Progress	□	□	□	□	□
I understand how my performance is evaluated	□	□	□	□	□
I am satisfied with the various employee recognition programs	□	□	□	□	□
I understand what I need to do to develop my career with Seminole County Government	□	□	□	□	□
Internal promotions are encouraged here	□	□	□	□	□
6. Cooperation and Coordination					
There is cooperation among departments to get the job done	□	□	□	□	□
Our organization does a good job of keeping us informed about current developments affecting the organization	□	□	□	□	□
I understand the organization's vision, mission, goals and values	□	□	□	□	□
There is a good feeling of teamwork in my work group with people working well together	□	□	□	□	□
I am aware of my responsibilities and the procedures to follow in resolving a conflict	□	□	□	□	□
7. Job Skills and Training					
Members are given challenging work that provides opportunities to learn new skills	□	□	□	□	□
Members have the skills to do their work well	□	□	□	□	□
My performance has improved as a result of attending training programs	□	□	□	□	□
I receive ongoing training to keep my skills current	□	□	□	□	□
8. Human Resources Department					
The Human Resources staff is accessible and easy to work with	□	□	□	□	□
The Human Resources Department does a good job of representing all employees and acting as a employee advocate	□	□	□	□	□
I am comfortable seeking information and advice from Human Resources	□	□	□	□	□
I understand how the Employee Assistance Program works and its benefit to me	□	□	□	□	□
I can use the grievance procedures without fear of retribution	□	□	□	□	□
The Personnel Policies Manual is well organized and easy to work with	□	□	□	□	□
9. Benefits and Compensation					
Seminole County's insurance package (health, life and optional coverage) meets the needs of employees	□	□	□	□	□
I am paid as well as people in other organizations with similar jobs	□	□	□	□	□

10. Demographic Questions

Finally, the following questions are asked for the purpose of statistical analysis only. Please check the appropriate boxes.

1. What is your gender?	Male	1
	Female	2
2. What is your race?	Caucasian	1
	African American	2
	Asian or Pacific	3
	Native American	4
	Other	5
3. Are you of Hispanic origin?	Yes	1
	No	2
4. What is your department?	Administrative Services	1
	Community Services	2
	County Attorney's Office	3
	County Manager/BCC	4
	Environmental Services	5
	Fiscal Services	6
	Human Resources	7
	Information Technologies	8
	Judicial	9
	Library & Leisure Services	10
	Planning & Development	11
	Public Safety	12
	Public Works	13
	Tourism	14
5. What is your pay band?	Pay Band 1 or 2	1
	Pay Band 3 or 4	2
	Pay Band 5 or 6	3
	Pay Band 7, 8, or 9	4
6. How long have you worked for Seminole County Government?	Less than 1 year	1
	1+ to 3 years	2
	3+ to 5 years	3
	5+ to 10 years	4
	10+ years	5

Community Indicators

General Description

The Community Indicators dataset is developed from secondary data sources to provide a profile of cities across the United States. The 98 selected cities include many of the country's largest, as well as those from different states and a few located in rural settings. The data reflect different aspects of interest: demographic, educational, crime, environmental, transportation, sports, economic, and other concerns.

A challenge in gathering this information is that although some data are available at the city level, other data are available only at the county level, and the county boundaries do not always conform to the city boundaries. This is especially the case when data are collected not by the U.S. Census, but by other agencies. When data are collected by the U.S. Census, city estimates are available for demographic, social, educational, and housing data. We identify instances in which data were collected at the county level; many crime and health data are of this nature. In addition, we estimate, based on a visual analysis comparing maps, how well county boundaries match those of the city jurisdictions. In some cases, the fit is quite close, and in other instances it is not.

The main data sources are the U.S. Census (http://factfinder.census.go) and heath and crime data from websites that have since changed. Current websites pertaining to the variables of this website include U.S. Department of Justice, Uniform Crime Reporting (https://ucr.fbi.gov/ucr-publications, see also http://www.ojjdp.gov/ojstatbb/ezaucr/asp/ucr_display.asp), U.S. Department of Health and Human Services, Office on Women's Health (https://www.womenshealth.gov/statistics/government-in-action/qhdata-how-to-use-guide.pdf, see also http://52.207.219.3/qhdo/index.html), and U.S. Environmental Protection Agency (https://www3.epa.gov/airdata).

Summary of Variables

Each variable name appears after its label.

City name—City. This is the city name, as used by the census.

Caucasian population—Caucasian. The percent of the population that is white.

Number of persons under 18 years—UnderAge18. The number of persons in the city under age 18 years.

Number of persons from 18 through 44—Age18to44. The number of persons in the city with ages 18 through 44 years.

Number of persons from 45 through 64—Age45to64. The number of persons in the city with ages 45 through 64 years.

Number of persons over 65 years—OverAge65. The number of persons in the city over age 65 years.

Total population—Pop. Total population (estimated).

Percenta of persons with high school degrees or more—HSGradorHigher. Percent of population, over age 25 years, who have a high school degree or higher degrees.

Percent of persons with bachelor's degrees or more—BGradorHigher. Percent of population, over 25 years, who have a bachelor's degree or higher degrees.

Number of violent crimes—ViolentCrime. The total number of violent crimes. According to Uniform Crime Reporting, violent crimes involve force or threat. There are four types of violent crimes: murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault. In terms of the rate of offenses for each of the four violent crimes, aggravated assault had the highest rate, estimated at 291.1 offenses per 100,000 inhabitants. There were an estimated 136.7 robberies, 32.2 forcible rapes, and 5.5 murders for each 100,000 resident population in 2004.

Murder and non-negligent manslaughter—Murder. Number of murders and non-negligent manslaughters.

Forcible rape—Rape. Number of forcible rapes.

Burglary—Burglary. Number of burglaries.

Motor vehicle thefts—MotorTheft. Number of motor vehicle thefts.

Full-time law enforcement employees—FTLaw. Number of full-time law enforcement employees.

Performing arts—PerfArts. Number of performing arts companies. Performing arts companies comprise establishments engaged primarily in producing live presentations involving the performances of actors and actresses, singers, dancers, musical groups, and other performing artists.

Museums and historical sites—MusSites. Number of museums or historic sites. Museums are described as establishments engaged primarily in the preservation and exhibition of objects of historical, cultural, and/or educational value. Historical sites are establishments engaged primarily in the preservation and exhibition of sites, buildings, forts, or communities that describe events or persons of particular historical interest.

Spectator sports—Sports. Number of spectator sport teams. Comprises sports teams or clubs participating primarily in live sporting events before a paying audience.

Unemployment rate—Unempl. Unemployment rate (includes some county-level observations).

Household income—Income. Median household income in dollars.

Rentals—Rentals. Number of households that rent.

Management and professional jobs—MgtJobs. Number of jobs in management, professional, and related occupations.

Sales and office jobs—SaleOfficeJobs. Number of jobs in sales and office occupations.

Agriculture and related jobs—AgrJobs. Number of jobs in agriculture, fishing and hunting, and mining.

Construction jobs—ConsJobs. Number of jobs in construction.

Mean travel time to work—TravelTime. Mean travel time to work (in minutes) for workers ages 16 years and older.

Use of public transportation—PublicTrans. People using public transportation for going to work, excluding taxicabs. Workers ages 16 years and older.

County-based measures:

Juvenile crime—CountyJuvCrime. Total number of juvenile crimes committed in the county. Includes violent crimes, property crimes, and nonindex crimes such as alcohol violations, forgery, fraud, and various other crimes.

Death rate—CountyDeath. Death rate in the county (2002 data).

Infant mortality rate—CountyInfant. Infant mortality rate per 100,000 population, 1996–2000.

AIDS cases—CountyAIDS. Number of AIDS cases in the county (2002 data).

Cancer cases—CountyCancer. Number of cancer cases in the county (2002 data).

Match between city and county—CityCountyMatch. Our assessment of how well the county boundaries match city boundaries. 1 = close, 2 = fair, 3 = dissimilar.

Air quality index—Air. The Air Quality Index (AQI) is an index for reporting daily air quality. The Environmental Protection Agency calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide. The AQI runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern.

Air quality data source—AirSource. Some data are from cities; others are from counties. This variable shows the source of the AQI data only. 1 = MSA (metropolitan statistical area), 2 = county.

Counties—Counties. Counties included in the county-based measures.

Watershed

General Description

The Watershed dataset was developed from rating variables, explanations, descriptions, and specific data incorporated within the Environmental Protection Agency's (EPA's) Index of Watershed Indicators. The IWI is a compilation score determined by combining the impact of 16 measured variables used as indicators of general watershed conditions for more than 2,262 watersheds in the United States and Puerto Rico. These indicators measure the general condition of the nation's rivers, streams, wetlands, and estuaries. The website used for developing this dataset is no longer available and has been changed to https://www.epa.gov/waterdata/watershed-index-online.

Background

The 16 variables contained in the IWI are divided into two groups. The first seven variables collectively are termed *condition indicators,* and taken together they represent existing conditions of the specific watershed. (Below are the variable labels. The actual variable names appear later.) These seven condition indicators consist of "designated use attainment," "fish and wildlife consumption advisories," "drinking water impairment," "sediment contamination," "ambient water quality for four toxic pollutants," "ambient water quality for four conventional pollutants," and a "wetlands loss index" developed from two separate wetlands loss inventories. The remaining nine variables comprise the *vulnerability indicators,* which evaluate the potential for future degradation of a particular watershed. These vulnerability indicators are "aquatic/wetlands species at risk," "toxic pollutant loads discharged exceeding aggregate state permit allowance," "conventional pollutant loads discharged exceeding aggregate state permit allowance," "urban runoff potential," "index of agricultural runoff potential," "population change within the watershed," "hydrologic modifications to the watershed," "estuarine pollution susceptibility index," and an "atmospheric deposition factor." Condition indicators and vulnerability indicators are combined, yielding an IWI rating for the specific watershed.

This dataset contains four additional variables: "overall watershed characteristic score," "geographical region," "population of the watershed as of the 1990 national census," and a variable describing the "wetlands loss index."

The 122 watershed observations comprising this dataset consist of two select groups of observations. First, 96 watersheds throughout the United States were specifically identified by the National Sediment Inventory as containing "Areas of Probable Concern." Second, an additional 26 watershed observations—chosen from sparsely populated, remote areas throughout the United States—were selected to provide comparison data for the 96 observations noted as Areas of Probable Concern. It bears comment that finding watersheds in the continental United States without some indicated degree of impairment, or expressed concern relating to their overall rating, is a difficult task. Variable summaries provided below are intended to convey a practical understanding of the nature of each variable and why it constitutes a parameter of concern with regard to overall watershed conditions.

Summary of Individual Variables: Condition Indicators

Each variable name appears after its label.

Watershed characteristic—Wshedch. This overall evaluation describes the condition of aquatic resources for a specific watershed.

Designated use attainment—Useatnmn. This variable is the percentage of water sources within a specific watershed evaluated as meeting the requirements and conditions to fully support the water source's designated uses.

Fish and wildlife consumption advisories—Advisory. When excessive levels of toxic substances are identified within watershed species, states may issue advisories warning against using specific species as a food source. These advisories may be targeted at specific population groups such as the elderly, small children, or pregnant women. This variable is the number of advisories issued within the watershed advising against the consumption of fish or aquatic wildlife. This variable is considered to be a good indicator of the general condition of a watershed and of the extent of toxic substance buildup within a watershed food chain.

Drinking water impairment—Wtrimprd. This variable provides a partial indicator of the condition of water sources within the watershed that could potentially constitute a source of drinking water. These potential sources include both ground and surface water sources sampled prior to treatment or purification for use as drinking water.

Contaminated sediments—Sedicont. This variable is a measure of the potential risks to human health and the environment determined through chemical analysis of bottom sediments, sediment toxicity data, and fish tissue residue data.

Ambient water quality (toxic pollutants)—Toxicon. This variable measures the presence and amount of copper, chromium (hexavalent), nickel, and zinc within watershed water sources. This database continues to the next data layer that provides an actual percentage of samples taken exceeding allowed levels.

Ambient water quality (conventional pollutants)—Conpolut. This variable measures the presence and amount of ammonia, dissolved oxygen, phosphorus, and pH within watershed water sources. This database continues to the next data layer providing an actual percentage of samples taken exceeding allowed levels.

Wetland loss index—Wtlds92 and Wtlnlos2. By combining two indicators, this variable measures the amount of cumulative and continuing wetlands lost. The National Resources Inventory indicates the percentage of watershed wetlands lost during the period of 1982 to 1992, while the National Wetlands Inventory estimates the percentage of watershed wetlands lost from the 1780s to the 1980s. While these two measures are combined to a single index for use within the IWI, this database continues to the next data layer and splits them back out to individual measures, Wtlds92, and Wtlnlos2, respectively. The two individual measures provide a better indication of the extent of historical wetlands loss.

Summary of Individual Variables: Vulnerability Indicators

Aquatic/wetlands species at risk—Speatrsk. This indicator represents the number of species documented in a watershed that are classified by the Heritage Network as being critically imperiled and assesses the conservation of plant and animal species at the greatest risk of extinction within the watershed.

Pollution loads discharged beyond permit limits (toxic)—Toxkdisc. Within the IWI database, this variable is listed as an aggregate index by individual state. If the total toxic pollutant discharges within the state did not exceed the total amount allowed by all permits cumulatively, then the watershed was not considered at risk. This database continues to the next data layer and documents the actual number of incidents in which individual permit discharge limits have been exceeded.

Pollution loads discharged beyond permit levels (conventional)—Convdisc. Within the IWI database, this variable is listed as an aggregate index by individual state. If the total conventional pollutant discharges within the state did not exceed the total amount allowed by all permits cumulatively, then the watershed was not considered at risk. This database continues to the next data layer and documents the actual number of incidents in which individual permit discharge limits have been exceeded.

Urban runoff potential—Urbnrnof. This variable is an indicator of the percentage of a watershed consisting of impervious surfaces (roads, paved parking lots, roofs, and so on). As the amount of impervious surface within a watershed increases as a result of development, natural flow patterns and volumes within streams and rivers can be altered significantly, leading to an increased potential for flooding and an increased potential for runoff of pollutants such as fertilizers and insecticides, in turn leading to further degradation of water sources within the watershed.

Agricultural runoff potential—Agrirnof. Nitrogen runoff potential from fertilizers, pesticide runoff, and sediment delivery to streams and rivers are combined in this index variable. The value for a particular watershed is ranked among 2,110 watersheds evaluated for this variable.

Population change—Popuincr. Comparison of national census data from 1980 and 1990 determined the percentage of population increase within a watershed. Increasing populations place additional burdens on watershed environments through the impact of continued development, increased exploitation of available water sources, further loss of wetlands, increased sewage flow, and additional impervious surfaces.

Hydrological modifications—Hydromod. This index is a measure of the relative amount of reservoir impoundment volume within a watershed. Dams alter natural river and stream flow rates and, when built, can lead to the loss of wetlands. Also, as water accumulates behind the dam, the resulting lake or impoundment can serve as a repository for runoff pollutant introduction, concentration, and subsequent sediment contaminant buildup.

Estuarine pollution susceptibility—Polusucp. Measuring an estuary's susceptibility to contaminant introduction and pollution buildup, this indicator is defined as an estuary's relative vulnerability to concentrations of dissolved and particulate substances. Coastal lands continue to be developed as population densities continue to shift to these regions. It should be noted that a value of 4 for this variable indicates no estuary present within the watershed for evaluation.

Atmospheric deposition—Atmosdep. Nitrogen is a primary nutrient that can cause algae blooms and other problems within watershed water sources. Atmospheric deposition primarily through precipitation is a significant source of nitrogen to these water sources. The units are kilograms per hectare per year. (One kilogram equals approximately 2.2 pounds; one hectare is 100 by 100 meters square, and approximately equal to 2.47 acres.)

Watershed population—Wtrshpop. These are the population figures as of the 1990 census.

Geographical region—Region. This is the location of the watershed within the continental United States:

Northeast: Maine, Massachusetts, New Jersey, New York, Rhode Island
Southeast: Alabama, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee
Midwest: Illinois, Indiana, Kansas, Michigan, Minnesota, Missouri, Ohio, Wisconsin
Northwest: Idaho, Montana, North Dakota, Oregon, South Dakota, Utah, Washington
Southwest: California, Texas

Methodological Note: Classified as "Northeast" are one watershed at the joint boundaries of Pennsylvania, Maryland, and West Virginia; one watershed at the joint boundaries of New York, Ohio, and Pennsylvania; and one watershed at the joint boundaries of Ohio, Pennsylvania, and West Virginia. Classified as "Southeast" is one watershed at the joint boundaries of Arkansas, Kentucky, Mississippi, Missouri, and Tennessee.

Productivity

General Description

This dataset contains hypothetical data about a government organization. It was created to illustrate interrelationships between employee perceptions of their work; organizational environment; and work-center, departmental, and overall organization productivity. Such data are usually obtained through employee surveys and internal administrative reports of departmental productivity.

Background

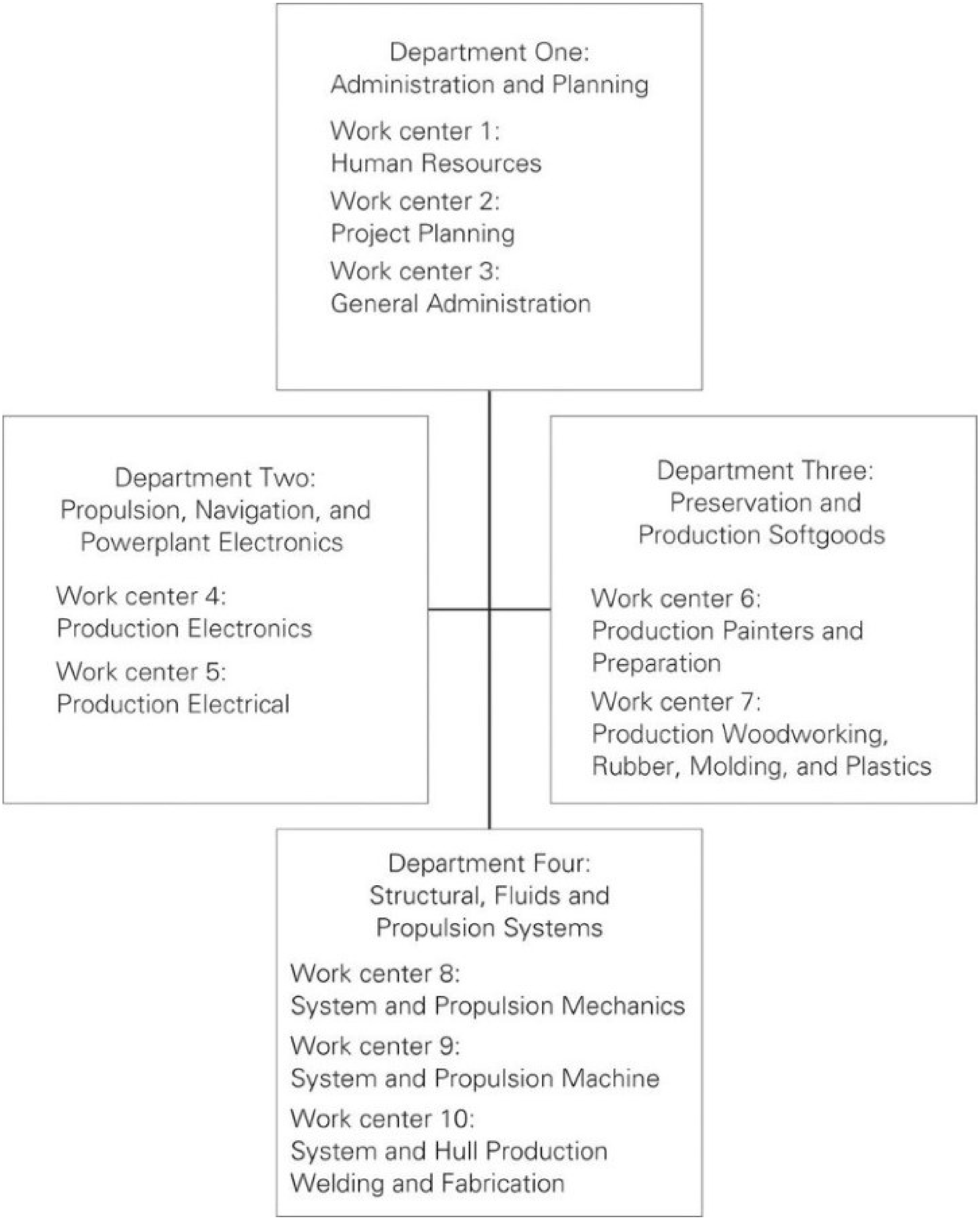
Biltrite Shipbuilding and Marine Repair is a small government shipbuilding and marine vessel repair facility located in Charleston, South Carolina. The organizational workforce comprises primarily federal General Schedule (GS) and Wage Grade (WG) employees. During the 1980s, Biltrite specialized in the manufacture of small coastal patrol and riverine vessels for the United States Army and the United States Coast Guard. During the 1990s the organization was involved primarily with repair of existing craft, with virtually no new construction contracts.

While Biltrite was able to remain economically viable during this period, the organization was subjected to a series of Reduction in Force (RIF) personnel actions. These RIFs resulted in a 31 percent personnel cutback over 7 years. Biltrite accomplished the required cutbacks through normal employee attrition, normal retirement, and early retirement incentives. After an extended period of government defense-spending cutbacks and facilities closures, Biltrite Shipbuilding and Marine Repair now anticipates a resurgence of government military building contracts. To ensure that Biltrite is positioned as competitively as possible, the company reviewed its past productivity evaluations and compared them with those of its competition. While Biltrite's productivity does not lag behind industry averages, Biltrite does not stand out as an industry leader.

Biltrite is now evaluating alternative productivity improvement strategies, which will enhance its opportunities of being awarded future government contracts. As part of this effort, Biltrite contracted with Deeterman & Associates, a recognized leader in productivity studies within the industry. Deeterman evaluated all 10 Biltrite work centers with its four departments (see Figure W20.1).

Deeterman personnel performed the overall organizational study over a period of 2 months. Individual employee interviews were performed during a 2-week period. All 321 Biltrite employees below the senior management level were interviewed. The survey instrument utilized was designed to assess multiple contributing aspects of three index variables and three single concept variables. Additionally, data were collected on length of employment with Biltrite and number of workdays missed because of illness over the past 12 months.

Figure W20.1 Biltrite Organizational Structure


Department One:
Administration and Planning
Work center 1:
Human Resources
Work center 2:
Project Planning
Work center 3:
General Administration
Department Two:
Propulsion, Navigation, and
Powerplant Electronics
Work center 4:
Production Electronics
Work center 5:
Production Electrical
Department Three:
Preservation and
Production Softgoods
Work center 6:
Production Painters and
Preparation
Work center 7:
Production Woodworking,
Rubber, Molding, and Plastics
Department Four:
Structural, Fluids and
Propulsion Systems
Work center 8:
System and Propulsion Mechanics
Work center 9:
System and Propulsion Machine
Work center 10:
System and Hull Production
Welding and Fabrication

Summary of Variables

To measure the different aspects of employee satisfaction, Deeterman personnel used a seven-point Likert scale, ranging from “Strongly Disagree” to “Strongly Agree.” Index variables were formed by averaging the responses to items with strong item-to-total and inter-item correlations. All variables were rescaled on a 10-point scale in which 1 = low and 10 = high.

The 12 variables studied here are listed by variable name and include the variable type and the corresponding statements from the survey.

Jobknowl. Single variable. Type: continuous. The statement evaluated by the respondent: “I have adequate skills and knowledge to perform my job.”

Wkrtrtmt. Single variable. Type: continuous. The statement evaluated by the respondent: “At Biltrite, I am treated fairly and with respect.”

Teamwork. An index variable (Cronbach alpha = 0.81). Type: continuous. The statements evaluated by the respondents: “In my work center, teams are used effectively to accomplish job assignments.” “I am satisfied with the functioning of my work team.” “People from different work centers work well together to accomplish job assignments.” “People from different departments work well together to accomplish job assignments.”

Jobauthr. Single variable. Type: continuous. The statement evaluated by the respondent: “I have adequate decision-making authority to do my job well.”

Recogawd. An index variable (Cronbach alpha = 0.74). Type: continuous. The statements evaluated by the respondents: “On a day-to-day basis I understand what is expected of me by my supervisors.” “In my work center, supervisors’ expectations of worker performance are fair.” “In my department, employee recognition and performance cash incentives are awarded fairly.” “People in other departments don’t have to work as hard as I do to receive the same recognition or performance cash awards.” “Work done excellently is recognized at Biltrite.”

Wrkdyssk. Number of workdays missed because of illness during the past 12 months. Type: continuous.

Inthempl. Length of time as Biltrite employee. Type: continuous.

Workcntr. Employee’s work center. Type: categorical.

Productivity. An index variable (Cronbach alpha = 0.78). Type: continuous. The statements evaluated by the respondents: “In my work center, employees are motivated to work.” “The productivity of my work center is high.” “Work activities are well planned.” “My supervisor is always available to help keep projects on track.” “We are always customer oriented.”

Wkctrpro. Work-center productivity determined by Deeterman & Associates by examining past work-center performance data, including production planning, material procurement, availability of needed material, materials expended, man-hours scheduled and expended per production task, total and job-specific overtime man-hours expended, and percentage of production tasks requiring rework. These values were then weighted and compared with industry standards. Type: continuous.

Dprtment. The department the employee's work center is located within. Type: categorical.

Deptprod. Department productivity determined by Deeterman & Associates. Type: continuous.

Crime

General Description

This time series dataset includes data from a fictional upper midwestern U.S. metropolitan area with a population of approximately 1 million residents. It was created to approximate the juvenile crime rate and trends in that rate, as documented by the Office of Juvenile Justice and Delinquency Prevention (OJJDP) from 1990 to 1998 (*OJJDP Statistical Briefing Book*, which can be found online at www.ojjdp.gov/ojstatbb).

The metropolitan area known as Normalton is representative of national population demographic averages. The juvenile population is estimated by applying age distributions obtained from the *Statistical Abstract of the United States, 1998,* 118th edition. Juvenile arrest data contained within this dataset represent the average number of juvenile criminal arrests per 100,000 juveniles. Juveniles are defined as being male or female, 10 to 17 years of age.

Arrests for juvenile curfew violations are not contained within the juvenile criminal arrest variables. Rather, they are represented within the specific curfew violation arrest variable Curfviol.

Unemployment figures averaged over the year represent national rates obtained from the Bureau of Labor Statistics. A variation was inserted to represent the hiring practices associated with areas that evidence a seasonal tourist economic component.

Background

In June 1993 the city council of Normalton and the newly elected mayor, the Honorable I. M. Worthyman, received from the chief of police the annual report on Normalton crime. The report indicated that, as in the rest of the nation, juvenile crime in Normalton had continued to rise. This report lent further support to an identified upward trend in juvenile crime that began in 1989.

In July 1993 an Australian family was violently attacked and robbed outside their Normalton hotel by a gang of youths. Their daughter was paralyzed as a result of the injuries received during the attack. This family represented only five of the typical 1.2 million tourists drawn to Normalton each spring, summer, and fall by its widely heralded natural beauty, ecological diversity, and three world-class roller coaster theme parks. The case drew national attention.

Feeling that the city must take action to halt what was perceived as an upwardly spiraling juvenile crime rate, the mayor and city council proposed a strict juvenile curfew. Following a survey of city residents, which indicated broad support for the initiative, the city council drafted and unanimously passed a juvenile curfew to take effect on January 1, 1994. The curfew affected all juveniles between the ages of 10 and 17, requiring them to be off the streets between the hours of 8 p.m. and 1 a.m., unless they were in transit to or from an official recreational event or a church function, or were accompanied by an adult. All juveniles were required to be at home after 1 a.m.

The city council also passed a parental responsibility law, whereby parents found negligent in exercising proper control and supervision of their children could be held legally responsible for their children's legal transgressions, including curfew violations. This law took effect on March 1, 1994.

By 1999 there were questions about the effectiveness of the curfew, as the mayor was preparing to run for another term.

Summary of Variables

The listing below shows the variable names, a brief definition of each of the variables, and the variable type. The dataset also includes the first-order differences of these variables, shown as Crimar_1, Juvars_1, Curfi_1, Parenr_1, Curfew_1, Juvnco_1, and Unempl_1.

Crimarst. Number of total juvenile criminal arrests by month from January 1, 1990, through December 31, 1998. Type: continuous.

Juvarsts. Number of monthly juvenile criminal arrests between 8 p.m. and 1 a.m., January 1, 1990, through December 31, 1998. Type: continuous.

Curfviol. Total juvenile arrests for curfew violations, by month. Type: continuous.

Policcur. Number of police per shift dedicated to juvenile code enforcement. Type: continuous.

Month. Month of year. Type: categorical.

Month. A dummy variable of each month. For example, Month1 denotes January, Month2 denotes February, and so forth. Type: categorical.

Befraftr. Number of daily juvenile criminal arrests between 8 p.m. and 1 a.m. 54 days before and 54 days after implementing juvenile curfew restrictions. Type: continuous.

Parenrsp. Number of parents held legally responsible for their children’s violations. Type: continuous.

Unemploy. Percentage of Normalton labor force unemployed. Type: continuous.

Juvncort. Juvenile criminal cases adjudicated in juvenile court. Type: continuous.

Time

General Description

All the variables in the Time dataset are hypothetical. The dataset was developed to illustrate problems relating to simple and multiple regression, as discussed in Chapters 15 and 17 of the textbook. The Time dataset is used for exercises in Chapters 15 and 17 of this workbook. Specifically, these exercises deal with the following problems:

1. Identifying and removing outliers
2. Performing regression with dummy variables
3. Nonlinear relations among variables
4. Autocorrelation

Each problem is associated with a different exercise, and the variables associated with each exercise are defined in the next section.

Summary of Variables

The dataset contains four different sets of variables:

1. **Two variables, Fishcon and Contam1, represent measurements taken from different parts of a large lake in order to test the hypothesis that alleged water pollution is affecting the stock of certain fish.**
   Fishcon = The concentration of fish.
   Contam1 = The concentration of a water pollutant.
2. **The Time dataset also contains observations from 35 hypothetical cities regarding the use of citizen focus groups in various departments (Focus). The data are based on a survey. Most variables are index variables taken from different survey questions. The variables are defined as follows:**
   Focus = A composite measure of the breadth and depth of the use of citizen focus groups in a city. Varies from 0 (low) to 20 (high).
   Mgrint = A measure of the city manager's interest in obtaining citizen-based feedback. Varies from 1 (low) to 4 (high).
   Pubcompl = A measure of public complaints about the quality and effectiveness of a wide range of municipal services. Varies from 1 (low) to 8 (high).
   Budget = Indicates whether municipal budgets have increased in the past 2 years. Values: –1 = decrease in budget; 0 = no change in budget; 1 = increase in budget.
   Size = City size. Varies from 1 = small to 7 = large.
   Region = An indicator variable of the region in which the city is located. Values: 1 = Northeast; 2 = South; 3 = Midwest; 4 = West.
3. **It is commonly hypothesized that crimes are more frequent in large cities. The dataset contains the following variables:**
   Nvcrime = Index of nonviolent crimes in a given year.
   Citysize = City size.
4. **Time series data are common in program evaluation and public policy. These variables examine the impact of a law, which, among other things, increases jail time for driving under the influence (DUI). The dataset contains the following variables:**
   Fatal = Traffic fatalities per 100,000 miles driven.
   Year = Year of traffic fatalities measured.
   Short = A dummy variable identifying when the policy intervention occurred, namely, in 1980 when a law was passed that requires mandatory jail time for DUI. Values: 0 = pre-law adoption (pre-1980), and 1 = post-law adoption (post-1980).
   Long = A dummy variable identifying the number of years of post-policy adoption. Values: 0 = for pre-law adoption, 1 = for 1980, 2 = for 1981, 3 = for 1982, and so forth. This variable gives weight to long-term effects of policy.
   Jailtime = Days of jail time served by offenders because of the law.

Florida County Conservation Spending Database

General Description

This database includes annual spending data on conservation as well as socioeconomic and demographic information of county governments in Florida. The spending data of 1999 to 2008 were collected to examine spending patterns of the government. Included in the spending data are total spending and conservation spending. Conservation spending is the cost associated with controlling pollution and protecting natural resources such as water, air, soil, forest, wildlife, and minerals. Examples include purchases of environmentally sensitive lands, monitoring and control of air and water pollution, and the creation and restoration of natural habitats for wildlife. Florida has 67 counties, but Duval County government was consolidated with the city of Jacksonville and was therefore excluded from this base.

The socioeconomic and demographic variables in this database include poverty rate, household income, vote for Democratic presidential candidates in elections, education, population, land size, manufacturing employment, farm earing, farm land size, water use, costal status, and some other variables.

Background

The database was used to examine socioeconomic or demographic impact on conservation spending. Data were collected from Florida Local Government Electronic Reporting (LOGER) System from Florida Department of Financial Services, Comprehensive Annual Financial Reports, and the U.S. Census.

Summary of Variables

1. Variable Name “County”: County Name
2. Variable Name “Cons99”: 1999 Conservation Spending. The same coding scheme is used for “Cons00” (2000 Conservation Spending), “Cons01” . . . and “Cons08.”
3. Variable Name “Tot99”: 1999 Total County Spending. The same coding scheme is for “Tot00” . . . and “Tot08.”
4. Variable Name “Vote2008”: Percent vote in 2008 election for the Democratic Presidential Candidate Obama
5. Variable Name “Vote2004”: Percent vote in 2004 election for the Democratic Presidential Candidate Kerry
6. Variable Name “Vote2000”: Percent vote in 2000 election for the Democratic Presidential Candidate Gore
7. Variable Name “Land”: Size of land in the county by square mile
8. Variable Name “Pop08”: Estimated County Population 2008. The same coding is used for “Pop07” (Estimated County Population 2007) . . . and “Pop00.”
9. Variable Name “ManufacturingFirms”: Number of Manufacturing Firms in 2002.
10. Variable Name “ManuEmployment”: Number of Manufacturing Employees in 2005
11. Variable Name “FarmEarn”: Farm Earning in 2005
12. Variable Name “FarmLand”: Total Acres of farm land in 2002
13. Variable Name “WaterTotal”: Total Water Use in million gallons per day in 2000
14. Variable Name “WaterUsePerCapita”: Total Water Use per resident in million gallons in 2000
15. Variable Name “Coastal”: Whether a county is coastal. 1 = yes. 0 = no.
16. Variable Name “MetroStatus”: Whether a county is granted a metro status in the Census. 1 = yes. 0 = no.
17. Variable Name “HighSchool”: Percent of high school graduates
18. Variable Name “College”: Percent of college graduates
19. Variable Name “Income”: Percent of Household income $75,000 or more in 2000
20. Variable Name “Poverty04”: Poverty Rate in 2004

CQ Press, an imprint of SAGE, is the leading publisher of books, periodicals, and electonic products on American government and international affairs. CQ Press consistently ranks among the top commercial publishers in terms of quality, as evidenced by the numerous awards its products have won over the years. CQ Press owes its existence to Nelson Poynter, former publisher of the St. Petersburg Times, and his wife Henrietta, with whom he founded Congressional Quarterly in 1945. Poynter established CQ with the mission of promoting democracy through education and in 1975 founded the Modern Media Institute, renamed The Poynter Institute for Media Studies after his death. The Poynter Institute (www.poynter.org) is a nonprofit organization dedicated to traning journalists and media leaders.

In 2008, CQ Press was acquired by SAGE, a leading international publisher of journals, books, and electronic media for academic, educational, and professional markets. Since 1965, SAGE has helped inform and educate a global community of scholars, practitioners, researchers, and students spanning a wide range of subject areas, including business, humanities, social sciences, and science, technology, and medicine. A privately owned corporation, SAGE has offices in Los Angeles, London, New Delhi, Singapore, and Boston in addition to the Washington DC office of CQ Press.