The Handbook of

Data Analysis

Paperback Edition

Edited by

Melissa Hardy and
Alan Bryman

Handbook of Data Analysis

Advisory Board for the Handbook of Data Analysis

Handbook of Data Analysis

Edited by

Melissa Hardy and Alan Bryman

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

First published 2004
This paperback edition first published 2009

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
33 Pekin Street #02-01
Far East Square
Singapore 048763

Library of Congress Control Number: 2008926859

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-84860-116-1 (pbk)

Typeset by C&M Digitals (P) Ltd., Chennai, India
Printed in India at Replika Press Pvt Ltd

Contents

Preface

As is the case with any edited text, this book represents the culmination of exchanges with authors past and present. We are fortunate to have persuaded so many well-established data analysts to contribute chapters. Their investment of time and thought is reflected in the quality of the discussions that fill these pages. We are most appreciative of the support and assistance we received from Sage and would like to give special thanks to Chris Rojek, Kay Bridger and Ian Antcliff. We would like to thank Richard Leigh for his meticulous copyediting, which has greatly improved the book. We would also like to thank the members of our Advisory Board and several colleagues who provided us with advice on chapters, Chardie Baird who helped manage the multiple drafts and reviews, and our spouses for their support and encouragement.

Our intention was to put together a set of resource chapters that described major techniques of data analysis and addressed noteworthy issues involved in their application. The list of techniques included here is not exhaustive, but we did try to cover a wide range of approaches while providing reference to an even broader set of methods. With that in mind, we decided to include techniques appropriate to data of different sorts, including survey data, textual data, transcripts of conversations, and longitudinal information. Regardless of the format of the original data, analysis requires researchers to develop coding schemes, classification protocols, definitional rules, and procedures for ensuring reliability in the application of all of these tools. How researchers organize the information they will use in their analyses should be informed by theoretical concerns. Even so, this process of organization is also one of creation and, as such, it can be accomplished in a variety of ways and analyzed by different approaches.

Data analysts must concern themselves with the criteria they use to sort between the systematic component of their observations and the stochastic elements, or random influences, that are also reflected in these observations. The randomness of events is something we acknowledge, but we often behave as though we can exert considerable control over the way our lives unfold.

That point is often driven home in unanticipated ways. During the time we dedicated to the production of this book, we made frequent adjustments to modify a once reasonable schedule that had become impossible to meet. These unanticipated events reflect the fabric of people's lives, and forecasting life's events that would occur a year or two into the future was sometimes tragically inaccurate. Prominent among our initial list of authors were Lee Lillard and Aage Sørensen, both greatly respected by the scientific community, admired by their peers, and loved by their friends and families. Both men died unexpectedly while this volume was under way. We make note here of the substantial contributions they made to this field of inquiry and to this volume through their published work, their teaching, and their involvement in too many discussions of these issues to count.

Melissa Hardy and Alan Bryman

Notes on Contributors

Andrew Abbott is the Gustavus F. and Ann M. Swift Distinguished Service Professor in the Department of Sociology and the College at the University of Chicago. Known for his ecological theories of occupations, Abbott has also pioneered algorithmic analysis of social sequence data. His recent books include studies of academic disciplines and publication (*Department and Discipline*, 1999) and of fractal patterns in social and cultural structures (*Chaos of Disciplines*, 2001), as well as a collection of theoretical and methodological essays in the Chicago pragmatist and ecological tradition (*Time Matters*, 2001). He is currently writing on social science heuristics and developing a major research project on the life course. Abbott is also a past Editor of the *American Journal of Sociology*.

Paul Allison is Professor of Sociology at the University of Pennsylvania where he teaches graduate methods and statistics. He is the author of *Missing Data* (2001), *Logistic Regression Using the SAS® System* (1999), *Multiple Regression: A Primer* (1999), *Survival Analysis Using the SAS® System* (1995), *Event History Analysis* (1984), and numerous articles on regression analysis, log-linear analysis, logit analysis, latent variable models, missing data, and inequality measures. He is a member of the editorial board of *Sociological Methods and Research*. A former Guggenheim Fellow, Allison received the 2001 Lazarsfeld Award for distinguished contributions to sociological methodology.

Douglas L. Anderton is Professor of Sociology at the University of Massachusetts Amherst, Director of the Social and Demographic Research Institute and a Fellow of the American Statistical Association. His research emphasizes quantitative-historical analysis of population and environment interactions. He is the author of over fifty journal articles and has co-authored several books, including: *Demography: Study of Human Populations* (2001), *Population of the United States* (1998), *Fertility on the Frontier* (1993), and an edited series, *Readings in Population Research Methodology* (1997). He is currently an editor of the ASA Rose Monograph Series in policy studies.

Paul Atkinson is Research Professor in Sociology at Cardiff University, UK. He is co-director of the ESRC Research Centre on Social and Economic Aspects of Genomics. His main research interests are the sociology of medical knowledge and the development of qualitative research methods. His publications include: *Ethnography: Principles in Practice* (with Martyn Hammersley), *The Clinical Experience, The Ethnographic Imagination, Understanding Ethnographic Texts, Medical Talk and Medical Work, Fighting Familiarity* (with Sara Delamont), *Making Sense of Qualitative Data* (with Amanda Coffey), *Sociological Readings and Re-readings, Interactionism* (with William Housley) and *Key Themes in Qualitative Research* (with Sara Delamont and Amanda Coffey). He was one of the editors of the *Handbook of Ethnography*. Together with Sara Delamont he edits the journal *Qualitative Research*. His recent ethnographic study of an international opera company will be published in 2004 as *Everyday Arias*.

Peter M. Bentler is Professor of Psychology and Statistics at University of California, Los Angeles, and former Chair of the Department of Psychology, and has over 400 publications in methodology, psychometrics, and statistics as well as in applied fields such as personality, attitudes, drug abuse, health, sexuality and related topics. He has been an elected president of the Society of Multivariate Experimental Psychology, the Psychometric Society, and the Division of Evaluation, Measurement, and Statistics of the American Psychological Association (APA). He is also a recipient of the Distinguished Scientific Contributions Award from the APA Division of Evaluation, Measurement, and Statistics.

Ronald L. Breiger is Professor of Sociology, University of Arizona. With Linton Freeman, he edits the journal *Social Networks*. His interests include social network analysis, stratification, mathematical models, theory, and measurement issues in cultural and institutional analysis. He has recently written with Philippa Pattison on lattices and dimensional representation of network structures, with David Stark and Szabolcz Kemény on ownership patterns among banks and firms in Hungary, and with John Mohr on the dual aggregation of social categories.

William Browne is Professor of Biostatistics, School of Clinical Veterinary Science, University of Bristol.

Alan Bryman is Professor of Organisational and Social Research in and currently Head of the School of Management, University of Leicester. His main research interests lie in research methodology, leadership, organizational analysis, and Disneyization. He is author or co-author of many books, including: *Quantity and Quality in Social Research* (Routledge, 1988), *Social Research Methods* (OUP, 2001, 2004, 2008), *Business Research Methods* (OUP, 2003, 2007), and *Disneyization of Society* (Sage, 2004). He is co-editor of *The SAGE Encyclopedia of Social Science Research* (Sage, 2004 and the forthcoming *Sage Handbook of Organizational Research Methods* (Sage, 2008).

Eric Cheney is a Ph.D. candidate at the University of Massachusetts Amherst. His research interests include economic sociology, organizations, social networks, statistics, and quantitative methodology. He is currently completing his dissertation on the topic of social structure and economic exchange.

Simon Cheng is an Assistant Professor in the Department of Sociology at the University of Connecticut. His substantive research is in race and ethnicity, family–school relationships, and political-economic development, where he has recently published articles in *Sociology of Education*, *Social Forces*, and other journals. His dissertation, 'Standing in the middle of interracial relations: The educational experiences of children from multiracial backgrounds', examines differences between biracial and monoracial families in a variety of family and student outcomes. He is also studying the small-sample behavior of tests of the independence of irrelevant alternatives assumption in the multinomial logit model.

Steven E. Clayman is Professor of Sociology and is affiliated with the Communication Studies Program at the University of California Los Angeles. His research concerns the intersection of talk, interaction, and mass communication. He has studied broadcast news interviews, presidential press conferences, newspaper editorial conferences, the dynamics of quotability, and collective audience behavior in political speeches and debates. His articles have appeared in *American Sociological*

Review, American Journal of Sociology, Language in Society, Journal of Communication, and *Media, Culture, and Society*. He is the author (with John Heritage) of *The News Interview: Journalists and Public Figures on the Air*.

Duncan Cramer is Professor of Psychological Health in the Department of Social Sciences at Loughborough University in England. He received his doctorate in 1973 from the Institute of Psychiatry in London. His main research interests and publications lie in the fields of close relationships, personality, psychological health, counselling and psychotherapy. He has authored and co-authored a number of books including *Personality and Psychotherapy* (1992, Open University Press), *Close Relationships* (1998, Arnold) and *Advanced Quantitative Data Analysis* (2003, Open University Press). Currently he is an Associate Editor of *Psychology and Psychotherapy* (2002–) and the *Journal of Social and Personal Relationships* (2004–). Previously he has been an Associate Editor (1993–1995, 2000–2001) and a Joint Editor (1995–2000) of the *British Journal of Medical Psychology*.

Barbara Czarniawska is Professor of Management Studies at Gothenburg Research Institute, School of Business, Economics and Law at University of Gothenburg, Sweden. Her research applies a constructionist perspective on organizing, with the focus on action nets. Her methodological interests concern fieldwork techniques and the application of narratology to social science studies. Recent books in English: *A Tale of Three Cities* (2002), *Narratives in Social Science Research* (2004), *Shadowing and Other Techniques for Doing Fieldwork in Modern Societies* (2007), and *A Theory of Organizing* (2008). She has recently edited *Actor-Network Theory and Organizing* with Tor Hernes (2005), *Global Ideas* with Guje Sevón (2005), *Organization Theory* (2006) and *Management Education & Humanities* with Pasquale Gagliardi (2006).

Sara Delamont is Reader in Sociology at Cardiff University. She was the first woman to be president of the British Education Research Association, and the first woman Dean of Social Sciences at Cardiff. Her research interests are educational ethnography, Mediterranean anthropology, and gender. Her most famous book is *Interaction in the Classroom* (1976 and 1983), her favourites *Knowledgeable Women* (1989) and *Appetites and Identities* (1995). She is co-editor of the journal *Qualitative Research*.

Tonya Dodge is a pre-doctoral student in the Psychology Department at the University of Albany, State University of New York. Her research interests focus on attitudes and decision-making, with a particular interest in attitude ambivalence. She also conducts research on the effects of athletic participation on adolescent risk behavior.

Nigel G. Fielding is Professor of Sociology and Associate Dean of Arts and Human Sciences at the University of Surrey. With Ray Lee, he co-directs the CAQDAS Networking Project, which provides training and support in the use of computers in qualitative data analysis. His research interests are in new technologies for social research, qualitative research methods, and mixed method research design. He has authored or edited 20 books, over 50 journal articles and over 200 other publications. In research methodology his books include a study of methodological integration (*Linking Data*, 1986, Sage; with Jane Fielding), an influential book on qualitative software (*Using computers in qualitative research*, 1991, Sage; editor, with Ray Lee), a study of the role of computer technology in qualitative research (*Computer Analysis and Qualitative Research*, 1998, Sage, with Ray Lee) and a four volume set, *Interviewing* (2002, Sage; editor). He is presently researching the application of high performance computing applications to qualitative methods.

Roberto P. Franzosi is Professor of Sociology and Linguistics at Emory University, having previously taught at the University of Wisconsin Madison (1983–93), Rutgers University (1994–95), the University of Oxford (1995–1999), and the University of Reading (1999–2006). Franzosi's long-standing research interest has been in the area of social conflict with several articles and a book (*The Puzzle of Strikes: Class and State Strategies in Postwar Italy*, Cambridge UP, 1995). Since the early 1980s Franzosi has been involved in the development of a new linguistic- and computer-based approach to content analysis applied to the study of historical processes, publishing several articles and two books (*From Words to Numbers*, Cambridge UP, 2004; *Quantitative Narrative Analysis*, Sage, 2009). He has edited the 4-volumes book Content Analysis for Sage Benchmark in Social Research Methods Series (2008). He is currently completing a book on *Quantitative Narrative Analysis* for Sage Quantitative Applications in the Social Sciences.

Vincent Kang Fu is Assistant Professor in the Department of Sociology at the University of Utah. He earned a PhD from UCLA in 2003. He is a social demographer studying racial and ethnic inequality in the United States. His research examines the incidence and consequences of racial and ethnic intermarriage. Past work has investigated the relationship between education and intermarriage, the effects of age and remarriage on intermarriage, as well as trends and geographic variation in intermarriage. Current and future work includes studies of the impact of intermarriage on fertility and divorce, and the incidence of intermarriage in Brazil. Other interests include immigration, demographic methods, and statistics.

Virginia Teas Gill is an Associate Professor in the Department of Sociology and Anthropology at Illinois State University. She received her Ph.D. from the University of Wisconsin, Madison. Her research focuses on the organization of social interaction in clinical settings, especially processes of persuasion and resistance in regard to diagnoses, labels, and medical interventions. It includes studies of physician-patient interaction in various primary care settings and clinician-parent interaction in a childhood developmental disabilities clinic. Her work has been published in journals such as *Social Psychology Quarterly, Social Problems*, and *Research on Language and Social Interaction*. She is co-editor (with Alison Pilnick and Jon Hindmarsh) of *Communication in Healthcare Settings: Policy, Participation and New Technologies* (forthcoming).

Guang Guo is Professor of Sociology, Department of Sociology, University of North Carolina at Chapel Hill. He has published methodological work in the *Journal of the American Statistical Association, Sociological Methodology, Annual Review of Sociology*, and *Behavior Genetics* on event-history analysis, multilevel analysis, and random-effects models for genetically related siblings. He has published substantive work in *American Sociological Review, Social Forces, Sociology of Education, Demography*, and *Population Studies* on child and infant mortality, poverty and children's intellectual development, sibsize and intellectual development, and heritability–environment interactions for intellectual development. His current interest lies mainly in the interactions between environment and heritability using both sibling and DNA data.

Melissa Hardy is Distinguished Professor of Human Development, Sociology and Demography and Director of the Gerontology Center at the Pennsylvania State University. She received her graduate degree in sociology from Indiana University in 1980. Her current research focuses on pensions and financial security in old age, health disparities, socio-political attitudes and inequality. She is the author of *Regression with Dummy Variables* (1993), editor of *Studying Aging and Social Change: Conceptual and Methodological Issues* (1997), and co-author of *Ending a Career in the Auto Industry: 30 and Out* (1997), 'Population Aging, Intracohort Aging, and Socio-Political Attitudes' in *American Sociological Review* (2007) and *Pension Puzzles: Social Security and the Great Debate* (2007).

Lawrence Hazelrigg is now Emeritus Professor, College of Social Sciences, Florida State University. Known for his research in stratification and his writing in social theory, he is currently engaged in a study of historical-cultural differences of the constitution and practice of selfhood. Recent and representative publications include: *Cultures of Nature* (1995), 'Individualism', in *Encyclopedia of Sociology*, 2nd edition (2000), 'Scaling the Semantics of Satisfaction' in *Social Indicators Research* (2000), 'Fueling the Politics of Age' in *American Sociological Review* (1999), 'Marx and the Meter of Nature' in *Rethinking Marxism* (1993), *Pension Puzzles: Social Security and the Great Debate* (2007), and *Hazelrigg Family History: North America c1635 to 1935* (2007).

Karen Henwood is a Reader in Social Sciences at the School of Social Sciences, Cardiff University. A social psychologist by training, her research addresses the role of culture, difference and life history in the formation of identity and subjectivity. She has undertaken research projects on adult–mother daughter relationships, masculinity and the body, and the meanings and non-economic values people attach to their natural environment. With Nick Pidgeon she has explored the role of qualitative methods in psychology and the social sciences. Her published work has appeared in journals such as the *British Journal of Social Psychology*, *British Journal of Psychology*, *Feminism and Psychology*, *Theory and Psychology*, *Journal of Environmental Psychology*, and *Social Science and Medicine*. Further collaborative work has appeared in practitioner journals such as the *Journal of Public Mental Health* and *BMJ*. She has completed ESRC funded projects on 'masculinities, identities and the transition to fatherhood', 'a consultation on qualitative research resources', and 'gender and risk perception: a secondary analysis'. She is currently involved in research exploring the role of narrative methodology and methods in researching risk, as part of the ESRC's 'Social contexts and responses to risk' (SCARR) network (2003–2008). From February 2007 she is investigating 'masculinities, identities and risk' as part of the major five year (Timescapes) network: 'Changing Lives and Times: Relationships and Identities across the Lifecourse', a network which is also showcasing qualitative longitudinal (QL) research.

John Hipp is an Assistant Professor in the departments of Criminology, Law and Society, and Sociology, at the University of California Irvine. His research interests focus on how neighbourhoods change over time, how that change both affects and is affected by neighbourhood crime, and the role networks and institutions play in that change. He approaches these questions using quantitative methods as well as social network analysis. He has published substantive work in such journals as *American Sociological Review*, *Criminology*, *Social Forces*, *Social Problems*, *Mobilization*, *Health & Place*, *City & Community*, *Crime & Delinquency*, *Urban Studies* and *Journal of Urban Affairs*. He has published methodological work in such journals as *Sociological Methodology*, *Psychological Methods*, and *Structural Equation Modeling*.

James Jaccard is Distinguished Professor of Psychology at the University at Albany, State University of New York. His primary research interest is in adolescent risk behavior, with an emphasis on understanding adolescent unintended pregnancy and adolescent drunk driving. His work has focused on family-based approaches to dealing with adolescent problem behaviors. He has authored four monographs on the analysis of interaction effects using a wide range of statistical models.

Mortaza (Mori) Jamshidian is Professor of Mathematics and Statistics at California State University, Fullerton. His main research area is computational statistics. He has made significant contributions in the area of EM estimation which has numerous applications in missing-data analysis. In particular his papers on acceleration of EM (JASA 1993, JRSS-B 1997) and EM standard error estimation (JRSS-B 2000) are important contributions. His other contributions include development of statistical methodological and computing algorithms in the fields of psychometrics, biostatistics, and general statistics.

Raymond M. Lee is Professor of Social Research Methods in the Department of Health and Social Care at Royal Holloway University of London. He has written extensively about a range of methodological topics, including the problems and issues involved in research on ‘sensitive’ topics, research in physically dangerous environments, the role of new technologies in the research process, and the history of the interview. He co-ordinates the UK Economic and Social Research Council’s Researcher Development Initiative, a nationwide programme to develop an advanced training infrastructure for social researchers.

J. Scott Long is Chancellor’s Professor of Sociology at Indiana University, Bloomington. His research focuses on gender differences in the scientific career, aging and labor force participation, and statistical methods. His recent research on the scientific career was published as *From Scarcity to Visibility.* He is past Editor of *Sociological Methods and Research* and the recipient of the American Sociological Association’s Paul F. Lazarsfeld Memorial Award for Distinguished Contributions in the Field of Sociological Methodology. He is author of *Confirmatory Factor Analysis, Covariance Structure Analysis, Regression Models for Categorical and Limited Dependent Variables,* and *Regression Models for Categorical and Limited Dependent Variables with Stata* (with Jeremy Freese), as well as several edited volumes.

Heather MacIndoe is Assistant Professor in the Department of Public Policy and Public Affairs at the University of Massachusetts Boston. She received her Ph.D. in Sociology from the University of Chicago in 2007. Her research interests include formal organizations, philanthropy, social change, and civil society. Her current research examines the relationships between philanthropic foundations and nonprofit organizations in American cities and how the use of performance/outcome measurement by nonprofit organizations influences organizational practice.

Peter K. Manning (Ph.D. Duke, 1966, MA Oxon. 1982) holds the Elmer V. H. and Eileen M. Brooks Chair in the College of Criminal Justice at Northeastern University, Boston, MA. He has taught at Michigan State, MIT, Oxford, and the University of Michigan, and was a Fellow of the National Institute of Justice, Balliol and Wolfson Colleges, Oxford, the American Bar Foundation, the Rockefeller Villa (Bellagio), and the Centre for Socio-Legal Studies, Wolfson College, Oxford. Listed in *Who’s Who in America*, and *Who’s Who in the World*, he has been awarded many contracts and grants, the Bruce W. Smith and the O.W. Wilson Awards from the Academy of Criminal Justice Sciences, and the Charles Horton Cooley Award from the Michigan Sociological Association. The author and editor of some 15 books, including *Privatization of Policing: Two Views* (with Brian Forst) (Georgetown University Press, 2000), his research interests include the rationalizing and interplay of private and public policing, democratic policing, crime mapping and crime analysis, uses of information technology, and qualitative methods. His most recent publications include *Policing Contingencies* (University of Chicago Press, 2003) and *The Technology of Policing: crime mapping, information technology and the rationality of crime control* (NYU Press, 2008).

Robert D. Mare is Distinguished Professor of Sociology at the University of California Los Angeles. His research interests include social mobility and inequality, demography, and quantitative methods. He has done extensive research on intergenerational educational mobility, assortative mating, youth unemployment, and methods for the analysis of categorical data. His recent work focuses on models for residential mobility and residential segregation and on links between educational stratification and marriage markets. His research has appeared in a number of journals, including the *American Sociological Review, American Journal of Sociology,* and *Sociological Methodology*. He is a former editor of *Demography*.

Mary Maynard is Professor in the Department of Social Policy and Social Work at the University of York, UK, where she was previously Director of the Centre for Women's Studies. She works and writes in the areas of gender, ethnicity, later life, social theory, and social research methodology. She has just completed a project focusing on what empowers older women, from a variety of ethnic groups, in later life.

Trond Petersen is a Professor at the University of California Berkeley, in the Department of Sociology and Haas School of Business. His research and teaching are in the areas of inequality and social stratification, organizations, human resource management, economic sociology, and quantitative methods. He has also taught at the University of Oslo and previously at Harvard University. Among his recent publications are: 'Offering a job: Meritocracy and social networks' in *American Journal of Sociology* (2000, with Ishak Saporta and Marc-David Seidel), 'Equal pay for equal work? Evidence from Sweden and a comparison with Norway and the U.S.' in *Scandinavian Journal of Economics* (2001, with Eva M. Meyersson Milgrom and Vemund Snartland), and 'The opportunity structure for discrimination' (with Ishak Saporta), to appear in *American Journal of Sociology*.

Nick Pidgeon is Professor of Applied Psychology at Cardiff University. He has substantive academic and public policy research interests in risk and its management, risk perception and risk communication in relation to a range of technological and environmental controversies, including: nuclear power, biotechnology and nanotechnologies. His most recent research is looking at societal responses to climate risks, and the preconditions for developing sustainable energy futures. As a part of this he has also worked on issues of public participation and engagement in relation to technological controversies. In methodological terms he has expertise in both survey research and qualitative methods, and where appropriate utilizes mixed methods approaches in his own research work. He was elected a Fellow of the Society for Risk Analysis (international) in 2003 and was a member of the Royal Society working group on nanotechnologies which published its influential report in 2004. He has published over 100 journal articles and scholarly papers on risk issues and is co-author of *Man-Made Disasters* (2nd Edn, with B.A. Turner 1997), *The Social Amplification of Risk* (with R. Kasperson and P. Slovic, 2003) and *The GM*

Debate: Risk, Politics and Public Deliberation (with T. Horlick-Jones, J. Walls, G. Rowe, W. Poortinga and T. O'Riordan).

Jonathan Potter is Professor of Discourse Analysis at Loughborough University. He has studied scientific argumentation, descriptions of crowd disorder, current affairs television, racism, and relationship counselling, and is currently studying calls to a child protection helpline. His most recent books include *Representing Reality* (1996), which attempts to provide a systematic overview, integration and critique of constructionist research in social psychology, postmodernism, rhetoric, and ethnomethodology, *Focus Group Practice* (Sage, 2003, with Claudia Puchta), which analyses interaction in market research focus groups, and Conversation and Cognition (CUP, 2005, with Hedwig te Molder) in which a range of different researchers consider the implication of studies of interaction for understanding cognition. He is co-editor of the journal *Theory and Psychology*.

Jon Rasbach is now Professor in Graduate School of Education, University of Bristol.

John R. Reynolds is Associate Professor of Sociology and an associate of the Claude Pepper Institute on Aging and Public Policy at Florida State University. His current research seeks to explain broad shifts in achievement ideology, such as the trend toward increasingly unrealistic career plans among teenagers, and to explicate the long-term consequences of such trends for gender, race, and class inequalities. Recent publications include 'Have adolescents become too ambitious? High school seniors' educational and occupational plans, 1976 to 2000' in *Social Problems*, 'Educational expectations and the rise in women's post-secondary attainments' in *Social Science Research*, and 'Mastery and the fulfillment of occupational expectations' to appear in *Social Psychology Quarterly*.

Dennis Smith is Professor of Sociology in the Department of Social Sciences at Loughborough University. He has written several articles and books, including *The Rise of Historical Sociology* (1990), *Zygmunt Bauman: Prophet of Postmodernity* (1999), *Norbert Elias and Modern Social Theory* (2000), *Conflict and Compromise: Class Formation in English Society 1830–1914* (1982), *Capitalist Democracy on Trial: The Transatlantic Debate from Tocqueville to the Present* (1991), *Barrington Moore: Violence, Morality and Political Change* (1983), *The Chicago School: A Liberal Critique of Capitalism* (1988), and has been a contributing editor of *Whose Europe? The Turn towards Democracy* (1999), *The Civilized Organization* and *Norbert Elias and the Future of Organization Studies* (2002). His latest book is entitled *Globalization: The Hidden Agenda* (2006). He is a past vice-president of the European Sociological Association (2001–3), one-time editor of *Sociological Review* and currently editor of *Current Sociology*, journal of the International Sociological Association.

Michael Sobel is a professor at Columbia University. His research interests include causal inference and new applications of the theory of financial decision-making to the social sciences. He is a previous co-editor of *Sociological Methodology* and a co-editor of the *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (1995).

Ross M. Stolzenberg is Professor of Sociology at the University of Chicago. He was a recent editor of *Sociological Methodology*, and his current research concerns the connection between family and labor market processes in stratification systems. He has held academic posts in university programs in social relations, sociology, population dynamics, and applied statistics at Harvard University, Johns Hopkins University and the University of Illinois at Urbana. He has held nonacademic posts at the Rand Corporation, the Graduate Management Admission Council, and as a consultant in complex litigation and other matters. He has served on editorial boards or held editorial postions at seven refereed academic journals.

Nancy Brandon Tuma obtained her Ph.D. in sociology from Michigan State University in 1972, and is currently a Professor of Sociology at Stanford University. She is a leading sociological methodologist, focusing primarily on the study of change. Best known for her pioneering work on event-history analysis and as co-author of *Social Dynamics: Models and Methods* (1984), she has published studies of life careers and social inequalities in the United States, China, Germany, Poland, the former Soviet Union, and various countries formerly in the Soviet Union. Her current primary research interest is the impact of the transition from socialism on people's life careers. She has served as editor of *Sociological Methodology* and also as associate editor of the *Journal of the American Statistical Association*. In 1994 she received the Lazarsfeld award for her contributions to sociological methodology.

Jodie B. Ullman is an Associate Professor of Psychology at California State University, San Bernardino. She earned her Ph.D. in measurement and psychometrics from the University of California Los Angeles in 1997. Her primary research interests are in applied multivariate statistics with a particular emphasis on structural equation modeling and hierarchical linear modeling. She is particularly interested in applications of complex statistical techniques to substance use questions. Her recent research includes evaluations of the Drug Abuse Resistance Education program, longitudinal examinations of cigarette sales to minors, and reduction of HIV/AIDS risk behaviors in homeless populations.

Christopher Winship is Professor of Sociology in the Kennedy School of Government at Harvard University. He did his undergraduate work in sociology and mathematics at Dartmouth College and received his graduate degree from Harvard in 1977. He is currently doing research on the Ten Point Coalition, a group of black ministers who are working with the Boston police to reduce youth violence; statistical models for causal analysis; the effects of education on mental ability; causes of the racial difference in performance on elite colleges and universities; and changes in the racial differential in imprisonment rates over the past sixty years.

'With the appearance of this handbook, data analysts no longer have to consult dozens of disparate publications to carry out their work. The essential tools for an intelligent telling of the data story are offered here, in thirty chapters written by recognized experts. While quantitative methods are treated, from basic statistics through the general linear model and beyond, qualitative methods are by no means neglected. Indeed, a unique feature of this volume is the careful integration of quantitative and qualitative approaches.' **Michael S. Lewis-Beck, F. Wendell Miller Distinguished Professor of Political Science at the University of Iowa.**

This book, which will rapidly be recognized as a social research bible, provides a peerless guide to key issues in data analysis, from fundamental concerns such as the construction of variables, the characterization of distributions and the notions of inference, to the more advanced topics of causality, models of change and network analysis.

No other book provides a better one-stop account of the field of data analysis. Throughout, the editors encourage readers to develop an appreciation of the range of analytic options available for a wide variety of data structures, so that they can develop a suitable analytic approach to their research questions.

Scholars and students can turn to it for teaching and applied needs with confidence, while specialists will find the provision of up to date expositions on a wide range of techniques invaluable.

Melissa Hardy is Distinguished Professor of Human Development and Family Studies, Sociology, and Demography and Director of the Gerontology Center at The Pennsylvania State University; Alan Bryman is Professor of Social Research, University of Loughborough.

'The book provides researchers with guidance in, and examples of, both quantitative and qualitative modes of analysis, written by leading practitioners in the field. The editors give a persuasive account of the commonalities of purpose that exist across both modes, as well as demonstrating a keen awareness of the different things that each offers the practising researcher.' **Clive Seale, Department of Sociology, Goldsmiths College, London.**

'This is an excellent guide to current issues in the analysis of social science data. I recommend it to anyone who is looking for authoritative introductions to the state of the art. Each chapter offers a comprehensive review and an extensive bibliography and will be invaluable to researchers wanting to update themselves about modern developments.' **Professor Nigel Gilbert, Pro Vice-Chancellor and Professor of Sociology, University of Surrey.**

1

Introduction

Common Threads among Techniques of Data Analysis

MELISSA HARDY AND ALAN BRYMAN

In deciding the mix of topics to include in this *Handbook*, we wanted to provide a wide range of analytic options suited to many different research questions and different data structures. An early decision was to include both 'quantitative' and 'qualitative' techniques in a single volume. Within the current research environment, practitioners can hardly fail to notice the schism that exists between camps of qualitative and quantitative researchers. For some, this division is fundamental, leading them to pay little attention to developments in the 'other' camp. Certainly the assumption has been that practitioners of these different approaches have so little in common that any text on data analysis must choose between the two approaches rather than include both in a single text.

We believe that reinforcing this division is a mistake, especially for those of us who practice in the behavioral and social sciences. Discipline boundaries too often act as intellectual fences beyond which we rarely venture, as if our own field of research is so well defined and so much ours that we can learn nothing from other disciplines that can possibly be of use. Many of us may remember our first forays into literature searches on a given research topic, which we too often defined in the narrowest of terms, only to learn from our advisors that we had missed mountains of useful publications arrayed across a variety of fields, time periods, and (perhaps) languages. One of the major costs of dividing and subdividing fields into an increasing number of specializations is that we may inadvertently limit the kinds of intellectual exchanges in which we engage. One learns more from attempting to view a subject through a variety of different lenses than from staring at the same page through the same pair of glasses. And so it can be with analytic techniques.

Researchers run the gamut from technical experts who speak in equations and spin out table after table of numerical results to those who have tried to devise an alternative to page enumeration, so averse to 'numbers' were they. Most of us are somewhere in the middle, interested in a particular research question and trying to formulate as systematic and as persuasive an answer as possible.

Both approaches attempt to 'tell a story' from the data. Quantitative researchers generally refer to this process as hypothesis testing or 'modeling' the data to determine whether and to what extent empirical observations can be represented by the motivating theoretical model. Qualitative researchers may or may not invoke models. Whether the method of *analysis* will be quantitative or qualitative is not so much an issue of whether the information/data at hand are organized through classifications, rank-ordered relative to some notion of magnitude, or assessed at the interval or ratio level of measurement. The choice can involve assumptions about the nature of social reality, how it should be studied, the kinds of research questions that are of interest, and how errors of observation, measurement, estimation, and conclusion should be addressed.

Because this is a text in data analysis rather than data collection, each author assumes a certain structure of data and a certain range of research questions. To be sure, many decisions have been made before the researcher begins analysis, although active researchers seldom march through the stages of design, data collection, and data analysis as if they were moving through security checkpoints that allowed mobility in only one direction. Instead, researchers typically move back and forth, as if from room to room, taking what they learn in one room and revisiting what was decided in the previous room, keeping the doors open.

However, if the researcher is relying on secondary data – data collected to serve a broad range of interests, often involving large national samples – key features such as the sampling design and questionnaire must be taken as given, and other types of information – how long it took the respondent to settle on a response, whether the respondent took some care to frame the response within a particular context even though what was recorded was simply a level of agreement with a statement, for example – are not retrievable. Researchers who collect their own data use a variety of sampling procedures and collection tools that are designed to illuminate what they seek to understand and to provide information best suited to their research interests. But once the data are in hand, the evidence that may be required to address the research problem will be limited to interpretations, reconfigurations, or creative combinations of this already collected information.

This distinction between measuring amounts and distinguishing categories is sometimes referred to as the distinction between quantitative and qualitative variables, and it is only one of the arenas in which 'quantity' and 'quality' are counterposed. Another contrast that is made between qualitative and quantitative approaches involves the use of *statistical* methods of analysis, where quantitative implies using statistics and qualitative, in some quarters, means eschewing statistical approaches. But not all research that is classified as quantitative relies only on statistical approaches. Certainly in coding interview information, any researcher must make decisions about the boundaries of classification, must determine 'like' and 'unlike' things, and these decisions are already shaping any analysis that will follow. In similar fashion, not all qualitative researchers reject statistics, although reliance on inferential statistics is not common. Does the fact that a researcher calculates a correlation coefficient or bases a conclusion on differences in the counts of events suddenly toss the research into the quantitative camp? Does it matter, so long as the procedures are systematic and the conclusions are sound?

THE BASICS

We begin the volume with some basic issues that require a researcher's attention. The novice researcher is often dismayed when first using a given data set, since the correspondence between the concepts he or she has in mind is seldom there simply to be plucked from a list. Issues of reliability and validity loom large in the enterprise of analysis, for the conclusions that can be drawn on the basis of an analysis, regardless of how simple or complex, are contingent on the utility of the information on which the analysis is based. It is the instrumentality of measurement – measure as organizing tool that relates observation to concept to theory – that is a common thread of all analysis. Having made that most fundamental recognition, however, we must also note that it is often through debates over procedures of *analysis* that concerns about the limitations of measurement are played out. The value of a measure is its utility for improving our understanding of some social process, whether such a measure emerges through the manual sifting of data, or whether it serves as the framework for data collection.

Defining variables is therefore an exercise in establishing correspondence. Part of our everyday activities involves organizing the steady flow of information that our senses feed to our brains. The manner in which we accomplish this organization is not a random process. Rather, we categorize, we classify, we monitor frequency and intensity, we note repetition, stability, change, and amount of change, along a variety of dimensions. We fudge the boundaries of these categories with phrases such as 'kind of' and 'sort of'. And whereas our classification schemes may be quite functional for our own use, they may not sit well with the schemes others use.

In our everyday conversations we either gloss over disagreements, or we may pursue the issue by defending how we make sense of a situation. But in taking this next step, we move closer to scientific practice, in that our original statement must then be argued on the basis of empirical evidence, rules of assignment, what counts as 'similar' versus 'different', and which traits trump others in making such assignments. In other words, such statements – such classifications – have to be reproducible on the basis of the rules and the evidence alone. Then the issue is how convincing others find our approach.

Once we have defined the terms of our analysis, the temptation for statistical analysts is to move quickly to the most complex procedures, but that step is premature. We can learn much by studying the distributions of the variables we observe. And once we have good basic information on the univariate distributions, we should spend some time examining simple associations among variables, two at a time. Although this stage can be time-consuming, it is essential to gradually build our understanding of the data structures on which more complex associations will rely. These insights prove valuable when one must translate the finding into some reasoned argument that allows others to grasp what has been learned.

THE UTILITY OF STATISTICS

In many of these early chapters, basic statistical procedures are explained and illustrated. As Duncan (1975: 4) noted:

> There are two broad kinds of problems that demand statistical treatment in connection with scientific use of [models] … One is the problem of inference from samples … Statistical methods are needed to contrive optimal estimators and proper tests of hypotheses, and to indicate the degree of precision in our results or the size of the risk we are taking in drawing a particular conclusion from them. The second, not unrelated, kind of problem that raises statistical issues is the supposition that some parts of the world (not excluding the behavior of scientists themselves, when making fallible measurements) may realistically be described as behaving in a stochastic (chance, probabilistic, random) manner. If we decide to build into our models some assumption of this kind, then we shall need the aid of statistics to formulate appropriate descriptions of the probability distributions.

A major benefit of even 'fallible' measurement as the method of organizing our observations within some comparative framework is that it serves as a tool of standardization, which provides some assurance that both we, as well as others who attempt to replicate our work, can reliably identify equivalences and differences. 'Better' measurement is often taken to mean 'more precise' measurement, but the increase in precision must have utility for the question at hand; otherwise, such efforts simply increase the amount of 'noise' in the measure. For example, a public opinion researcher may decide that she can better capture variability in people's view of a certain taxation policy by moving beyond a Likert scale of agreement or disagreement to a set of possible responses that range from 0 (I see no redeeming value in such a policy) to 100 (I see this policy as the perfect response to the need). In testing this new measurement strategy, however, the researcher may discover that the set of actual responses is far more limited than the options available to respondents and, for the most part, these responses cluster at the deciles (10, 20, 30, …, 90); the respondents effectively reduce the choice set by focusing on multiples of 10 rather than increments of one. However, the researcher may also observe the occasional response of 54 or 32. What is she to make of that additional variability? Can she be confident that the difference between a response of 32 and one of 30 represents a reliable distinction with regard to tax policy? Or is the 32 response perhaps more a reflection of 'a tendency toward non-conformity'?

But this issue of precision/reliability/variability is not in itself a function of a statistical versus a non-statistical approach. The issue of precision, as Duncan notes, is one of assessing the likelihood of erroneous conclusions and the role played by 'chance' in our research activities. Error is inescapable. Error as mistaken observation, error as blunder, error as bias – how do we systematically manage error within the range of techniques available to us? The question at hand is how we manage error when using 'quantitative' or 'qualitative' techniques of analysis.

In sum, any analysis of data, however it proceeds, is a sorting process of information that contains errors – however it was collected. Further, this sorting process by which we sift 'good' information from 'error' also allows us to sort for logical patterns, for example, Y only occurs when X is present, but when X is present, Y does not always occur. And by identifying certain patterns, noting their frequency, determining the contexts under which they occur always, sometimes, or never, we make sense of the data. And that is our goal – to make 'sense' of the data.

SIMILARITIES BETWEEN QUANTITATIVE AND QUALITATIVE DATA ANALYSIS

It is easy to assume that the different preoccupations and inclinations of their respective practitioners mean that as research strategies, quantitative and qualitative research are totally different. Indeed, they *are* different, reflecting as they do distinctive intellectual traditions. However, this does not signal that they are so different they do not share any common features. It is worth reflecting, therefore, on the ways in which quantitative and qualitative data analysis may be said to have common characteristics. In doing so, we begin to raise issues about what data analysis is and also what constitutes a good data analysis, whether quantitative or qualitative.

Both are concerned with data reduction

Although data analysis is something more than data reduction, it is also true to say that paring down and condensing the vast amounts of data that we frequently collect in the course of fieldwork is a major preoccupation of all analysts. Indeed, it would be surprising if this were *not* the case since dictionary definitions of 'analysis', such as that found in *The Concise Oxford Dictionary*, refer to a process of resolving into simpler elements. Therefore, to analyze or to provide an analysis will always involve a notion of reducing the amount of data we have collected so that capsule statements about the data can be provided.

In quantitative research, we are often confronted with a large array of data in the form of many cases and many variables. With small amounts of quantitative data, whether in terms of cases or variables, we may be able to 'see' what is happening. We can sense, for example, the way in which a variable is distributed, such as whether there is bunching at one end of the scale or whether a particular value tends to recur again and again in a distribution. But with increasing numbers of cases and variables our ability to 'see' tails off. We begin to lose sight of what is happening. The simplest techniques that we use to summarize quantitative data, such as frequency tables and measures of central tendency and dispersion, are ways of reducing the amount of data we are handling. They enable us to 'see' our data again, to gain a sense of what the data show. We may want to reduce our data even further. For example, we might employ factor analysis to establish whether we can reduce the number of variables that we are handling.

Similarly with qualitative data, the researcher accumulates a large amount of information. This information can come in several different forms. Ethnographers are likely to amass a corpus of field notes based on their reflections of what they heard or saw. Researchers who use qualitative interviews usually find that they compile a mountain of transcripts of tape-recorded interviews. As Lee and Fielding remark in Chapter 23, the transcription of such interviews is frequently the source of a major bottleneck in qualitative research, because it is so time-consuming to produce. However, transcripts frequently constitute a kind of double bottleneck because, in addition to being time-consuming to generate, they are daunting to analyze. Most approaches to analyzing ethnographic fieldnotes, qualitative interview transcripts, and other qualitative data (such as documents) comprise a coding approach that segments the textual materials in question. Not all approaches to qualitative data analysis entail this approach; for example, narrative analysis, which is discussed in Chapter 29 by Czarniawska, involves a preference for emphasizing the flow in what people say in interviews. But whatever strategy is adopted, the qualitative researcher is keen to break his or her data down so that it is more manageable and understandable. As Lee and Fielding show, the growing use of computer-aided qualitative data analysis software is a means of making that process easier (in terms of the coding, retrieval, and management of data) in much the same way as statistical software can rapidly summarize large quantities of data.

Both are concerned with answering research questions

While the precise nature of the relationship between research questions and data analysis may be different among quantitative and qualitative researchers, both are concerned with answering research questions. In quantitative research, the stipulation of research questions may be highly specific and is often translated into hypotheses which are outlined either at the beginning of an investigation or as we begin to analyze our data. This process is often depicted as indicative of the hypothetico-deductive method with which quantitative research is often associated. Stipulating research questions helps to guide the collection and analyses of data, but having such organizing questions also serves to ensure that the research is about *something* and that the something will make a contribution to our understanding of an issue or topic.

Qualitative researchers are often somewhat circumspect about devising research questions, or perhaps more precisely about the timing of their formulation. In qualitative research there is frequently a preference for an open-ended strategy so that the meaning systems with which participants operate are not closed off by a potentially premature confinement of what should be looked at. In addition, qualitative researchers frequently revel in the flexibility that the open-endedness offers them. Consequently, it is not unusual to find accounts of the qualitative research process which suggest that the investigation did not start with any concrete research questions. Not all qualitative research is like this; many practitioners prefer to begin with the relatively clear focus that research questions provide. Nonetheless, there is a strong tradition among practitioners which enjoins them not to restrict their field of vision too early in the research process by orienting to research questions. Some versions of grounded theory, for example, specifically encourage the deferment of research questions, as Pidgeon and Henwood observe in Chapter 28. But all this is not to say that research questions do not get asked in some versions of qualitative research. Instead, they tend to emerge in the course of an investigation as the researcher gradually narrows the area of interest. The research questions may even be developed into hypotheses, as in grounded theory. Deferring the asking of research questions has the advantage for qualitative researchers of enabling them to develop an understanding of what is important and significant from the perspective of the people they are studying, so that research questions that may be irrelevant to participants are less likely to be asked, if it is the perspective of relevance that matters. It also offers greater flexibility in that interesting insights gleaned while in the field can be used as a springboard for new research questions.

Thus, while the stage at which the formulation of research questions occurs frequently differs between quantitative and qualitative research, and the nature of the research questions may also be somewhat different, data analysis is typically oriented to answering research questions regardless of whether the research strategy is quantitative or qualitative.

Both are concerned with relating data analysis to the research literature

This point is closely related to the previous one but nonetheless deserves separate treatment. An important aspect of any data analysis is to relate the issues that drive and emerge from it to the research literature. With quantitative data analysis, the literature tends to provide an impetus for data analysis, in that it is invariably a key element in the formulation of a set of research questions. Quantitative research papers typically conclude by returning to the literature in order to address such issues as whether a hypothesis deriving from it is confirmed and how far the findings are consistent with it.

With qualitative data analysis, the existing literature may help to inform or at least act as a background to the analysis. This means, for example, that the coding of transcripts or fieldnotes will be partly informed by the literature. Existing categories may be employed as codes. In addition, the qualitative researcher will typically seek to demonstrate the implications of an analysis for the existing literature.

Thus, practitioners of both research strategies are highly attuned to the literature when conducting data analysis. This feature is indicative of the fact that practitioners are equally concerned with making a contribution to theory through their data analysis.

Both are concerned with variation

Variability between cases is central to quantitative data analysis. The goal of quantitative data analysis is to capture the amount of variation in a sample and to explain why that variation exists as it does and/or how it was produced. An attribute on which people (or whatever the nature of the cases) do not vary, and which is therefore a constant rather than a variable, is typically not of great interest to most analysts. Their toolkit of data analysis methods is geared to variability rather than to its absence. As noted above, even the most basic tools of quantitative data analysis – measures of central tendency and dispersion – are concerned to capture the variability that is observed.

But variation is equally important to qualitative researchers when they conduct their analyses. Variation is understood somewhat differently from quantitative research in that it relates to differences one observes but to which one does not necessarily assign a numerical value, but it is nonetheless central as an observation of relative magnitude (e.g., respondents differed more in their opinions on this than on that). In the course of carrying out an analysis of qualitative data, the researcher is likely to be attending to assorted issues that reflect an interest in variation: Why does a particular activity or form of behavior occur in some situations rather than others? Why are some people excluded from participation in certain activities? To what extent do differences in certain kinds of behavior vary because of the different meanings associated with the behavior in certain situations? How and why do people's behavior or meaning attributions vary over time? These are common issues that are likely to arise in the course of qualitative data analysis, and all of them relate in some way to variation and variability. The idea that meaning and behavior need to be understood contextually (e.g., Mishler, 1979) implies that the researcher is forced to consider the implications of contextual variation for his or her findings.

Conversation analysis might be assumed to belie this point about qualitative data analysis in that its emphasis on the ordered nature of talk in interaction could be taken to imply that it is a lack of variation that is of concern. However, the conversation analyst is also concerned with such issues as *preference organization,* which presumes that certain kinds of responses are preferred following an initial utterance and is at least implicitly concerned with the exploration of variation. Similarly, an interest in the use of *repair mechanisms* in conversations would seem to imply a concern with variation and responses to it. Thus, once again, while it is addressed in different ways in quantitative and qualitative data analysis, the exploration of variation is an important component of both strategies.

Further, an initial understanding of patterns of variability may inform the collection of data. In the formal application of sampling theory, populations may be viewed as comprised of different strata, and each stratum may be assigned a different sampling ratio. In this way, the researcher ensures that sufficient variability of important minority characteristics occurs in the sample. Similarly, in deciding where and whom to observe, qualitative researchers may choose sites and/or groups they expect to differ, thereby building into the research design variability of observed behavior and/or observational context.

Both treat frequency as a springboard for analysis

That issues of frequency are important in quantitative data analysis is neither surprising nor illuminating. In the course of quantitative data analysis, the practitioner is bound to be concerned with issues to do with the numbers and proportions of people holding certain views or engaging in different types of behavior. The emphasis on frequency is very much bound up with variation, since establishing frequencies is a common way of expressing variation.

However, frequency is a component of qualitative data analysis as well. There are two ways in which this occurs. Firstly, as some commentators remark when they write up their analyses, qualitative researchers often use quantitative terms, such as 'most', 'many', 'often', and 'sometimes' (Becker, 1958). In many ways, these are very imprecise ways of conveying frequency and, given their ambiguity, it is usually difficult to know what they mean. Qualitative researchers are not alone in this regard, however. In spite of the fact that they use apparently more precise yardsticks for gauging frequency, quantitative researchers also resort to such terms as embellishments of their quantitative findings, although the actual values are generally reported as well. Moreover, when quantitative researchers do employ such terms, they apply to widely different indicators of frequency (Ashmore et al., 1989). Silverman (1985) recommends that qualitative researchers use limited quantification in their analyses rather than rely excessively on vague adjectival terms.

Frequency can be discerned in relation to qualitative data analysis in another way. As Bryman and Burgess (1994) observe, when they code their unstructured data, qualitative researchers are likely to rely on implicit notions of frequency. This can occur in at least two ways. They may be impressed by the frequency with which a theme appears in their transcripts or fieldnotes and may use this as a criterion for deciding whether to apply a code. Themes that occur very infrequently may be less likely to receive a distinct code. In addition, in developing codes into concepts or categories, they may use frequency as a method of deciding which ones are worth cultivating in this way.

Both seek to ensure that deliberate distortion does not occur

Although few social scientists nowadays subscribe to the view that we are objective, value-free observers of the social world, this recognition makes it more important that we proceed in ways that are explicitly defined and therefore replicable. There is evidence in certain quarters of the emergence of avowedly partial research. For example, Lincoln and Guba (1985) recommend that one set of criteria by which research should be judged involves the issue of *authenticity*. This set of criteria relates to the political dimension of research and includes such principles as *catalytic authenticity*, which enjoins researchers to ask whether their research has motivated members to engage in action to change their circumstances, and *tactical authenticity*, which asks whether the research has empowered members to engage in action. In spite of the use of such criteria, which are political in tone and which are a feature of much writing from a feminist standpoint, qualitative researchers have not suggested that the distortion of findings during data analysis should accompany political ambitions. There are plenty of opportunities for researchers to twist findings intentionally during data analysis – whether quantitative or qualitative. However, by and large, they are committed to presenting an analysis that is faithful to the data. Of course, there is a far greater recognition nowadays that both quantitative and qualitative researchers employ a variety of rhetorical strategies for convincing readers of the authenticity of their analyses (see Bryman, 1998, for a review of some of these writing techniques). However, this is not to suggest that data analysis entails distortion, but that through their writings researchers have to win over their readers to the credibility of what they are trying to say. In essence, what is guarded against in most quantitative and qualitative data analysis is what Hammersley and Gomm (2000) call *willful bias*, that is, consciously motivated misrepresentation.

Both argue the importance of transparency

Regardless of the type of research being conducted, the methodology that is used should not eclipse the data, but should put the data to optimal use. The techniques of analysis should be sufficiently transparent that other researchers familiar with the area can recognize how the data are being collected and tested, and can replicate the outcomes of the analysis procedure. (Journals are now requesting that authors provide copies of their data files when a paper is published so that other researchers can easily reproduce the analysis and then build on or dispute the conclusions of the paper.) Whether they also agree about what those outcomes mean is a different issue. Much of the disagreement that occurs in the research literature is less with analysis-as-process and more with the specification or the context in which the question is being addressed and the interpretation of the findings. In arguing a certain ‘story line’, a quantitative researcher may try to demonstrate the ‘robustness’ of findings by showing that certain key results persist when evaluated within a variety of contexts of specifications.

If we take as an exemplar of quantitative research the analysis of national survey data, transparency in the data collection process is generally high. Sampling procedures are well documented; comparative analysis of how the sample compares to the population on known characteristics is reported; the researcher is provided with a codebook and questionnaire that provide details about the questions asked, the range of responses given, and frequency distributions, so researchers can be confident they are reading the data correctly. Improvements in computer technology have made this process considerably easier, faster, and more reliable. In addition, the general availability of software packages to perform a wide range of analyses removes the mystery of what algorithm was used and what calculations were made.

But one issue of 'transparency' in quantitative research involves the use of statistical tools that, from some perspectives, 'distance' the researcher from the data. For example, missing values are imputed, cases are weighted, parameter estimates have confidence intervals that change with each specification, sometimes achieving the status of statistical 'significance' and sometimes falling short. Estimates of effects to the first, second, occasionally third decimal point – how can anyone 'see' the original data behind this screen of computational complexity? But to say that the procedures are sufficiently complex to require computer assistance in their application is *not* to say that they are opaque. The sampling framework that generates the case weights is derived from sampling theory, an ample literature that provides rules for both selection and adjustment, as well as the likely consequence of proceeding other than 'by the rules'. The algorithms on which sample estimates are based are derived from estimation theory, their properties tested through simulations and statistical experiments so that researchers can understand the conditions under which their use will yield desirable and reliable results. The process is neither convoluted nor impenetrable, but it is complex, and it is reasonable to assume that practitioners who use quantitative methods are not always well acquainted with the details of sampling, estimation, or statistical theories that provide the rationale for the practice. To acknowledge that building an understanding of the theoretical foundations for this practice is a challenging task is one thing; to reject this literature because it is challenging is quite another.

With qualitative research, an absence of distance and, until rather recently, limited use of technological innovation for organizing and analyzing information can create a different dilemma for replication. Observational data may rely on one person's recollections as fieldnotes are written; transcriptions of taped interviews or coded segments of videotape that anyone can evaluate provide more the type of exactitude that many quantitative types find reassuring. And clear rules that govern who, what, and when we observe; justifications for the chosen procedure over alternatives; rules of coding; logical relationships; analytical frameworks; and systematic treatments of data can combine to produce consistent and reproducible findings.

Conversation analysis (Chapter 26) takes a somewhat different line on this issue from most forms of qualitative data analysis, in that practitioners have always exhibited a concern to demonstrate the transparency of their data and of their analysis. Qualitative researchers generally have few guidelines about how to approach their data other than the need to address their research questions through their data. One of the great appeals of grounded theory (Chapter 28) has been that it provides a framework, albeit at a far more general level than statistical techniques provide, for thinking about how to approach qualitative data analysis. It is also worth bearing in mind that one of the arguments frequently employed in favor of computer-assisted qualitative data analysis is that it forces researchers to be more explicit about the way they approach their data, so that, in the process, the transparency of the analytic process may be enhanced.

Indeed, we begin to see here some of the ways in which quantitative and qualitative data analysis differ. Not only is there a difference in most instances in the transparency of the process, but also quantitative data analysts have readily available toolkits for the examination of their data. Conversation analysis comes closer to a toolkit approach than many other forms of qualitative data analysis, although semiotics (see Chapter 25) and to a certain extent discourse analysis (see Chapter 27) come close to providing this kind of facility. A further difference is that in analyzing secondary data, quantitative researchers usually conduct their analyses at the end of the research process, since data collection occurred elsewhere. However, in analyzing primary data, both quantitative and qualitative researchers intersperse data collection with data analysis. Quantitative researchers need to pilot-test their measures to ensure that the information collected meets criteria of both validity and reliability. And many writers on qualitative data analysis, particularly those influenced by grounded theory, advocate that data collection and analysis should be pursued more or less in tandem. As Coffey and Atkinson (1996: 2) suggest: 'We should never collect data without substantial analysis going on simultaneously. Letting data accumulate without preliminary analysis along the way is a recipe for unhappiness, if not total disaster.' Coffey and Atkinson (1996: 2) go on to say that there 'is no single right way to analyze data'. While this comment is made in relation to the analysis of qualitative data, it applies equally well in relation to quantitative data analysis. On the other hand, there are plenty of ways in which data can be wrongly or inappropriately analyzed, and a book such as this will help to steer people away from potential mistakes.

Both must address the question of error

The manner in which quantitative and qualitative approaches manage the effects of error may well be the most central point of difference. Quantitative research can be viewed as an exercise in managing error, since variability-as-observed-difference is both a function of empirically distinct characteristics and error in the empirical process of observing those distinctions. One context in which the utility of statistical information and the acknowledgment of error come into conflict is the courtroom. Statisticians asked to give expert testimony are inevitably asked by opposing counsel whether they are 'certain' of their findings. Regardless of whether they acknowledge a 5% margin for error, a 1% margin for error, or a 0.1% margin for error, they can never say with absolute certainty that 'this' occasion cannot possibly be an error. In contrast, for many years eyewitness testimony was the gold standard of evidence, since a 'good' eyewitness would deny uncertainty, testifying to no doubt, no possibility of error – testifying with certainty. And so they may have believed. But the frequency with which recently utilized DNA evidence is proving exculpatory has given everyone pause. If we cannot trust our own eyes, how can we be sure of anything? One answer is that absolute certainty was always an illusion, whether it was asserted in scientific enterprise or everyday life. Even so, we know many things, and in so knowing, we can accomplish many tasks. And in trying to accomplish, we can learn much more. So if our choice is between drowning in doubt or acting on best information, we act. Neither judge nor jury can ever be certain, in the sense that they cannot claim that error is impossible; but they can draw conclusions by weighing the evidence. And so they do.

Within the framework of behavioral and social science, both quantitative and qualitative analysts acknowledge that error is an unavoidable aspect of data collection, measurement, coding, and analysis procedures. And both agree that error cannot always be assumed to be random, such that the summary influences of error on our conclusions simply 'cancel out'. Much of the development in quantitative research that has occurred over the past three decades has been oriented toward better managing error. In particular, attention has been focused on developing procedures to address error as a confounding source in the data while preserving the substantive focus and the structural relations of interest. In fact, we can look at the chapters in this text as representing advancements in the analysis of error.

The early chapters on constructing variables, describing distributions, and dealing with missing data involve the exposition of techniques for using already collected bits of information and combining them, reconfiguring them, transforming them in ways that create a better match between the measure and the concept. The variance has been called the 'mean squared error' because it provides the average weighted distance of observations from the midpoint of the distribution. This measure of inequality, of observed difference, provides the problematic for further analysis designed to answer the question: what produced the differences?

Missing data can create problems of error, since the missing information may occur at higher frequency in one or another part of the distribution (creating truncated distributions), or the pattern of missing data may be correlated with other factors. Chapter 4, on inference, underscores the complications introduced by sampling error, or generally by procedures designed to allow statements about the whole using only partial information. What this and other early chapters share is an emphasis on process. Dealing with missing information through some kind of imputation procedure requires that we theorize about the process that created the data gaps in the first place. Why do some people answer this question, while other respondents refuse? What is it about the question, the kind of information the question tries to elicit, and the known characteristics of the respondent that makes 'refusal' more likely?

For example, collecting income information is notoriously difficult. People generally consider their household income or the amount they have saved or invested to be private information. Although respondents often like to offer their opinions, they are less pleased – and sometimes angered – by questions of 'fact' that appear to invade their privacy. But techniques for collecting information in wide categories, coupled with information about relationships among observed characteristics of respondents and the piece of missing information, have allowed improvements in imputations. To ask someone to report last year's gross annual income may elicit a refusal. But to follow up with a question that asks the respondent to report whether it was 'above $50 000' creates a benchmark. Once the respondent supplies that first benchmark, it is often possible to channel them through a progressive series of categories, so that the gross annual income is eventually known to be between $25 000 and $35 000. The exact income is still 'missing', but imputation procedures can now utilize the range of values in which it falls.

In similar fashion, Chapter 4 links the adjustments we make for sampling error (e.g., the building of confidence intervals around estimates by using information on the error of those estimates) to the selection procedures that generated the sample (the part) from the population (the whole). Again, we rely on the theory of probability to move from the population to the sample, and then again to move back from the sample estimate to the population parameter. If the selection process was not according to some known probability process, then probability theory is of no use to us, and we are left with a description of a set of observations that do not generalize to any known population. Later chapters on selection models take these issues further by suggesting approaches that explicitly model mechanisms of sample selection as part of the system of equations testing structural relationships.

The process of constructing variables also introduces error. Are single indicators sufficient? If we combine indicators, what type of weighting scheme should we employ? And even at our best, we realize that there is some slippage between the concepts as abstractions and the variables that we use as the informational repositories of their meaning. But errors in measurement attenuate measures of association, making it more difficult to take that next step of describing underlying processes that produce what we observe. And in trying to represent that process, we are limited to our success in finding information that maps well the conceptual space we have defined. Missing pieces of information – missing for everyone rather than missing selectively – create specification error, which can introduce bias into our conclusions. The chapters on regression, structural equation models, models for categorical data, etc. all address these issues of error that complicate the task of the researcher, providing guidance on proper procedures when we attempt to explain the variability in dependent variables measured in different ways (e.g., by interval scale, by dichotomy, by polytomous classification) and within different levels of complexity (e.g., single equation versus multiple equation models motivated by concerns of endogeneity).

And if we are really interested in the underlying process, don't we need to look at process? In other words, shouldn't we be analyzing longitudinal data, following individuals over time so we know how changes in one aspect of their lives may be linked to subsequent changes in other aspects of their lives? But then we have the complication of correlated errors, since multiple observations on one respondent are likely to be characterized by similar observational errors at each point in time. Or perhaps our longitudinal frame should be the comparison of same-aged people over time to determine whether opinions in the aggregate have changed over time, for example? Further, as social scientists, we know that context is important, that processes may unfold one way under one set of circumstances, but unfold differently under different organizational or institutional constraints. How do we analyze information that describes the individual *within* the organizational context? Over time? These are the issues that event-history models, hierarchical linear models, panel models, latent curve models, and other advanced techniques were designed to address.

The more complicated the questions we ask, the more complicated the error structure with which we must deal, but we are not without tools to tackle these tasks, although the tools become more complicated as well. Any carpenter who wants to saw a board into two pieces has a variety of tools at his or her disposal, the simplest being a handsaw. But to cut designs into the wood, or dovetail a joint, or fit rafters on a double-hipped roof, requires more sophisticated tools to produce the desired outcome.

In qualitative research, error has not been a notion that has great currency. Indeed, some qualitative researchers argue that the very idea of error implies a 'realist' position with which some versions of qualitative research, particularly those influenced by postmodernism (see Chapter 30), are uncomfortable. For these qualitative researchers, it is demonstrating the credibility of findings that is likely to be of roughly equivalent concern (Lincoln and Guba, 1985), although it may be implicit in some notions of validity in qualitative research (e.g., Kirk and Miller, 1986). Demonstrating credibility takes many forms, but a major feature is being able to show a close correspondence between one's data and one's conceptualization, a concern which can be translated into quantitative research as concerns with 'goodness of fit', or how well the theoretical model fits the empirical information.

For those who use statistics, the 'fit' can be assessed as prediction successes versus prediction errors. But interpreting whether a given level of fit, a given value of the statistic, is persuasive evidence of the correctness of the theory is open to dispute. And the terms of dispute on this point are likely to be similar for both qualitative and quantitative researchers. Are your observations consistent with the predictions of the theory? Has the information been properly classified? Have you ignored other things that could change this picture? Do I believe your story? In both types of research, the richer the data, the more persuasive the conceptualization is likely to be.

Moreover, for the qualitative researcher, the emerging concepts must be demonstrably located in the data. The quantitative researcher refers to this as operationalization – whether the empirical variables fit the theoretical concepts. In the process of sorting through the vast amounts of information, many qualitative researchers must inevitably classify, which means they determine categories and group what they observed into 'like' and 'unlike' observations. Is there only one way this can be accomplished? Most researchers from either camp would answer 'no'. So both types of researchers may be accused of category 'errors', in that someone else working with these same observational data may define groups differently. Disputes such as these are not uncommon.

Has the researcher ignored something 'important' in his or her analysis? Not intentionally, but someone with a different perspective may argue a different 'story' by picking up a feature that the first researcher failed to consider. Quantitative researchers refer to this as specification 'error', which simply means that in developing your story, you have left out something relevant. This error of omission is among the most serious in quantitative research, since it means that the evidence on which you are basing your conclusions is incomplete, and it is difficult to say how the story may change once you take this new twist into account.

These sources of 'error' in qualitative and quantitative research – observational error, classification error, and specification error – can be introduced through the choices made by the researcher, who may fail to pick up important cues from his or her research participants or may misread in conceptual terms what is happening. Thus, even though error is a term that is unlikely to sit easily with the way many, if not most, qualitative researchers envision their work, it is not without merit. A major difference is that the quantitative researcher turns to sampling, measurement, and estimation theory to mathematically formalize how error is assessed and addressed; the qualitative researcher generally relies on rules of logic, but not on mathematics. Both researchers, however, must rely on argument and the strength of evidence they muster from their data to convince others of their story.

The trick for the qualitative researcher is one of balancing a fidelity to the data (in a sense, a commitment to naturalism) with a quest to say something meaningful to one's peers (in other words, to conceptualize and theorize). The advantage of fidelity to the data is that the researcher's emerging conceptual framework will be relatively free of error, but the problem is that it may be difficult to appear to have done anything other than act as a conduit for the world-view of the people who have been studied. The corollary of this position is that qualitative researchers must be wary of conceptualizing to such an extent that there is a loss of contact with the data, so that the credibility of their findings is threatened and therefore error creeps in.

ORGANIZATION OF THE BOOK

It is with the kinds of issues and considerations outlined above that the authors of the chapters in this volume have sought to come to terms. The quantitative–qualitative research distinction partly maps onto the organization of the book, but only partly. On the face of it, qualitative data analysis is covered in Part V. However, content analysis is essentially a quantitative approach to the analysis of unstructured or qualitative data, while the chapters in Part I on feminist issues in data analysis (Chapter 6) and historical analysis (Chapter 7) transcend the distinction in having implications for and elements of both quantitative and qualitative approaches to data analysis. Part I provides some of the foundations of data analysis – the nature of distributions and their analysis; how to construct variables; the nature of observational and statistical inference; what missing data are and their implications; and, as has just been remarked upon, feminist issues and historical analysis.

Part II teaches the reader about the single-equation general linear model, its extensions, and its applicability to particular sorts of research questions. Although called the 'linear' model, it can accommodate a variety of functional forms of relationships, which can be used to test whether an association is monotonic, curvilinear, or proportional, for example.

Part III addresses the issue of studying change. Whereas in cross-sectional analysis we can describe how the outcome is associated with certain characteristics, in longitudinal analysis we introduce the timing of the outcome relative to the timing of changes in the characteristics. Introducing time into the research design creates another layer of complications, which must be addressed through both theory and technique. It also requires a different data structure, which factors time into both the procedures and the content of data collection.

Part IV introduces the reader to some recently developed but well-established approaches to data analysis. Many of these approaches address the issue of endogeneity, which is the complication that some of the factors we view as predictors of a certain outcome are also at least partly determined *within* the same system of relationships. In such circumstances, single-equation models are not sufficient.

Part V, as previously noted, is devoted to the analysis of qualitative data. In Chapter 23, some of the main elements of qualitative data analysis are outlined, along with the issues involved in the use of computer software for the management and analysis of qualitative data. Chapter 24 deals with content analysis, which, although an approach for the analysis of qualitative data, employs an analytic strategy that is very much in tune with quantitative research. Chapters 25–27 deal with approaches to qualitative data analyses that emphasize language and its significance in the construction of social reality. Chapter 28 discusses grounded theory, which has been referred to several times in this introduction and which has become one of the major frameworks for organizing qualitative data analysis. Chapter 29, in presenting narrative analysis, provides a discussion of an approach that is attracting a growing number of adherents and which in many ways provides an alternative to the coding approach to the initial analysis of qualitative data that is characteristic of grounded theory and many other approaches to the analysis of qualitative data. Finally, Chapter 30 provides an outline of the highly influential postmodernist approach, particularly in relation to qualitative data. In many ways, the postmodernist mind-set entails an inversion of many of our cherished beliefs about how social research should be carried out and about how to understand its written products.

SUMMARY

The approaches explicated in this *Handbook* are not exhaustive of the range of approaches available to the researcher. As we explained earlier, we chose to build on basics, yet address some of the most difficult and complicated issues researchers face. Some of the most recent innovations in approaches are, at best, mentioned parenthetically, with reference to other sources of information the interested reader is encouraged to pursue. Our goal is to help readers do 'good research'.

Good research shares some common features. It does not violate ethical guidelines. It is not based on 'fictionalized' data, but rather on information collected according to rules of observation and recording. It describes with fidelity and, at its best, explains how what was observed came to be as it was rather than otherwise. In building this text, we hope to allow interested researchers to learn from one another about a wide range of approaches to data analysis. New techniques are in the process of development; techniques already in use find new advocates and new critics. Here is a place to take up the journey.

REFERENCES

Ashmore, M., Mulkay, M. and Pinch, T. (1989) *Health and Efficiency: A Sociology of Health Economics*. Milton Keynes: Open University Press.

Becker, H.S. (1958) 'Problems of inference and proof in participant observation', *American Sociological Review*, 23: 652–60.

Bryman, A. (1998) 'Quantitative and qualitative research strategies in knowing the social world', in T. May and M. Williams (eds), *Knowing the Social World*. Buckingham: Open University Press.

Bryman, A. and Burgess, R.G. (1994) 'Reflections on qualitative data analysis', in A. Bryman and R.G. Burgess (eds), *Analyzing Qualitative Data*. London: Routledge.

Coffey, A. and Atkinson, P. (1996) *Making Sense of Qualitative Data: Complementary Research Strategies*. Thousand Oaks, CA: Sage.

Duncan, O.D. (1975) *Introduction to Structural Equation Models*. New York: Academic Press.

Hammersley, M. and Gomm, R. (2000) 'Bias in social research', in M. Hammersley (ed.), *Taking Sides in Social Research: Essays in Partisanship and Bias*. London: Routledge.

Kirk, J. and Miller, M.L. (1986) *Reliability and Validity in Qualitative Research*. Newbury Park, CA: Sage.

Lincoln, Y.S. and Guba, E. (1985) *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Mishler, E. (1979) 'Meaning in context: is there any other kind?', *Harvard Educational Review*, 49: 1–19.

Silverman, D. (1985) *Qualitative Methodology and Sociology: Describing the Social World*. Aldershot: Gower.

PART I

Foundations

2

Constructing Variables

ALAN BRYMAN AND DUNCAN CRAMER

The process of quantitative research is frequently depicted as one in which theory is employed in order to deduce hypotheses which are then submitted to empirical scrutiny. Within the hypothesis will be two or more concepts that will require translation into empirical indicators. These indicators are frequently referred to as *variables* and represent the fundamental focus of all quantitative research. While some writers might question the degree to which quantitative research necessarily follows such a linear progression and indeed how far it is driven by hypotheses (as against simply research questions), there is no doubt that the variable represents a major focus (Bryman, 2001). It constitutes a crucial bridge between conceptualization and findings.

Essentially, the quantitative researcher is concerned to explore variation in observed values among units of analysis and the correlates and causes of variation. All techniques of quantitative data analysis – from the most basic methods to the most advanced – are concerned with capturing variation and with helping us to understand that variation. The variable is crucial because it is the axis along which variation is measured and thereby expressed. Indeed, so central is the variable to the discourse of quantitative research that it has to all intents and purposes become synonymous with the notion of a concept. Variables are, after all, supposed to be measures or indicators that are designed to quantify concepts, but frequently writers of research papers and methodology texts refer to the process of measuring variables. In the process, concepts and variables become almost indistinguishable. The variable is also frequently the focus of attention for critics of quantitative research (e.g., Blumer, 1956), in large part because it is emblematic of the research strategy.

The variable can be usefully contrasted with the idea of a *constant*. The latter occurs when there is no variation in observed values among units of analysis, as when all members of a survey sample reply to a questionnaire item in the same way. Uncovering constants is relatively unusual and is likely to require a somewhat different strategy on the part of the researcher, since techniques of quantitative data analysis are typically concerned with exploring variation rather than its absence.

LEVELS OF MEASUREMENT

One of the most fundamental issues in quantitative data analysis is knowing which types of technique can be used in relation to particular levels of measurement. It is fundamental because each statistical technique presumes that the levels of measurement to which it is being applied are of a certain type or at least meet certain basic preconditions. This means that if a technique is applied to variables which do not meet its underlying assumptions, the resulting calculation will be meaningless. Therefore, being able to distinguish between the different levels of measurement is basic to the art and craft of quantitative data analysis.

Writers often refer to different 'types of variables' as a shorthand for different levels of measurement. As such there is an array of different types of variables or levels of measurement. This array reflects the fact that the four levels of measurement to be discussed are on a continuum of degrees of refinement. There are four types of variables which are typically presented in terms of an ascending scale of refinement: nominal; ordinal; interval; and ratio.

Nominal variable

The *nominal variable*, often also referred to as the *categorical variable*, is the most basic level of measurement. It entails the arbitrary assignment of numbers (a process referred to as *coding*) to the different categories that make up a variable. The different categories simply constitute a classification. We cannot order them in any way – they are simply different. The numbers that are different have no mathematical significance; instead, they act as tags which facilitate the computer processing of the data. Thus, if we asked a question in a social survey on religious affiliation, we would assign a number to each type of affiliation and record each respondent's affiliation with the appropriate number. Similarly, in an experiment on asking questions, Schuman and Presser (1981) asked:

> The next question is on the subject of work. People look for different things in a job. Which of the following five things would you *most* prefer in a job?

The five options which could be chosen were:

1 Work that pays well
2 Work that gives a feeling of accomplishment
3 Work where there is not too much supervision and you make most decisions yourself
4 Work that is pleasant and where the other people are nice to work with
5 Work that is steady with little chance of being laid off

In assigning numbers to each of these five possible answers, all we are doing is supplying a label to each type of response. We can only say that all those answering in terms of the first response differ from those answering in terms of the second, who differ from those answering in terms of the third, and so on.

Sometimes, we have just two categories, such as male/female or pass/fail. Strictly speaking such variables – often referred to as *dichotomous variables* or *binary variables* (e.g., Bryman and Cramer, 2001) – are nominal variables. However, sometimes such variables require a different approach to analysis from nominal variables with more than two categories and are therefore treated by some writers as a separate type of variable.

Ordinal variable

As we have seen, with a nominal variable we can say no more than that people (or whatever the unit of analysis) differ in terms of its constituent categories. If we are able to array the categories in terms of rank order then we have an ordinal variable. Thus, if we asked a sample of people how satisfied they were with their jobs and presented them with the following possible responses, we would have an ordinal variable:

1 Very satisfied
2 Fairly satisfied
3 Neither satisfied nor dissatisfied
4 Fairly dissatisfied
5 Very dissatisfied

In this case, although the numbers attached to each category are merely used to allow the answers to be processed, we can say that each number has a significance that is *relative* to the others, since they are on a scale from 1 (denoting very satisfied) to 5 (denoting very dissatisfied). Each number therefore represents a level of job satisfaction or dissatisfaction. What we cannot say is that, for example, the difference between being very satisfied and fairly satisfied is the same as the difference between being very dissatisfied and fairly dissatisfied. All we can say is that the respondents differ in terms of their levels of job satisfaction, with some respondents being more satisfied than others.

Interval variable

An interval variable is the next highest level of refinement. It shares with an ordinal variable

Table 2.1 *Summary of the characteristics of the four types of variable*

	Is there a true zero point?	Are the distances between categories equal?	Can the categories be rank-ordered?
Ratio variable	Yes	Yes	Yes
Interval variable	No	Yes	Yes
Ordinal variable	No	No	Yes
Nominal variable	No	No	No

the quality of the rank ordering of the categories (which should more properly be called *values*) but differs in that with an interval variable, the distances between the categories are equal across the range of categories. Thus, we can say that the difference between a temperature of 43°F and 44°F is the same as the difference between 24°F and 25°F. As such, the values that an interval variable can take are genuine numbers rather than the scoring or coding process associated with the quantification of the categories of nominal and ordinal variables, where the number system is essentially arbitrary. However, interval variables are relatively unusual in the social sciences, in that most apparently interval variables are in fact ratio variables.

Ratio variable

A ratio variable represents the highest level of measurement. It is similar to an interval variable, but in addition there is a true zero point. In measurement theory, a true zero point implies an absence of the quality being measured, that is, you cannot have less than none of it. This feature means that not only can we say that the difference between an income of $30 000 a year and an income of $60 000 a year is the same as the difference between an income of $40 000 and an income of $70 000 a year (that is, a difference of $30 000), but also we can say that the income of $60 000 a year is double that of $30 000 a year. This means that we can conduct all four forms of arithmetic on ratio variables. Similar qualities can be discerned in such common variables as age, years in full-time education, size of firm, and so on.

In the social sciences, because most apparently interval variables are ratio variables, it is common for writers to prefer to refer to them as interval/ratio variables (e.g., Bryman and Cramer, 2001). Moreover, the vast majority of statistical techniques which require that the variable in question is at the interval level of measurement can also be used in relation to ratio variables. Therefore, the crucial distinctions for most purposes are between nominal, ordinal and interval/ratio variables.

Table 2.1 seeks to bring together the key decision-making principles that are involved in deciding how to distinguish between different kinds of variables.

MEASURES AND INDICATORS

A distinction is often drawn between measures and indicators. Measures constitute direct quantitative assessments of variables. For example, we could say that a question on respondents' incomes in a survey would provide us with a measure of the variable income. As such, reported income is a very direct estimate of income. This can be contrasted with a situation in which the quantitative assessment of a variable is or has to be indirect. An example is the previously cited question on job satisfaction. While the question asks directly about job satisfaction, we do not know whether it does in fact tap that underlying variable. In this case, we are using the question as an *indicator* of job satisfaction. Whether it does in fact reflect respondents' levels of job satisfaction is an issue to do with whether it is a *valid* indicator, about which more will be said below. The issue of whether something is an indicator or a measure is not to do with an inherent quality: if respondents' answers to a question on their incomes are employed as a proxy for social class, it becomes an indicator rather than a measure as in the previous illustration.

CODING

A key step in the preparation of data for processing by computer is *coding*. As has already been suggested in relation to nominal and ordinal variables, precisely because these variables are not inherently numerical, they must be transformed into quantities. Illustrations of the coding process have already been provided in relation to Schuman and Presser's (1981) question on work motivation and an imaginary example of a question on job satisfaction. In each case, the numbers chosen are arbitrary. They could just as easily start with zero, or the direction of the coding could be the other way around.

Coding in relation to social surveys arises mainly in relation to two kinds of situations. Firstly, in the course of designing a structured interview or self-administered questionnaire, researchers frequently employ *pre-coded questions*. Such questions include on the instrument itself both the categories from which respondents must choose and the code attached to each answer. Coding then becomes a process of designating on the completed questionnaires which code an answer denotes. The second kind of context arises in relation to the post-coding of open questions. Coding in this context requires that the researcher derives a comprehensive and mutually exclusive set of categories which can denote certain kinds of answer.

What is crucial is that the coding should be such that:

- the list of categories is mutually exclusive so that a code can only apply to one category;
- the list of categories is comprehensive, so that no category or categories have been obviously omitted; and
- whoever is responsible for coding has clear guidelines about how to attach codes so that their coding is consistent (often called *intra-coder reliability*) and so that where more than one person is involved in coding the people concerned are consistent with each other (*inter-coder reliability*).

The first two considerations are concerned with the design of pre-coded questions and with the derivation of categories from open questions. The third consideration points to the need to devise a coding frame which pinpoints the allocation of numbers to categories. In a sense, with pre-coded questions, the coding frame is incorporated into the research instrument. With open questions, the coding frame is crucial in ensuring that a complete list of categories is available and that the relevant codes are designated. In addition, it is likely to be necessary to include a detailed set of instructions for dealing with the uncertainties associated with the categorization of answers to open questions when the appropriate category is not immediately obvious. With techniques like structured observation and content analysis, the design of such instructions – which is often in a form known as a *coding manual* – is a crucial step in the coding of the unstructured data which are invariably the focus of these methods.

A further consideration is that researchers quite often *recode* portions of their data. This means that their analyses suggest that it is likely to be expedient or significant to aggregate some of the codes and hence the categories that the codes stand for. For example, in the coding of unstructured data, the researcher might categorize respondents into, for example, nine or ten categories. For the purposes of presenting a frequency table for that variable, this categorization may be revealing, but if the sample is not large, when a contingency table analysis is carried out (e.g. cross-tabulating the variable by age), the cell frequencies may be too small to provide a meaningful set of findings. In response to this situation, the researcher may group some of the categories of response so that there are just five categories. Such recoding of the data can only be carried out if the recoded categories can be meaningfully combined. There is the risk that the process of recoding in this way might result in combinations that cannot be theoretically justified, but recoding of data is quite common in the analysis of survey and other kinds of data.

SCALE CONSTRUCTION

One of the crucial issues faced in the measurement process in social research is whether to employ just one or more than one (and in fact usually several) indicators of a variable. Employing more than one indicator has the obvious disadvantage of being more costly and time-consuming than relying on one indicator. However, there are certain problems with a dependence on single indicators:

1. A single indicator may fail to capture the full breadth of the concept that it is standing in for. This means that important aspects of the concept are being overlooked. The use of more than one

indicator increases the breadth of the concept that is being measured.

2. In surveys, a single indicator may fail to capture a respondent's attitude to an issue or behaviour. This may be due to a variety of factors, such as lack of understanding or misinterpretation of a question. By using several indicators, the effect of such error may be at least partly offset by answers to other questions which serve as indicators and which are not subject to the same problem.
3. When more than one indicator is employed and the score on each indicator is then combined to form a total score for each respondent (as occurs with the use of summated scales – see below), much greater differentiation between respondents is feasible than when a single indicator is employed. For example, with the imaginary job satisfaction indicator used above, respondents could only be arrayed along a scale from 1 to 5. If more than one indicator is used and scores are aggregated, much finer quantitative distinctions become possible.

In other words, for any single respondent, reliance on a single indicator increases the likelihood of measurement error.

The recognition of the importance of multiple-indicator measures has resulted in a growing emphasis on the construction of scales. There are different approaches to scale construction, but most researchers employ *summated scales*, which entail the use of several items which are aggregated to form a score for each respondent. This allows much finer distinctions between respondents to be made (see point 3 above). One of the most common formats for this type of scale is the *Likert scale*, whereby respondents are presented with a series of statements to which they indicate their levels of agreement or disagreement.

To illustrate this approach to scale construction, consider an attempt by a researcher interested in consumerism to explore (among other issues) the notion of the 'shopaholic'. The following items might be used to form a Likert scale to measure shopaholicism:

1 I enjoy shopping.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

2 I look forward to going shopping.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

3 I shop whenever I have the opportunity.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

4 I avoid going shopping if I can.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

5 When I visit a town or city I don't know well, I always want to see the shops.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

6 Shopping is a chore that I have to put up with.
Strongly agree   Agree   Neither agree nor disagree   Disagree   Strongly disagree

Each reply will be scored. Various scoring mechanisms might be envisaged, but let us say that we want 5 to represent the highest level of shopaholicism represented by each answer and 1 the lowest, with 3 representing the neutral position. Notice that two of the items (4 and 6) are 'reverse items'. With the four others agreement implies a penchant for shopping. However, with items 4 and 6, agreement suggests a dislike of shopping. Thus, with items 1, 2, 3 and 5, the scoring from strongly agree to strongly disagree will go from 5 to 1, but with items 4 and 6 it will go from 1 to 5. This reversal of the direction of questioning is carried out because of the need to identify respondents who exhibit *response sets*, which have been defined as 'irrelevant but lawful sources of variance' (Webb et al., 1966: 19). An example of a response set to which Likert and similar scales are particularly prone is *yeasaying* or *naysaying*, whereby respondents consistently answer in the affirmative or negative to a battery of items apparently regardless of their content. Consequently, if a respondent answered strongly agree to all six items, we would probably take the view that he or she is not paying much attention to the content of the items, since the answers are highly inconsistent in their implications.

The scale would have a minimum score for any individual of 6 (presumably indicating a 'shopaphobe') and a maximum of 30 (a total 'shopaholic'). Most will be arrayed on the 23 points in between. A respondent scoring 5, 4, 4, 5, 3, 5, producing a score of 26, would be towards the shopaholic end of the continuum. A further feature of such scales is that essentially they produce ordinal variables. We cannot really say that the difference between a score of 12 and a score of 13 is equal to the difference between a score of 15 and a score of 16. However, most writers are prepared to treat such scales as interval/ratio variables on the grounds that the large number of categories (25 in this case) means that they approximate to a 'true' interval/ratio variable. Certainly, summated scales are routinely treated as though they are interval/ratio variables in journal papers reporting the results of research.

With a Likert scale, respondents indicate their degrees of agreement. While a five-point scale of agreement is employed in the above example, some researchers prefer to use seven-point scales (very strongly agree, strongly agree, agree, etc.) or even longer ones. Other types of response format for summated scales include the binary response format:

I enjoy shopping Agree Disagree

the numerical response format:

I enjoy shopping 5 4 3 2 1

(where 5 means Strongly agree and 1 means Strongly disagree)

and the bipolar numerical response format:

I enjoy shopping 7 6 5 4 3 2 1 I hate shopping

Once a scale has been devised and administered, the researcher needs to ask whether the resulting scale measures a single dimension. There are three highly related aspects to this question.

1. Is there an item (or are there items) showing a different pattern of response from those associated with the other constituent items? If there are, the offending item or items need to be eliminated from the scale. One way of checking for this possibility is to search out information on the *item–total correlations*. An inter-item correlation relates scores on each item to scores on the scale overall. If an inter-item correlation is much lower or higher than other inter-item correlations, it becomes a candidate for exclusion from the scale.
2. Is the scale internally reliable? This issue, which will be elaborated upon below, is concerned with the overall internal coherence of the items. Eliminating items which show a different pattern of response from the rest will enhance internal reliability.
3. Does the scale contain more than one dimension? If there are items which show a different pattern of response, it may be that there is a systematic quality to this variation such that the scale is not measuring a single dimension but possibly two or more. When this occurs, the nature of the underlying dimensions needs to be identified and named. Factor analysis is the most appropriate means of exploring this issue and will be given greater attention below.

The second of these aspects is concerned with the more general issue of the reliability of variables, which, along with validity, is a crucial issue in the evaluation of the adequacy of a variable.

RELIABILITY AND VALIDITY OF VARIABLES

Reliability and validity are crucial criteria in the evaluation of variables. In spite of the fact that these two terms are often used interchangeably in everyday speech, they refer to different aspects of the qualities of variables.

Reliability

Reliability is concerned with the consistency of a variable. There are two identifiable aspects of this issue: *external* and *internal reliability*. If a variable is externally reliable it does not fluctuate greatly over time; in other words, it is stable. This means that when we administer our scale of shopaholicism, we can take the view that the findings we obtain are likely to be the same as those we would find the following week. The most obvious examination of external reliability is to test for *test–retest reliability*. This means that sometime after we administer our scale, we readminister it and examine the degree to which respondents' replies are the same for the two sets of data. The chief difficulty with this method is that there are no guidelines about the passage of time that should elapse between the two waves of administration. If the passage of time is too great, test–retest reliability may simply be reflecting change due to intervening events or respondents' maturation. Furthermore, testing for test–retest reliability can become a major data collection exercise in its own right, especially when large samples are involved and when there are several variables to be tested.

Internal reliability is an issue that arises in connection with multiple-indicator variables. If a variable is internally reliable it is coherent. This means that all the constituent indicators are measuring the same thing. There are several methods for assessing internal reliability, one of which – item–total correlations – was briefly mentioned above. A further method is *split-half reliability*. This entails randomly dividing the items making up a scale into two halves and establishing how well the two halves correlate. A correlation below 0.8 would raise doubts about the internal coherence of the scale and perhaps prompt a search for low item–total correlations. In the case of the shopaholicism scale, the scale would be divided into two groups of three items, and respondents' scores on the two groups of items would be assessed. Nowadays, the most common method of estimating internal reliability is *Cronbach's alpha* (α), which is roughly equivalent to the average of all possible split-half reliability coefficients for a scale (Zeller and Carmines, 1980: 56). The usual formula is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{1}{\sigma_x^2}\sum\sigma_i^2\right),$$

where k is the number of items; $\sum\sigma_i^2$ is the sum of the total variances of the items; and σ_x^2 is the variance of the total score (Pedhazur and Schmelkin, 1991: 93). If alpha comes out below 0.8, the reliability of the scale may need to be investigated further. Computer software programs such as SPSS include a facility whereby it is possible to request that the alpha for the scale be computed with a particular item deleted. If there is a sharp rise in the level of alpha when any item is deleted, that item will then become a candidate for exclusion from the scale.

An important consideration in the measurement process is that resulting variables will contain *measurement error* – variation that is separate from true variation in the sample concerned. Such measurement error is an artefact of the measurement instruments employed and their administration. For many researchers, assessing internal reliability is one way in which they can check on the degree of measurement error that exists in summated scales, although it cannot exhaust the range of possible manifestations of such error.

Validity

Validity is concerned with the issue of whether a variable really measures what it is supposed to measure. Can we be sure that our scale of shopaholicism is really to do with shopaholicism and not something else? At the very least, we should ensure that our scale exhibits *face validity*. This will entail a rigorous examination of the wording of the items and an examination of their correspondence with the theoretical literature on consumption. We might also submit our items to judges and invite them to comment on the wording of the items and on the goodness of fit between the items and what we might take shopaholicism to entail. However, face validity is only a first step in validity assessment.

Criterion-related validity assesses a scale in terms of a criterion in terms of which people are known to differ. This form of validity assessment can be viewed in terms of two forms. Firstly, testing for *concurrent validity* relates a variable to a contemporaneous criterion. Thus, we might ask respondents who are completing our shopaholicism scale how frequently they go shopping. If we found that there was no difference between shopaholics and shopaphobes in terms of the frequency with which they go shopping, we might question how well the scale is measuring the underlying concept. Equally, if the two types of shoppers clearly differ, our confidence is enhanced that the scale is measuring what it is supposed to be measuring. Secondly, testing for *predictive validity* relates a variable to a future criterion. Some months after we administer the shopaholicism scale we might recontact our respondents and ask them about the frequency with which they have been shopping in the previous month. Again, we would expect the shopaholicism scale to be able to discriminate between the frequent and occasional shoppers. Alternatively, we might ask our respondents to complete a structured diary in which they report the frequency with which they go shopping and the amounts of time spent on their expeditions.

Testing for *construct validity* entails an examination of the theoretical inferences that might be made about the underlying construct. It means that we would have to stipulate hypotheses concerning the construct (shopaholicism) and then test them. Drawing on theories about the consumer society and consumerism, we might anticipate that shopaholics will be more concerned about the sign value of goods than their use value. Consequently, we might expect they will be more concerned with the purchase of goods with designer labels. We could therefore design some questions concerned with respondents' predilection for designer brands and relate these to findings from our shopaholicism scale. Of course, the problem here is that if the theoretical reasoning is flawed, the association will not be forged and this is clearly not a product of any deficiencies with our scale.

These are the major forms of validity assessment. Other methods, such as *convergent validity*, whereby a different method is employed to measure the same concept, are employed relatively rarely because they constitute major projects in their own right.

One final point on this issue is that validity presupposes reliability. If you have an unreliable variable, it cannot be valid. If a variable is externally unreliable, it fluctuates over time and therefore cannot be providing a true indication of what it is supposed to be measuring. If it is internally unreliable, it is tapping more than one underlying concept and therefore is not a genuine measure of the concept in question.

DUMMY VARIABLES

One way of examining the association between a nominal or categorical variable (such as religious affiliation or nationality) and a non-nominal variable (such as income or life satisfaction) is to code the different categories of the categorical variable in a particular way called dummy coding (Cohen and Cohen, 1983). This procedure will be explained in terms of the following example. Suppose we wanted to determine the association between nationality and life satisfaction. To enable the relevant statistics to be computed, a small sample of fictitious data has been created and is presented in Table 2.2.

Table 2.2 *Life satisfaction in three nationalities*

	American	British	Canadian
	9	8	7
	7	5	7
	6	4	4
Mean	7.33	5.67	6.00

The categorical variable consists of three nationalities, American, British and Canadian. Each group consists of three people. The non-categorical variable comprises a 10-point measure of life satisfaction varying from 1 to 10, with higher scores representing greater life satisfaction. From the mean score for each nationality, we can see that the Americans have the greatest life satisfaction, followed by the Canadians and then the British. What we are interested in is not the association between particular nationalities and life satisfaction (e.g., being American and life satisfaction) but the association between the general variable reflecting these nationalities and life satisfaction (i.e., the association between nationality and life satisfaction).

The simplest way of expressing the association between the general variable of nationality and life satisfaction is in terms of the statistical coefficient called *eta squared*. Eta squared is the variance in life satisfaction attributed to the variable of nationality as a proportion of the total variance in life satisfaction. It can be worked out from an unrelated one-way analysis of variance. In this case eta squared is 0.194. This method does not involve dummy coding.

The dummy coding of a categorical variable may be used when we want to compare the proportion of variance attributed to that variable with the proportion of variance attributed to non-categorical variables (such as age) together with any other categorical variables (such as marital status). The method usually used to determine these proportions is multiple regression. Multiple regression can be represented by the following regression equation:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

The dependent or criterion variable is often designated y and in our example is life satisfaction. The independent or predictor variables are usually signified by x_1 to x_k. One of the predictor variables in our example is nationality. Another predictor might be age. The contribution or weight of each predictor is normally the partial regression coefficient, which is generally symbolized as b_1 to b_k. The a is the intercept and may be referred to as the constant.

Multiple regression assumes that the predictor variables are dichotomous or noncategorical. Dichotomous variables (such as gender) have two categories (female and male) and may be treated as if they are noncategorical in that one category is arbitrarily assumed to be higher than another. For example, females may be coded 1 and males 2. This cannot be done with categorical variables having more than two categories because the numbers will be seen as reflecting an ordinal scale at the very least. For instance, if we coded Americans 1, Britons 2 and Canadians 3, multiple regression will assume that Americans have the highest value and Canadians the lowest, which might not be the case. We cannot order nationalities in terms of their mean score on life satisfaction (with Americans coded 1, Canadians 2 and Britons 3) because this order might not be the same for the other predictor variables (such as age). Consequently, we have to treat the categorical variable as if it were a series of dichotomous variables.

The simplest form of coding is *dummy coding*, where we assign a 1 to the units of analysis belonging to that category and 0 to units not belonging to that category. So, for example, we could code the three nationalities as shown in Table 2.3. Here we use one dummy variable to code all Americans as 1 and all non-Americans as 0. We use another dummy variable to code all Britons as 1 and non-Britons as 0. In this scheme Americans are represented by a 1 on the first dummy variable and a 0 on the second dummy variable. Britons are denoted by a 0 on the first dummy variable and a 1 on the second dummy variable. We do not need a third dummy variable to code Canadians because Canadians are represented by a 0 on both dummy variables. The category denoted by all 0s is sometimes known as the reference category. Thus, only two dummy variables are needed to represent these three categories.

Table 2.3 *Dummy variable coding of three nationalities*

Nationalities	d_1	d_2
American	1	0
British	0	1
Canadian	0	0

The number of dummy variables required to code a categorical variable is always one less than the number of categories. So, if there are four categories, three dummy variables are necessary. It does not matter which category is denoted by 1s and 0s. In our example, Americans could have been coded 0 0, Britons 1 0 and Canadians 0 1. The results for the dummy variables taken together will be exactly the same. If the reference category is also coded in 1s and 0s, then one less than the total number of dummy variables will be entered into the multiple regression because one of them is redundant. The reference category is represented by the intercept a in the regression equation. So, the multiple regression equation for regressing the criterion of life satisfaction on the dummy coded categorical variable of nationality is:

$$\underset{(y)}{\text{Life satisfaction}} = \underset{(a)}{\text{Canadian}}$$

$$+ \underset{(b_1x_1)}{b_1 \times \text{American}} + \underset{(b_2x_2)}{b_2 \times \text{British}}$$

The multiple correlation squared is 0.194, which is the same value as that for eta squared. Dummy coded variables representing a particular categorical variable need to be entered together in a single step in a hierarchical multiple regression analysis.

EFFECTS AND CONTRAST CODING

Two other ways of coding categorical variables are effects and contrast coding. Both these methods will explain exactly the same proportion of variance by the categorical variable as dummy coding. However, the partial regression coefficients may differ insofar as they represent different comparisons. If information on particular comparisons is also needed, the required comparisons have to be specified with the appropriate coding. With dummy coding, the constant is the reference category. In our example on nationality, the unstandardized partial regression coefficient for the first dummy variable essentially compares the mean life satisfaction of Americans with that of Canadians. Similarly,

Table 2.4 *Effects coding of three nationalities*

Nationality	e_1	e_2
American	1	0
British	0	1
Canadian	–1	–1

Table 2.5 *Contrast coding of three nationalities*

Nationality	c_1	c_2
American	1	–½
British	–1	–½
Canadian	0	1

the unstandardized partial regression coefficient for the second dummy variable compares the mean life satisfaction of Britons with that of Canadians. See Cohen and Cohen (1983) for further details.

With effects coding, the constant is the mean of all equally weighted group means, which is produced by coding one of the categories as –1 instead of 0, such as the Canadians as shown in Table 2.4. In this case, the unstandardized partial regression coefficient for the first effects-coded variable compares the mean life satisfaction of Americans with that of all three groups. The unstandardized partial regression coefficient for the second effects-coded variable contrasts the mean life satisfaction of Britons with that of the three nationalities.

Contrast coding enables other kinds of comparisons to be made provided that the comparisons are independent or orthogonal. As with dummy and effects coding, the number of comparisons is always one less than the number of groups. For example, if we wanted to compare Americans with Britons and Americans and Britons combined with Canadians, we would code the groups as indicated in Table 2.5. For the comparisons to be independent, the products of the codes for the new contrast-coded variables have to sum to zero, which they do in this case:

$$1 \times (-\tfrac{1}{2}) + (-1) \times (-\tfrac{1}{2}) + 0 \times 1$$
$$= -\tfrac{1}{2} + \tfrac{1}{2} + 0 = 0.$$

FACTOR ANALYSIS

Factor analysis is commonly used to determine the factorial validity of a measure assessed by several different indices. Factorial validity refers to the extent to which separate indices may be seen as assessing one or more constructs. Indices that measure the same construct are grouped together to form a factor. Suppose, for example, we were interested in determining whether people who said they were anxious were also more likely to report being depressed. We made up three questions for assessing anxiety (A1–A3) and three questions for measuring depression (D1–D3):

A1 I get tense easily
A2 I am often anxious
A3 I am generally relaxed

D1 I often feel depressed
D2 I am usually happy
D3 Life is generally dull

Each question is answered on a five-point Likert scale ranging from ‘Strongly agree’ (coded 1) through ‘Neither agree nor disagree’ (coded 3) to ‘Strongly disagree’ (coded 5).

The anxiety questions appear to ask about anxiety and the depression questions seem to be concerned with depression. If people can distinguish anxiety from depression and if people who are anxious tend not to be depressed as well, then answers to the anxiety questions should be more strongly related to each other than to the answers to the depression questions. Similarly, the answers to the depression questions should be more highly associated with each other than with the answers to the anxiety questions. If this turns out to be the case, the three items measuring anxiety may be combined together to form a single index of anxiety, while the three items assessing depression may be aggregated to create a single measure of depression. In other words, the anxiety items should form one factor and the depression items should form another factor.

However, the way the answers to these six questions are actually grouped together may differ from this pattern. At one extreme, each answer may be unrelated to any other answer so that the answers are not grouped together in any way. At the other extreme, all the answers may be related and grouped together, perhaps representing a measure of general distress. In between these two extremes the range of other possible patterns is large. For example, the two positively worded items (A3 and D2) may form one group of related

Table 2.6 *Coded answers on a 5-point scale to six questions*

Cases	A1(Tense)	A2 (Anxious)	A3 (Relaxed)	D1 (Depressed)	D2 (Happy)	D3 (Dull)
1	5	3	2	3	4	2
2	2	1	4	3	2	4
3	4	3	2	4	1	4
4	3	5	1	2	3	2
5	2	1	5	4	2	4
6	3	2	4	3	4	1

Table 2.7 *Triangular correlation matrix for six variables*

Variables	A1 (Tense)	A2 (Anxious)	A3 (Relaxed)	D1 (Depressed)	D2 (Happy)	D3 (Dull)
A1 (Tense)	1.00					
A2 (Anxious)	0.51	1.00				
A3 (Relaxed)	–0.66	–0.94	1.00			
D1 (Depressed)	–0.04	–0.61	0.51	1.00		
D2 (Happy)	0.33	0.22	–0.11	–0.59	1.00	
D3 (Dull)	–0.36	–0.45	0.29	0.63	–0.91	1.00

items and the remaining four negatively worded items may comprise another group of related items. We use factor analysis to see how the items group together.

Correlation matrix

The first step in looking at the way the answers are related to each other is to correlate each answer with every other answer. To illustrate our explanation we will use the small sample of fictitious data in Table 2.6. This table shows the coded answers of six people to the six questions on anxiety and depression. So, case number 1 answers 'strongly disagree' to the first question (A1) and 'neither agree nor disagree' to the second question (A2). Correlating the answers of the six cases to the six questions results in the triangular correlation matrix shown in Table 2.7.

Correlations can vary from –1 through 0 to 1. The sign of the correlation indicates the direction of the relationship between two variables. A negative correlation represents high scores on one variable (e.g., 5) being associated with low scores on the other variable (e.g., 1). For instance, from Table 2.7 we can see that the correlation between the answers to the questions about being anxious (A2) and being relaxed (A3) is –0.94. In other words, people who agree they are anxious have a strong tendency to disagree that they are relaxed (and vice versa). A positive correlation indicates high scores on one variable being associated with high scores on the other variable and low scores on one variable going together with low scores on the other variable. For example, in Table 2.7 we can see that the correlation between the answers to the questions about being tense (A1) and being anxious (A2) is 0.51. In other words, individuals who agree that they are tense have a moderate tendency to agree that they are anxious.

The strength of the association between two variables is indicated by its absolute value (i.e., disregarding the sign of the correlation). The correlation between being anxious and being relaxed (–0.94) is stronger than that between being tense and being anxious (0.51) because it is bigger. Conventionally, correlations in the range of 0.1 to 0.3 are usually described verbally as being weak, small or low; correlations in the range of 0.4 to 0.6 as being moderate or modest; and correlations in the range of 0.7 to 0.9 as being strong, large or high. The correlations in the diagonal of the matrix can be ignored or omitted as they represent the correlation of the variable with itself. This will always be 1.0 as there is a perfect positive relationship between two sets of the same scores.

From Table 2.7 it can be seen that the absolute size of the correlations among the three anxiety answers ranges from 0.51 to 0.94, suggesting that these answers go together. The absolute size of the correlations among the three depression answers ranges from 0.59 to 0.91, indicating that these answers go together. The data were deliberately generated to be associated in this way. In data that have not been so made up, the pattern may be less obvious. Even in these data, the pattern of results is not clear-cut. The absolute size of the correlation between being anxious (A2) and being depressed (D1) is 0.61, larger than the 0.51 between being tense (A1) and being anxious (A2). Furthermore, the correlation between being relaxed (A3) and being depressed (D1) is 0.51, the same as that between being tense (A1) and being anxious (A2). Consequently, it is possible that the answers to D1 may be more closely associated with the three anxiety items than with the other two depression items. Thus, the way the items are grouped may not be sufficiently apparent from simply looking at the correlations among the items. This is more likely to be the case the larger the number of variables. Factor analysis is used to make the way variables are grouped together more obvious.

Factor analysis is a set of statistical procedures that summarize the relationships between the original variables in terms of a smaller set of derived variables called factors. The relationship between the original variable and the factors is expressed in terms of a correlation or *loading*. The larger the absolute size of the correlation, the stronger the association between that variable and that factor. The meaning of a factor is inferred from the variables that correlate most highly with it. Originally, factor analysis was used to *explore* the way in which variables were grouped together. More recently, statistical techniques have been designed to determine whether the factors that have been obtained are similar to or *confirm* those that were either hypothesized as existing or actually found in another group. Consequently, when developing a series of indices to measure a variable, it may be more appropriate to use an exploratory rather than a confirmatory factor analytic technique. If we want to compare our results with those already obtained, then confirmatory factor analysis may be preferable.

Exploratory factor analysis

There are a number of different procedures for exploratory factor analysis. The two most commonly used are *principal components* and *principal factors* or *axes*. Factor analysis is the term used to describe all methods of analysis but may also refer to the particular technique called principal factors. In principal components all the variance in a variable is analysed. Variance is a measure of the extent to which the values of a variable differ from the mean. In principal components, this variance is set at 1.0 to indicate that all the variance in a variable is to be analysed. This will include any variance that may be due to error rather than to the variable being measured. In principal axes only the variance that the variable shares with all other variables in the analysis is analysed. This shared variance or covariance is known as *communality* and will be less than 1.0. Communality is also sometimes used to refer to the variance in principal components.

Often both procedures will give similar results so that it does not matter which procedure is selected. Tabachnick and Fidell (2001) have suggested that principal components should be used when an empirical summary of the data is required, whereas principal axes should be applied when testing a theoretical model. One problem with principal axes is that the communalities may not always be estimable or may be invalid (e.g., having values greater than 1 or less than 0), thereby requiring one or more variables to be dropped from the analysis. Consequently, we will use principal components to illustrate the explanation of factor analysis.

Initial factors The number of factors initially extracted in an analysis is always the same as the number of variables, as shown in Table 2.8. For each variable, the entries in the table represent its loading or correlation with each factor; the square of each entry is a

Table 2.8 *Initial principal components*

	1	2	3	4	5	6
A1 (Tense)	−0.62	0.39	0.67	−0.13	−0.02	0.00
A2 (Anxious)	−0.84	0.44	−0.23	0.20	0.09	0.00
A3 (Relaxed)	0.79	−0.61	0.10	0.05	0.03	0.00
D1 (Depressed)	0.77	0.26	0.54	0.21	0.04	0.00
D2 (Happy)	−0.69	−0.68	0.24	−0.08	0.11	0.00
D3 (Dull)	0.80	0.53	−0.15	−0.22	0.10	0.00
Eigenvalues	3.42	1.53	0.88	0.16	0.03	0.00
Eigenvalues as proportion of total variance	0.57	0.26	0.15	0.03	0.01	0.00

measure of variance. So, the variance of A1 is −0.62 squared, which is about 0.38. The amount of variance accounted for by a factor is called the *eigenvalue* or latent root, and is the sum of the squares of each entry in a column, that is, the sum of the variances for each variable. The first factor has the highest loadings and extracts or reflects the greatest amount of variance in the variables. It has an eigenvalue of 3.42. Subsequent factors represent decreasing amounts of variance. The second factor has an eigenvalue of 1.53, while the sixth factor has an eigenvalue of 0. The eigenvalues should sum to the number of factors, which in this case is 6 (allowing for rounding error). The variance that each factor accounts for can also be expressed as a proportion of the total variance. Thus, the first factor explains 3.42/6.00 = 0.57 of the total variance, and the second factor 1.53/6.00 = 0.26.

Number of factors to be retained Because the number of factors extracted is always the same as the number of variables that are analysed, we need some criterion for determining which of the smaller factors should be ignored as the bigger ones account for most of the variance. One of the main criteria used is the Kaiser or Kaiser–Guttman criterion, which was suggested by Guttman and adapted by Kaiser. This criterion ignores factors that have eigenvalues of 1 or less. The maximum variance that each variable explains is set at 1, so that factors having eigenvalues of 1 or less explain less variance than that of one variable on average. In other words, according to this criterion, only factors that account for the variance of more than one variable are retained for further analysis. In our example, only the first two factors have eigenvalues of more than 1, while the other four factors have eigenvalues of 1 or less. Thus, according to this criterion, we would keep the first two factors for further analysis. It should be noted that a cut-off at 1 may be somewhat arbitrary when there are factors which fall close to either side of this value. According to this criterion, a factor with an eigenvalue of 1.01 will be retained while one with an eigenvalue of 0.99 will be dropped, although the difference in the eigenvalues of these two factors is very small. In such cases it may be worthwhile extracting both more and fewer to see whether these factors, when rotated, are more meaningful than those retained according to Kaiser's criterion.

A second criterion is the graphical scree test proposed by Cattell (1966), who suggested that the Kaiser criterion may retain too many factors when there are many variables and too few factors when there are few variables. Child (1990) has specified 'many variables' as more than 50 and 'few' as less than 20. In the scree test the eigenvalue of each factor is represented by the vertical axis of the graph while the factors are arranged in order of decreasing size of eigenvalue along the horizontal axis, as shown in Figure 2.1.

Scree is a geological term for the rubble and boulders lying at the base of a steep slope and obscuring the real base of the slope itself. The number of factors to be extracted is indicated by the number of factors that appear to represent the line of the steep slope itself where the scree starts. The factors forming the slope are seen as being the substantial factors, while those comprising the scree are thought to be small error factors. The number

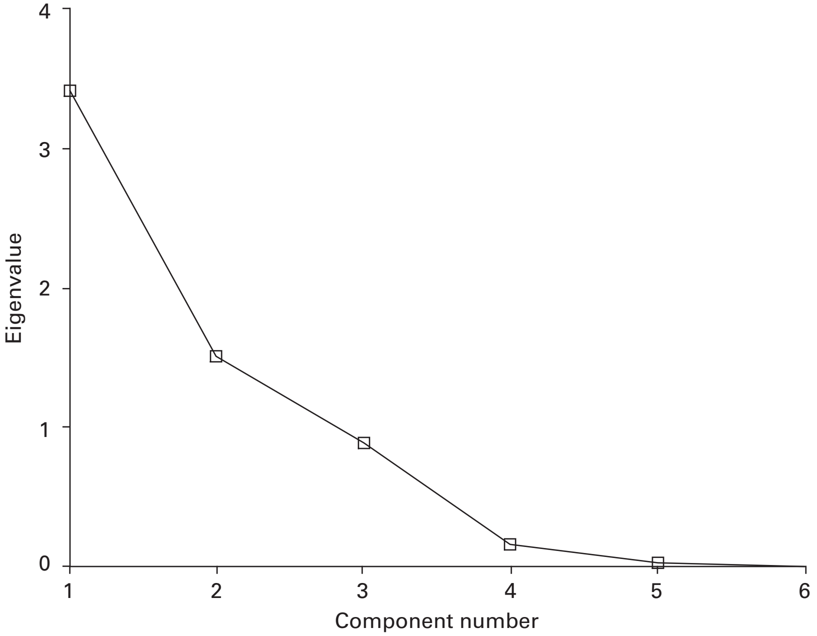

Figure 2.1 *Cattell's scree test*

of the factor identifying the start of the scree indicates the number of factors to be kept.

The scree factors are usually identified by being able to draw a straight line through or very close to their points on the graph. This is not always easy to do, as shown in Figure 2.1. In this case it is unclear whether the scree begins at factors 2, 3 or 4, and so whether the number of factors to be retained for further analysis should be 2, 3 or 4. Thus, one problem with the scree test is that determining where the scree begins may be subjective, as in this example. When this occurs, it may be useful to extract both fewer and more factors around the number suggested by the scree test and to compare their meaningfulness when rotated. If more than one scree can be identified using straight lines, the number of factors to be retained is minimized by selecting the uppermost scree.

Factor rotation As already explained, the first factor in a factor analysis is designed to represent the largest amount of variance in the variables. In other words, most of the variables will load or correlate most highly with the first factor. If we look at the absolute loadings of the variables on the first factor in Table 2.8, we see that they vary from 0.62 to 0.84. The second factor will reflect the next largest amount of variance. As a consequence, the loadings of the variables on the second factor will generally be lower. We see in Table 2.8 that they range in absolute value from 0.26 to 0.68. The loading of variables on two factors can be plotted on two axes representing those factors, as shown in Figure 2.2. These axes are called reference axes. In Figure 2.2 the horizontal axis represents the first factor and the vertical axis the second factor. The scale on the axes indicates the factor loadings and varies in steps of 0.2 from –1.0 to +1.0. The item on anxiousness (A2), for example, has a loading of –0.84 on the first factor and of 0.44 on the second (see Table 2.8).

It may be apparent that the two axes do not run as close as they could to the points representing the variables. If we were to rotate the axes around their origin, then these two axes could be made to pass nearer to these points, as shown in Figures 2.3 and 2.4.

The effect of rotating the axes is generally to increase the loading of a variable on one of the factors and to decrease it on the others, thereby making the factors easier to interpret. For example, in Table 2.9 we can see that the effect of rotating the two axes is to increase the loading of the item on anxiousness from –0.84 to –0.91 on the first rotated

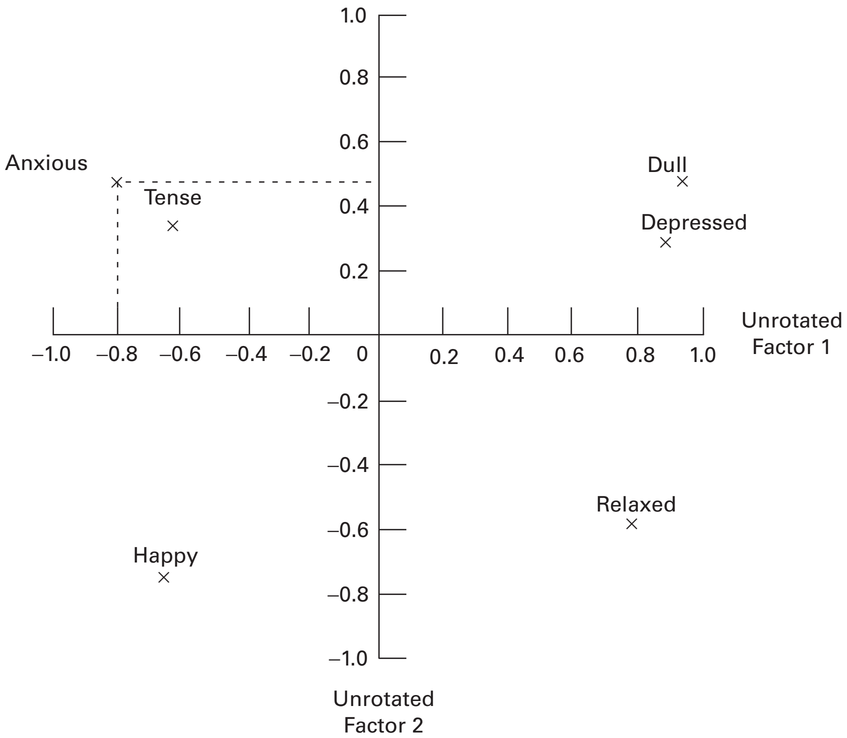

Figure 2.2 *Plotting variables on two unrotated factors*

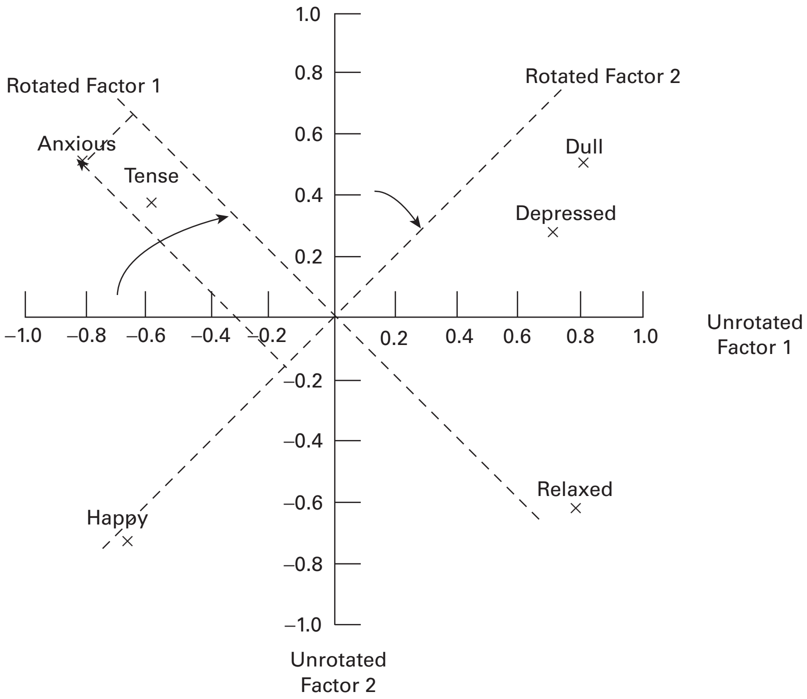

Figure 2.3 *Initial factors orthogonally rotated*

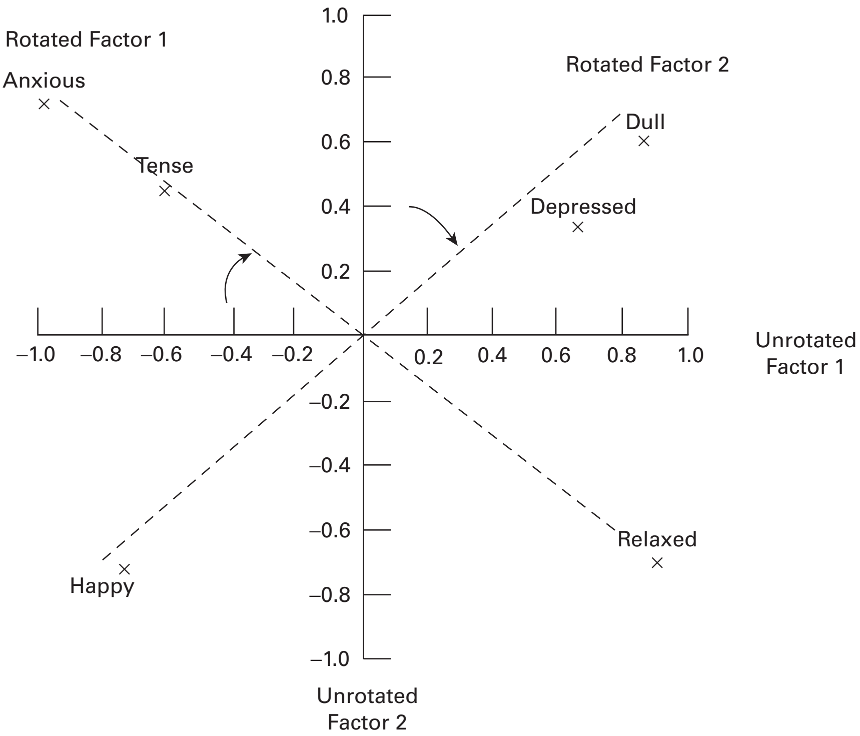

Figure 2.4 *Initial factors obliquely rotated*

Table 2.9 *First two orthogonally rotated principal components*

	1	2
A1 (Tense)	–0.72	–0.15
A2 (Anxious)	–0.91	–0.27
A3 (Relaxed)	0.99	0.11
D1 (Depressed)	0.37	0.72
D2 (Happy)	0.03	–0.96
D3 (Dull)	0.22	0.94
Eigenvalues	2.51	2.43
Eigenvalues as proportion of total variance	0.42	0.41

factor and to decrease it from 0.44 to 0.27 on the second rotated factor.

Axes may be rotated in one of two ways. First, they may be made to remain at right angles to each other, as is the case in Figure 2.3. This is known as *orthogonal* rotation. The factors are independent of or uncorrelated with one another. The advantage of this approach is that the information provided by the factors is not redundant. Knowing the values on one factor (e.g., anxiety) does not enable one to predict the values of another factor (e.g., depression) as the factors are unrelated. The disadvantage is that the factors may be related to one another in reality and so the factor structure does not accurately represent what occurs.

Second, the factors may be allowed to be related and to vary from being at right angles to one another, as illustrated in Figure 2.4. This is known as *oblique* rotation. The advantage of this method is that the factors may more accurately reflect what occurs in real life. The disadvantage is that if the factors are related, knowledge about the values of one factor may allow one to predict the values of other factors. The results of the two methods may be similar, as in this example.

The most widely used form of orthogonal rotation is *varimax*, which maximizes the variance within a factor by increasing high loadings and decreasing low loadings. The loadings shown in Table 2.9 were derived using this method. Comparing the results of Tables 2.8 and 2.9, we can see that orthogonal rotation has increased the loadings of three variables for the first (A1, A2 and A3) and second factor (D1, D2 and D3). It has decreased the loadings of three variables for the first (D1, D2 and D3) and second factor (A1, A2 and A3). The variables loading highest on the first factor are being relaxed (0.99), not anxious (–0.91) and not tense (–0.72) respectively, indicating that this factor represents anxiety. The variables loading highest on the second factor are not being happy (–0.96), finding life dull (0.94) and being depressed (0.72) respectively, showing that this factor reflects depression. These results suggest that the three items on anxiety (A1, A2 and A3) can be aggregated to measure anxiety and the three items on depression (D1, D2 and D3) can be grouped together to assess depression. Orthogonal rotation also has the effect of spreading the variance across the factors more equally. The variance accounted for by the first factor is 0.57 when unrotated and 0.42 (2.51/6.00) when rotated. For the second factor it is 0.26 when unrotated and 0.41 (2.43/6.00) when rotated.

The results of an oblique rotation using a method called *direct oblimin* are presented in Table 2.10. The findings are similar to those for varimax. The variables loading highest on the first factor are being relaxed (0.99), not anxious (–0.94) and not tense (–0.73), respectively. The variables loading highest on the second factor are finding life dull (0.96), not being happy (–0.95) and being depressed (0.78), respectively. The results indicate that the three anxiety items (A1, A2 and A3) can be combined together, as can the three depression items (D1, D2 and D3). The two factors were found to have a correlation of 0.36 with one another. As negative values on the first factor indicate anxiety and positive values on the second factor depression, the positive correlation between the two factors means that depression is associated with low anxiety. Because the factors are correlated, the proportion of variance explained by each factor cannot be estimated as it is shared between the factors.

Table 2.10 *First two obliquely rotated principal components*

	1	2
A1 (Tense)	–0.73	–0.28
A2 (Anxious)	–0.94	–0.43
A3 (Relaxed)	0.99	0.29
D1 (Depressed)	0.50	0.78
D2 (Happy)	–0.20	–0.95
D3 (Dull)	0.39	0.96

Combining items to form indices The results of the factor analysis are used to determine which items should be combined to form the scale for measuring a particular construct. Items loading highly on the relevant factor (e.g., anxiety) and not on the other factors (e.g., depression) should be used to form the scale. The direction of scoring for the scale needs to be established. Generally higher scores on the scale should indicate greater quantities of the variable being measured. For example, if the scale is assessing anxiety, it is less confusing if high scores are used to denote high anxiety rather than low anxiety. The numerical codes for the responses may have to be reversed to reflect this. For instance, the numerical codes for the anxiety items A1 and A2 need to be reversed so that strong agreement with these items is recoded as 5. The scale should have adequate alpha reliability. Items not contributing to this should be omitted.

CONCLUSION

In this chapter, we have moved fairly rapidly from some very basic ideas concerning variables to some fairly complex approaches to their creation and assessment. However, in another sense, the entire chapter deals with issues that are fundamental to the analysis of quantitative data, since the variable is the basic reference point. We have explored several ways in which variables are created, both in terms of such strategies as summated scales, which are common in the measurement of attitudes, and in terms of the ways in which analysts seek to refine and improve the quality of variables. Since the variable is fundamental to all quantitative data analysis, the material covered in this chapter constitutes an important starting point for many of the chapters in this book that deal with various aspects of quantitative data analysis.

REFERENCES

Blumer, H. (1956) 'Sociological analysis and the "variable"', *American Sociological Review*, 21, 683–90.

Bryman, A. (2001) *Social Research Methods*. Oxford: Oxford University Press.

Bryman, A. and Cramer, D. (2001) *Quantitative Data Analysis with SPSS Release 10 for Windows: A Guide for Social Scientists*. London: Routledge.

Cattell, R.B. (1966) 'The scree test for the number of factors', *Multivariate Behavioral Research*, 1, 245–76.

Child, D. (1990) *The Essentials of Factor Analysis* (2nd edition). London: Cassell.

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schuman, H. and Presser, S. (1981) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego, CA: Academic Press.

Tabachnick, B.G. and Fidell, L.S. (2001) *Using Multivariate Statistics* (4th edition). New York: HarperCollins.

Webb, E.J., Campbell, D.T., Schwartz, R.D. and Sechrest, L. (1966) *Unobtrusive Measures: Nonreactive Measures in the Social Sciences*. Chicago: Rand McNally.

Zeller, R.A. and Carmines, E.G. (1980) *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge: Cambridge University Press.

3

Summarizing Distributions

MELISSA HARDY

Statistical analysis is similar to any number of summarizing activities we perform each day. We describe a book as 'fascinating', a meal as 'delicious', a co-worker as 'kind'. In each case this single word communicates a central feature of the object which it describes, while ignoring much else. But we can expand our descriptions. For example, we can say that the book captured our attention in the first paragraph and held it to the last word, or perhaps that it took a few chapters to get into the story but thereafter was difficult to put down. We can recount the main story line, characterize the protagonist, discuss the use of language, liken it to other novels, and at some point, as a listener who had not read the book, you could gain an understanding of this text. So it is with statistical analysis.

When we work with a data set, our goal is to tell its story – or one of its stories. We use the data to formulate an answer to a question, to illustrate a point, to test a theory. And we need tools by which to accomplish these tasks. Staring at pages and pages of numbers, even numbers already organized into the necessary data matrices, will accomplish little. What we require are shorthand ways to represent the data arrays, a practice that helps us visualize what each variable 'looks like'. In general, we need methods of summarizing the information, and we need techniques of assessing how well and how consistently the summary suits the data. But any summary measure is designed to succinctly portray a specific feature of the data, not to provide a detailed description. Therefore, it is important to choose measures suited to the question at hand and to the nature of the data.

CLASSIFYING, COUNTING, AND MEASURING

The tools that we use must be suited to the type of information we wish to analyze. In the same way that a carpenter learns that different types of saws with different types of blades are best suited to cutting different sorts of materials, so the analyst learns that the first task is to identify the nature of the information at hand. Initially, we can categorize variables into two types: discrete and continuous. Discrete variables can take a finite number of values, whereas continuously measured variables can take on any value within a much more detailed range of possible values. In other words, the possible values for discrete variables are countable; the possible values for continuous measures

Table 3.1 *Descriptive statistics for variables in the NLSY79 data extract*

Variable name	Valid *N*	Mode	Median	Mean	Range	Variance	St. Dev.	Skew	Kurtosis
Gender	8889	1	1	0.5	1	0.25	0.5	–[a]	–[a]
Employment status	8889	1	1	0.84	1	0.134	0.366	–[a]	–[a]
Race/ethnicity	8889	3	–[a]	–[a]	3	–[a]	–[a]	–[a]	–[a]
Marital status	8884	1	–[a]	–[a]	6	–[a]	–[a]	–[a]	–[a]
Gender role attitude	8831	2	2	2.03	3	0.7	0.84	0.6	–0.11
Region of residence	8679	3	–[a]	–[a]	3	–[a]	–[a]	–[a]	–[a]
Age at interview	8889	31	33	32.98	8	5.01	2.24	0.135	0.052
Age at first marriage	6455	21	22	22.92	24	17.335	4.164	0.501	–0.231
Number of pregnancies	4411	2	2	2.34	14	3.09	1.76	0.834	1.234
Number of jobs	8882	6	8	8.8	44	27.78	5.27	1.05	1.784
Tenure at current job	7469	1	158	231.56	1045	49704.23	222.94	1.073	0.306
Highest grade completed	8884	12	12	12.98	20	5.98	2.44	0.172	1.319
Weight	8684	180	166	170.47	400	1548.71	39.35	0.823	1.393
Total net family income	7004	–[b]	33200	40883.12	189918	1.329E+09	36452.78	2.455	7.506
ln(family income)	7004	–[b]	10.41	10.12	12.15	2.6	1.61	–4.337	24.3

[a] Statistic is inappropriate for nominal variables.

[b] Measure is uninformative.

are not. Because of this difference, the statistical approaches to describing distributions of discrete and continuous measures utilize different branches of mathematics.

Among discrete variables, we also have different possible types. These types include nominal classifications, ordinal classifications, and counts. Continuous variables may be measured on either interval or ratio scales, the difference being that ratio scales have an absolute zero point. To facilitate our discussion of distribution statistics, we will rely on an exemplary data set extracted from the National Longitudinal Survey of Youth (NLSY) that was initiated in 1979. Table 3.1 lists the variables we will use in examples. Most of the variables are from the 1994 wave; a few are taken from the initial wave in 1979. The sample consists of men and women who were initially interviewed when they were aged 14–22. By 1994, the sample members were aged 29–37. In our data extract, respondents are classified by gender and by race/ethnicity. We know their employment status, marital status, and region of residence in 1994. We have one measure of gender attitudes taken in 1979, which records the level of the respondent's agreement/disagreement with the statement: 'A woman's place is in the home, not the office or the shop.' In 1994, we also know the highest grade of schooling they had completed, the number of jobs they had had, their age at first marriage (if ever married), the number of pregnancies (for the women), and number of weeks of tenure on their primary job (if employed). Finally, we know age at interview, weight, and total net family income. An example of a complete case is a 30-year-old African-American man who completed a bachelor's degree and was married at the age of 24 to a woman from whom he was divorced in 1993; the year 1994 saw him living in Massachusetts and employed in his second job, which he has held for 6 years; he disagreed with the statement that 'A woman's place is in the home', weighed 180 pounds, and had a total net income of $65 350.

Gender, employment, race/ethnicity, region, and marital status are all nominal classifications containing information that allows us to sort cases into categories. Gender and employment are binary items; the remaining variables in this list have several categories. Our attitude measure is ordinal. The remaining variables are treated as interval or ratio measures.

SINGLE-VARIABLE DISTRIBUTIONS

The most common way to display the pattern of observations for a given variable is to produce a frequency distribution, which displays the values of the observations, relative to the number of times each specific value is observed. In generating this distribution, we often use the standard geometry of the upper right quadrant of a two-dimensional space, which displays all positive values and is bounded below by the x-axis and to the left by the y-axis.[1] Here, the x-axis reports the case-by-case values of the variable, the y-axis the frequency of its observation. If the variable is measured continuously, then the frequency distribution is represented as a smooth curve. If the variable is a classification, then the frequency distribution is often a histogram, which displays vertical columns labeled by the category, rising to the number (frequency) of times it is observed.

Perusal of this simple type of distribution communicates much useful information. We view each value and each frequency relationally, within the full range of observed values (highest to lowest, if quantitative) and relative to how common or uncommon an observation is. In that way, we see the most likely observed value, and the range of possible values. We see the distributional form of each variable. Whereas a variable with a more limited number of possible category responses allows an observer to accurately assess this information through simple visualization, the more precisely a variable is measured, the less easily this is accomplished. With variables quantitatively measured, then, we require mathematics beyond simple counts to provide us with the summary information we desire.

Descriptive statistics

The statistics used to describe a distribution were developed to provide information about four features: a typical or most likely value in the distribution (a value at the midpoint of the distribution or one that is most often observed); the heterogeneity of the distribution (or the extent to which observations have different, perhaps widely different values); the symmetry of the distribution (whether observations are more heavily concentrated at values lower than the most likely value, higher than the most likely value, or equally divided between higher and lower values); and the peakedness of the distribution (or the extent to which observations are heavily concentrated around the most likely observation). The combination of these four types of measures provides a good picture of the entire distribution, which the researcher uses to decide how to proceed with further analysis. Therefore, regardless of the modeling technique or analysis approach that will ultimately be used to address research questions, the first step in analysis must always be to learn about the distributional properties of one's data.

Central tendency In summarizing an observed distribution, the most useful pieces of information tell us the most likely observed value and something about the differences among observed values, referred to as central tendency and dispersion. With classifications, the typical observation belongs to the category most frequently observed. We call this category the *mode* or *modal category*. It may happen that we have two or three categories observed equally frequently and more than any other, in which case we speak of the distribution as being bimodal or trimodal. Figure 3.1 shows the distributions of a subset of our variables.

Consider the bar chart for gender. Although the modal category is 'women', the sample is almost equally divided between men and women. Is it then correct to describe the 'typical' respondent as a woman? If the sample is (all but) equally divided between these two groups, such a description is misleading. If, on the other hand, we drew a simple random sample from a population with 70% women (as is the case at the oldest age ranges), our sample would be primarily older women. Turning to employment, again we have two categories, but in this case more than four of five respondents were employed at the time of interview; only 1422 (or 16%) were not. To describe the 'typical' respondent as

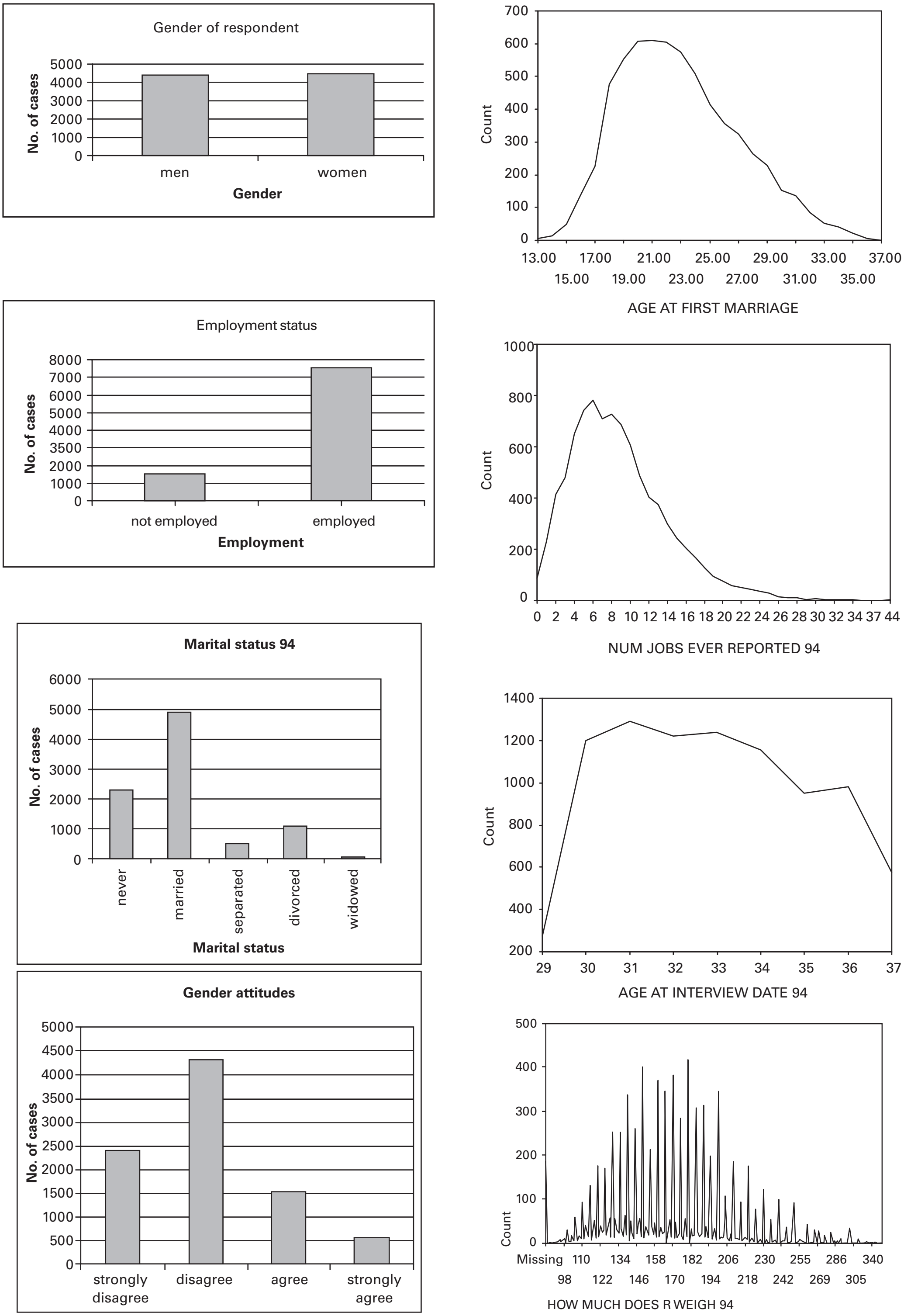

Figure 3.1 *Frequency distributions for binary, multinominal, and continuous variables*

'employed' is therefore appropriate. As for marital status, which is reported in five categories, the modal response is currently married at the time of the interview, a status which captures more than half the respondents.

The measure of gender attitudes allows for four responses, ranging from strongly disagree to strongly agree. Here the modal response is 'disagree', which captures almost half the respondents, so the 'typical' respondent disagrees with the statement: 'A woman's place is in the home, not the office or the shop.' Since these responses can be ordered, we can use a second measure of central tendency – the *median*. The median reports the middle value in an ordered array of numbers and is often referred to as the 50th percentile.[2] In this example, if we sorted (in ascending order) the data relative to responses on this question, the first 2420 (or 27.2% of) cases, coded 1, register strong disagreement. The following 4308 respondents (respondents numbered 2421 through 6728) disagreed with the statement. Given that 8831 respondents answered the question, the median value is associated with the 4416th case, 'disagree'.

Variables at a higher level of measurement can be characterized by the median or the mode, but a more useful measure is the *mean*. As precision of measurement increases, the mode becomes less and less useful, since the likelihood that multiple respondents are (precisely) 'the same' declines. In practice, commonality of responses on items that can be measured with precision is more likely a function of 'rounding' in the respondent's reporting.[3] The weight variable provides such an example here. One might have anticipated a symmetric, bell-shaped curve on weight. Instead, we see a set of spikes and toe holds, which undoubtedly result from people reporting their weight as 'roughly' 175, for example, rather than 176.882. The mean, as a measure of central tendency, is appropriate to interval and ratio data. It assumes that the values are meaningful quantities (rather than a shorthand way of denoting particular categories of qualitative information), and the mathematics of its formula uses this information of equally spaced numerical intervals in calculating its value, as in:

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

Special cases involve binary coded items (i.e., items coded 0 or 1) such as gender and employment status, in our extract. Mean values for these variables are reported in Table 3.1, but the information represented is actually the proportions of cases coded 1 (here, women and employed) rather than 0 on the variable in question. The binary coding transforms the summation into a counting process, which yields the frequency of 1s. The division by N relativizes the frequency for sample size, yielding the relative frequency of the category coded 1, or the proportion of the sample who are women (0.504) or who are employed (0.84).

Mean values for schooling, number of jobs, number of pregnancies, age at interview, age at first marriage, weight, net family income, and tenure are all reported in Table 3.1. In describing our sample, we would say that, on average, our respondents had completed 1 year of college beyond high school, had worked at 9 jobs, were aged 33 at interview, weighed 170 pounds, and had a net family income of $40 883. Women had experienced 2 pregnancies, on average. Among those ever married, the average age at first marriage was just shy of 23. Among those currently working, the average length of time with their primary employer was 232 weeks.[4]

How useful is this information? How accurate a picture do we now have of our sample of respondents? Answers to these questions require additional information about the distributions that the measures of central tendency were meant to describe. Are our respondents a relatively homogeneous group? Or are they widely divergent on the characteristics of interest to us? This is the issue of dispersion or variability.

Dispersion In the case of categorical measures, a measure of dispersion captures how observations are distributed across the various categories. First, we can visually inspect the charts in Figure 3.1 to draw some conclusions. Return to the graph of gender composition. We have two possible categories, and the distribution is virtually bimodal. In other words, we have two groups more or less equal in size. Does that suggest homogeneity (sameness) or heterogeneity (differences) among respondents in the sample? To put it another way, if a case were drawn at random from the sample, how confident are you that you could correctly guess the gender of the respondent? If your response is that you would be no more confident than in calling the toss of a fair coin, you would be correct.

Defining the limits of variability in a categorical measure leads us to the observation that, when all cases belong to a single category, the variance equals zero, which means our 'variable' is in fact a *constant*. But what defines the upper bound of variability? Imagine observations flowing from that single category into the second category, thereby increasing variability. At what point is variance maximized? Within this context, variability increases as the proportion of cases in the two categories moves to equality. Therefore, for gender and employment, we have less variability in employment status, more variability in gender in our sample. Again, since these items are binary, we can express the variance as:

$$s^2 = pq,$$

where p is the proportion coded 1, q its complement, and n the number of cases. Table 3.1 reflects this difference in variability: the variance for gender is 0.25, for employment 0.13.

The logic of a variance measure is the same when we have more than two categories. Variance is maximized when cases are equally distributed across all categories (the frequency graph of categories is rectangular). The *measure of qualitative variation* allows us to assess the degree of dispersion in nominal distributions. Based on the number of categories and their respective frequencies, the larger the number of categories and the smaller the differences in frequencies across categories, the larger the variance. The measure of qualitative variation compares the total number of differences in the distribution to the maximum number of possible differences for the same distribution. Calculation therefore requires an evaluation of observed differences relative to possible differences. The number of observed differences is

$$\text{Total observed differences} = \sum f_i f_j, \quad i \neq j,$$

where f refers to the frequency of categories i, j. We can calculate these differences for gender and for marital status, by way of illustration. For gender, the frequencies are 4409 for men and 4480 for women, so total observed differences are $4409 \times 4480 = 19\,752\,320$. To calculate the maximum number of possible differences (*MPD*), we use the formula

$$MPD = \frac{c\,(c-1)}{2}\left(\frac{n}{c}\right)^2,$$

where c is the number of categories and n is the number of observations. In the case of gender, we have

$$MPD_{Gender} = \frac{2 \times 1}{2}\left(\frac{8889}{2}\right)^2 = 4444.5^2$$
$$= 19\,753\,580.25.$$

The index of qualitative variation or IQV (Mueller and Schuessler, 1961: 177–9) is defined as the ratio of observed differences to maximum differences. Again, for gender, that ratio is

$$IQV_{Gender} = \frac{19\,752\,320}{19\,953\,580.25} = 0.9999,$$

which tells us that variability in gender in this sample is all but at maximum. Comparison of the value 0.25, which is the variance calculated as previously noted, to the maximum variance possible for a binary item, 0.5 × 0.5 = 0.25, shows consistency in the statistics.

Calculating *IQV* when there are more than two categories becomes an increasingly complex exercise, but it follows the same logic. In evaluating *IQV* for marital status, we use the same formula. To determine observed differences, we have 10 elements in the summation:

$$2327 \times 4915 + 2327 \times 513 + 2327 \times 1076 + 2327 \times 53 + 4915 \times 513 + 4915 \times 1076 + 4915 \times 53 + 513 \times 1076 + 513 \times 53 + 1076 \times 53 = 23\,964\,774.$$

For the denominator,

$$MPD = \frac{5 \times 4}{2}\left(\frac{8824}{5}\right)^2 = 10 \times 1776.8^2$$

$$= 31\,570\,182.4;$$

therefore,

$$IQV = \frac{23\,964\,774}{31\,570\,182.4} = 0.76.$$

A second issue raised by measures such as *IQV* is whether standardization clarifies or confuses the interpretation of the measure. With *IQV* as well as other like measures, the calculated level of diversity is expressed as a proportion of the maximum possible diversity, given the number of subgroups. Is this form of standardization desirable? Lieberson (1969) notes that when dealing with a single population at a single point in time, a standardized measure is appropriate. Also, if the researcher is making comparisons between different populations with the same number of qualitative subgroups in each population, either the standardized or an unstandardized index may be used. However, if the researcher is comparing two or more populations that differ in the number of qualitative subgroups, an unadjusted measure of diversity is preferable. In general, when the goal is to describe the actual level of diversity in a population, an unadjusted measure is preferred. Lieberson's diversity measure, A_w, is defined as the probability that randomly paired members of a population will differ on a specified characteristic.[5] Holding aside the modifications due to sampling without replacement and the standardization procedure just discussed, A_w is equivalent to the index of qualitative variation (Lieberson, 1969).

To calculate A_w for marital status, which has five categories, we let P_k be the proportion of respondents in the first through fifth statuses, such that $P_1 + P_2 + P_3 + P_4 + P_5 = 1.00$. If we assume sampling with replacement (for simplicity's sake) the proportion of pairs with each possible marital status combination is the square of the sum of the proportions, or $(P_1 + P_2 + P_3 + P_4 + P_5)^2$. Expanding this polynomial gives us the following expression:

$$P_1^2 + P_2^2 + P_3^2 + P_4^2 + P_5^2 + 2(P_1P_2 \\ + P_1P_3 + P_1P_4 + P_2P_3 + P_2P_4 + P_2 + P_5 \\ + P_1P_5 + P_3P_4 + P_3P_5 + P_4P_5) = 1.00.$$

The proportion of pairs with a common marital status, *S*, is the sum of the squares for all marital statuses. In this example, *S* equals the sum of the first five terms. The proportion of pairs with a different marital status, *D*, is the sum of the remaining terms, or twice the bracketed expression. Using the same information for marital status from the data set, $S = 0.262^2 + 0.553^2 + 0.058^2 + 0.121^2 + 0.006^2 = 0.3925$, and A_w, the probability of different marital statuses, equals $1.00 - S = 0.6075$.

A final lesson here is that with qualities or characteristics that are classifiable, the more lop-sided the distribution – the higher the proportion of observations that fall in a single category – the less variability we have, and the better a descriptor the mode becomes. The more equally divided observations are across categories, the greater the variability (which is maximized when all categories are equal), and the less efficient is a measure of central tendency as a summary of the distribution.

For interval/ratio variables, we have several measures from which to choose. The simplest is the *range*, which reports the difference between the lowest value and the highest value. For age the range is 8 years, a function of sample design. The range for years of schooling is 20, for weight 400. The range gives us a sense of the magnitude of individual level differences we might observe, but also has some limitations. Suppose, for example, that in this sample we had one person who weighed 450 pounds and that the second highest weight was 300 pounds. A range of 400 suggests a level of variation that may be misleading in this case. A derivative measure, the *interquartile range*, reports the difference between the value associated with the 25th percentile and the 75th percentile. Nevertheless, both these measures use only two data points in the distribution.

For interval/ratio variables, we would prefer a measure that tells us about aggregate variation, a measure that utilizes information on every observation, as the mean does for measures of central tendency. The two most common measures of dispersion are based on deviations from the arithmetic mean of the distribution.[6] A deviation is a measure of difference, in this case the difference between an observed value and the mean. The sign of the deviation, either positive or negative, indicates whether the observation is larger than or smaller than the mean. The magnitude of the value reports how different (in the relevant numerical scale) an observation is from the mean. One of the features of the mean is that the sum of the deviations across all observations must always equal zero. Hence, the mean is often referred to as the center of gravity of a distribution – the balancing point. By using the absolute value of the deviation score, one can calculate the average deviation as:

$$\text{Average deviation} = \frac{1}{n}\sum |(X - \bar{X})|.$$

However, the most common measures of dispersion are the *variance* and the *standard deviation* (which is the square root of the variance). The sample variance is built on the same concept of deviation score, but in this case the deviations are *squared* (an important distinction between the 'average' deviation and the variance/standard deviation), and then summed over all cases and divided by $n - 1$:

$$\text{Variance} = s^2 = \frac{1}{n - 1} \sum (X - \bar{X})^2.$$

Subtracting 1 from n in the denominator is necessary to adjust for degrees of freedom, which is a count of the remaining pieces of information on which you have imposed no linear constraints. Because we use sample statistics to estimate population parameters, we must be vigilant about keeping track of the circumstances in which we must use sample information (in this case, $\bar{X}$) to calculate other sample estimates of parameters.[7] The variance is also referred to as the 'mean squared error' in the context of prediction error. The mean is also the general least-squares estimate of central tendency, which means that the sum of the squared deviations around it (the numerator of the variance formula) is minimized, i.e., smaller than around any other measure of central tendency or any other value in the distribution. Therefore, the mean becomes the 'best' predictor in the absence of any information beyond the variable's distribution.

The standard deviation, *s*, is found by taking the square root of the variance, which accomplishes a return to the original unit of measurement. Although its value is generally close to that of the *average deviation*, it should not be confused with it in discussions. So, for example, it is *not* correct to say that, given a standard deviation value of 39.35 for weight, respondents differ from mean weight by 39.35 pounds on average.[8]

According to *Chebyshev's theorem*, it is possible to calculate the minimum proportion of observations that will fall within k (where $k > 1$) standard deviations of the mean. The formula for making this calculation, $1 - 1/k^2$, is applicable to *any* distribution, regardless of its shape. For example, at least 75% (1 – 1/4) of the observations of a distribution will fall within ±2 standard deviations of the mean. In other words, knowing nothing about the distribution but its mean and standard deviation, one can say that at least 75% of all observations lie with a range of ±2 standard deviations around the mean; at least 89% lie within ±3 standard deviations around the mean; and at least 94% lie within ±4 standard deviations around the mean.

One final measure of dispersion is the *coefficient of variation*, which relativizes the size of the standard deviation to the scale of measurement for the variable by dividing it by the mean:

$$\text{Coefficient of variation} = V = \frac{s}{\bar{X}}\cdot$$

For example, the standard deviation for schooling is 2.44 and the standard deviation for weight is 39.35. How can we make sense of that magnitude of difference? Using the *coefficient of variation*, we have 2.44/12.98 = 0.108 for schooling and 39.35/170.47 = 0.231 for weight. Clearly, the relative dispersion from the mean is larger in the case of weight than it is for schooling, but not nearly as much larger as we may have initially believed. We could also calculate V for number of pregnancies (1.76/2.34 = 0.752) indicating that, although the standard deviations for schooling and pregnancies were fairly close, the dramatically different means suggest that relative variation for pregnancies is much higher.

Shape In addition to the midpoint of a distribution and some notion of the degree of heterogeneity among respondents, information about the shape of the distribution can also be quite useful. Two measures that describe distribution shapes are skewness and kurtosis.

Skewness describes the degree of symmetry in a distribution, where symmetry refers to the balance between the number of observations that are above the mean and the number of observations below the mean. If we have an equal number of observations above and below, and the distribution is unimodal, the distribution is also symmetric. Since equality of the number of observations on either side of the mean is equivalent to saying the mean and median of the distribution are equal, Pearson developed a coefficient of skewness based on the *difference* between the mean $\bar{X}$ and the median $\tilde{X}$:

$$\text{Pearsonian coefficient of skewness} = Sk = \frac{3(\bar{X} - \tilde{X})}{s}.$$

Since the numerator is the simple difference between the two measures of 'average' value, the measure of skewness is signed: a distribution can be either positively skewed or negatively skewed, with the sign indicating which tail of the distribution contains the smaller proportion of observations. The mean is greater than the median when extreme positive values pull the mean in the direction of the right tail. Since the mean, unlike the median, uses information on the specific value, rather than simply noting its rank among other observations, even a relatively small number of very large observed values can shift a distribution away from symmetry.

Another measure of skewness is reported in Table 3.2, which contains the four sample *moments* around the mean of a distribution. The first moment is the midpoint, around which the sum of the deviations equals zero. The second moment is the variance, or mean squared deviation around the mean. The third moment is the average of the cubed deviations around the mean, which measures skewness when divided by the cube of the standard deviation. Since the variance is based on squared deviations around the mean, it must always be positive (with the standard deviation being defined as the positive square root). Cubing the deviations around the mean reintroduces the sign, positive or negative, to the measure.

Imagine a distribution of 1,2,2,3,3,3,4,4,5. The mean, median and mode are all equal to 3. Variance is equal to 1.5. Skewness is equal to 0. Now change the 5 to 10. The median and mode remain 3, but the mean is now 3.56. Variance is 6.78. Skewness is 2.21. We have five observations less than the mean; three observations greater than the mean, and a skewness value that tells us this fact: we have fewer observations to the right of the mean than to the left of the mean.

The frequency distribution of total family income is an example from our data set of a distribution that is positively skewed, with a value of 2.455. Income distributions are frequently skewed, which is to say that if one uses the arithmetic mean of an income distribution as its 'midpoint', one is indeed describing the center of gravity of the distribution. But since the distribution is asymmetrical, that balance point is such that more than half the cases lie below the mean (values lower than the mean are sampled in greater density because they are more likely observations). The relatively rare but very large values to the right of the mean disproportionately influence the 'balance point'. The greater the difference between the median value and the mean value, the more skewed the distribution; the more skewed the distribution, the more necessary it becomes to provide two measures of central tendency. The median will always be the proportional midpoint of the observations, the point above and below which 50% of the cases fall. The mean will be the numerical midpoint of the observed values in the distribution, a different meaning of 'midpoint'. In skewed distributions, that distinction is very important. We could not say, for example, that half the respondents in our sample have family income in excess of \$40 883. We can say, however, that half the respondents in our sample have family income in excess of \$33 200, since this is the median value of the distribution.

Kurtosis tells us whether the distribution is very peaked around the mean, or whether it is relatively flat. It is based on the fourth moment around the mean of a distribution. Since the mean deviations are now raised to the fourth power, measures of kurtosis will always be positive. In addition, by summing the mean deviations raised to the fourth power, observations that are far from the mean receive much more weight than they do in the calculation of the variance. A very peaked unimodal and symmetric distribution with observations compactly distributed around the mean is called leptokurtic ($k > 3$). A flatter unimodal and symmetric distribution with observations more widely dispersed around the mean is called platykurtic ($k < 3$). Mesokurtic describes a distribution that is neither excessively peaked nor excessively flat ($k = 3$).

THE NORMAL DISTRIBUTION

One type of symmetrical distribution is the *normal distribution*[9] or *normal curve*, which is a

Table 3.2 *Relationship between features of a distribution and moments around the mean*

Moments of a random variable		
First moment (mean)	$\mu = \mathrm{E}\{x\}$	
Second moment (variance)	$\sigma^2 = \mathrm{E}\{(x-\mu)^2\}$	
Third moment (skewness)	$\gamma_1 = \frac{1}{\sigma^3}\mathrm{E}\{(x-\mu)^3\}$	
Fourth moment (kurtosis)	$\gamma_2 = \frac{1}{\sigma^4}\mathrm{E}\{(x-\mu)^4\}$	
Sample moments around the mean		
First moment	$m_1 = \frac{\sum (X-\bar{X})}{n}$	
Second moment	$m_2 = \frac{\sum (X-\bar{X})^2}{n}$	
Third moment	$m_3 = \frac{\sum (X-\bar{X})^3}{n}$	
Fourth moment	$m_4 = \frac{\sum (X-\bar{X})^4}{n}$	
Statistics of a distribution		
Arithmetic mean	$\frac{\sum X_i}{n}$	Midpoint of the distribution
Variance	$\frac{\sum (X_i-\bar{X})^2}{n-1}$	Measure of dispersion around the mean
Skewness	$\frac{m_3}{m_2^{3/2}}$	Measure of symmetry around the mean
Kurtosis	$\frac{m_4}{m_2^2}$	Measure of peakedness at the mean

distribution of particular significance in data analysis. The *normal distribution* is both symmetrical and bell-shaped, with the three measures of central tendency (the mean, median, and mode) equal to the same numerical value; skewness is zero; kurtosis is equal to 3. Although, in theory, the normal distribution is asymptotic to the axis, in practice, applications generally have a finite number of observations. The exact shape of the normal distribution is determined by two parameters, the mean and the standard deviation of the distribution; its probability density function (pdf) is defined as:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)\,[(X-\mu)/\sigma]^2}$$

where μ is the population mean of X, σ is the population standard deviation of X, $\pi = 3.14159\ldots$, and $\mathrm{e} = 2.71828\ldots$. This formula allows one to calculate the value of the expected frequency (or density) of observation associated with a given value of x for a normal curve specified by a particular mean and standard deviation.

In estimating probabilities, the normal distribution is particularly useful, since the area under the curve, given by the cumulative density function (cdf), allows us to estimate the probability of a given range of outcomes.

The total area under the curve, ranging from $-\infty$ to $+\infty$, totals 1:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1.$$

Because x is continuous and the probability is defined by area, it is not possible to assess the probability of a specific outcome value. Rather, one defines a range of values, which may be small or large depending on the question at hand, to determine the probability:

$$\int_{b}^{a} f(x)\mathrm{d}x = P(a < x\ b),$$

where $f(x)$ dx is the probability associated with a small interval of a continuous variable, the interval $[a, b]$.

Being able to locate a specific observation in the normal distribution therefore allows one to determine the empirical probability of values less than or greater than the observation of interest. This practice is limited to variables that are measured at the interval/ratio level, with normal distributions. However, since any given normally distributed variable will present its own mean and standard deviation, calculation of these probabilities (through integration of areas under the curve) would be tedious. By making an adjustment to an observed distribution that would set the mean and standard deviation at standard values, we could utilize a single normal distribution, which is, in fact, how we proceed.

This standardization procedure is generally called the *z-transformation*, and it is appropriate to normally distributed interval/ratio variables. Calculating a *z-score*, or transforming the observed empirical distribution into the standardized normal distribution, is accomplished by dividing the unit deviation by the standard deviation:

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

This produces a distribution with mean equal to 0 and a standard deviation of 1. By calculating the *z*-scores, we can immediately view each observation in probabilistic terms.

A positive *z*-score means the observation is higher than the mean, which automatically signals that the respondent scored higher than at least half the respondents in the sample. How many more than half the sample? Although Chebyshev's theorem could prove useful here, we have more information than that theorem requires. Chebyshev's theorem is silent on the shape of the distribution; therefore, it is applicable to all distributions of all shapes and sizes. We are now working with a particular type of distribution – a normal distribution. Using this additional information, we can be more precise about the proportion of observations that lie within the range of *k* standard deviations around the mean. The empirical rule can be applied here, allowing us to say that 68.3% of the values will fall between ± 1 standard deviation around the mean; 95.5% will fall between ± 2 standard deviations around the mean; and 99.7% will fall between ± 3 standard deviations around the mean.

The *z*-transformation allows us to take advantage of tables that report already calculated areas under the normal curve (see the Appendix at the back of the book), rather than having to evaluate integrals in each distinct normal distribution we observe. Since we are dealing with a continuous distribution, the probabilities we can assess must be bounded by two values; we cannot ascertain the probability of observing a specific discrete value. Working through a few examples should make this clear.

Table 3.3 shows a small portion of the *z*-distribution included in the Appendix. The range of possible values for *z* lies between $-\infty$ and $+\infty$. Since the variance of the distribution is also 1, the sum of the area under the curve is equal to 1 as well, which is the upper limit of a probability. Given that the curve is symmetric, the mean divides the area in two, with 0.5 between the mean and $+\infty$, and 0.5 between the mean and $-\infty$. Tables that report the area under the normal curve either report the area that lies between the mean and a given *z*-score or the area that lies between the given *z*-score and infinity. Both the extract in Table 3.3 and the complete table in the Appendix report the area between the mean and the *z*-score. Finally, only positive *z*-scores are reported in the table. Again the symmetry of the curve allows the reader to determine the area between negative *z*-scores and the mean as easily as between positive *z*-scores and the mean.

Consider the quantitative portion of the Graduate Record Examination (GRE), one of the exams often required for admission to graduate school in the US. The highest possible score is 800. Suppose in a given year that the mean score was 480, the standard deviation was 100, and your score was 600. You can convert your score to a *z*-score by dividing 120 by 100, which is 1.2. Your score is 1.2 standard deviations above the mean, so you know you scored better than more than half of those who took the exam. How much better? Consider the extract from the *z*-distribution in Table 3.3. You find the area associated with your *z*-score by looking in the row headed 1.2 and the column headed 0.00. The value is 0.3849, which describes the area between the mean, 0, and your score, 1.2 (Figure 3.2 (a)). Add to that the area in the other half of the distribution, 0.5, and you gain the information that you scored better than 88.49% of those taking the exam. To determine the probability that someone picked at random scored better than you, simply subtract 0.8849 from 1: 0.1151 is your answer.

As a second example, consider a score of 350, which converts to a *z*-score of −1.3. The area between the mean and a *z*-score of 1.3 is 0.4032, therefore the area between the mean and a *z*-score of −1.3 is also 0.4032. In this case, 90.32% scored better than 350, and 9.68% scored worse (Figure 3.2 (b)).

As a final example, consider the area under the normal curve that corresponds to the difference between scoring 325 and scoring 628. The corresponding *z*-scores are −1.55 and 1.48 (Figure 3.2 (c)). The corresponding areas from the table are 0.4394 and 0.4306. Based on that information, we can say that 87% scored between 325 and 628; the probability that someone at random scored better than 628 is 0.0694; the probability that someone at random scored worse than 325 is 0.0606.

Table 3.3 *Extract from the table of the z-distribution*

	Second decimal place in *z*									
z	*0.00*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767

JOINT DISTRIBUTIONS AND MEASURES OF ASSOCIATION

An understanding of the main features of univariate distributions is an important preface to answering questions of relationships between or among variables. It is this notion of 'relationship' that is often of primary interest. Bivariate relationships can be assessed for different types of variables, thereby generating information about 'total' or 'gross' effects. But the complexity of the research questions we ask often requires us to assess 'partial' or 'net' relationships between variables. So how do we move from the characteristics of single distributions to those of joint distributions? One place to begin is with bivariate distributions.[10]

Bivariate distributions

Interval/ratio variables Given that we ended the last section with *z*-scores, observed values transformed into values from the standard normal distribution, let us begin this section with two standardized variables, Z_1 and Z_2, which are the *z*-transformed values for X_1 and X_2. The distributions of Z_1 and Z_2 are standard normal, with mean equal to 0 and variance and standard deviation equal to 1. What does their *bivariate* distribution, or *joint* distribution, look like? Somehow this third distribution must incorporate information from both univariate distributions in such a way that we can make judgments about whether the two variables are related, and if so, how they are related.

What does it mean to say two variables are related? We know what it means to say two people 'are related'. They belong to the same family: if 'closely related' they stem from the same portion of their 'family tree'. If 'distantly related', the branches of their respective nuclear families diverged some number of generations ago. Therefore, a close relationship can indicate a shared genetic structure (in a biological sense) but also shared likes and dislikes, similar attitudes, preferences, behaviors, mannerisms and so on (in a social sense). It also implies a certain predictability, which is the major reason family medical history is collected by physicians. So how do we translate this commonplace notion of 'relationship' to statistics?

Begin with the notion of predictability. How can an observed value for one variable be predicted by the value for the second variable? Suppose these values were equal. Suppose that, for respondent after respondent, the numerical value for Z_1 is the same as the numerical value for Z_2. This situation would allow us to perfectly predict the distribution of Z_2, if we knew Z_1 values, and vice versa, since each pair of observed values contains two identical numbers. To say that we can perfectly reproduce a second distribution by utilizing knowledge of the first distribution is to say that the two variables are related – perfectly related. Can we translate that statement into a statistic? We want to summarize, on a case-by-case basis, how the distributional position of the value of Z_1 corresponds to the distributional position of the value of Z_2. Recall that the variance measures heterogeneity in a univariate distribution by summing the squared mean deviations, case by case, and dividing by $n - 1$. Since *z*-scores are themselves indicators of distributional position (e.g., 1 standard deviation above the mean, 2.3 standard deviations below the mean), the variance of Z_1 and Z_2 would be written as $(\sum Z_1^2)/(n - 1)$ and $(\sum Z_2^2)/(n - 1)$, respectively.[11] What we want in this case is a measure of the mean difference

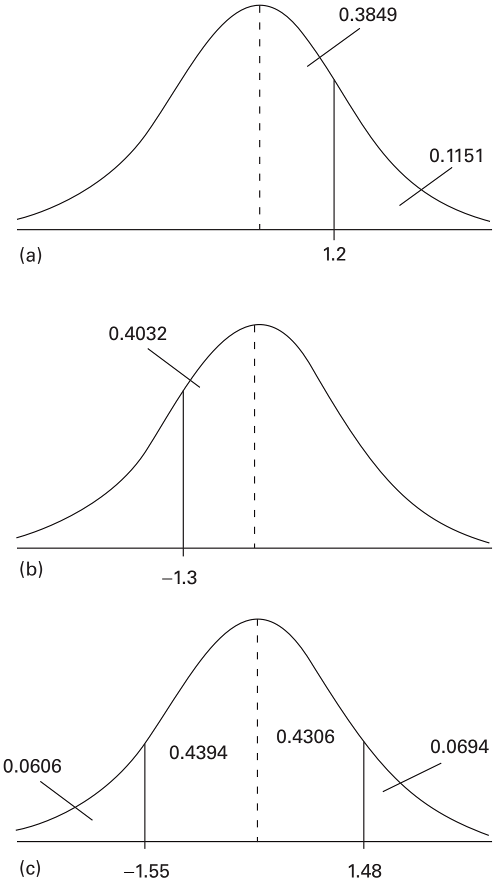

Figure 3.2 *Calculating probabilities of various outcomes using the area under the normal curve*

in the product of relative placement in the two distributions. That measure is named the *covariance* of Z_1 and Z_2 and is given by

$$\mathrm{cov}(Z_1, Z_2) = \frac{\sum Z_1 Z_2}{n-1}.$$

If we return to our invented data set of nine cases (1,2,2,3,3,3,4,4,5) and create values on a second variable as 5,10,10,15,15,15, 20,20,25, then transform both distributions to z-scores, we have the two identical distributions in Z_1 and Z_2. Calculating the covariance by summing the products across all cases and dividing by $n - 1$ yields a value of 1 (Table 3.4).

Since we have now linked the value of 1.00 with a 'relationship' of identity, we have also established a limit on the positive value of the covariance between two *z*-distributions. If we reverse the signs of observed values for Z_2 to produce Z_2^* and repeat the calculation of the covariance, we get a value of −1.00. What does this mean? It means that on a case-by-case basis, the relative position in the distribution of Z_1 is the reverse of, or opposite to, the position of the respondent in Z_2^*. If, as before, we assume these scores to be evaluations of respondents' performances on two tests, we can say, for example, that if the ninth respondent performed better than 94.84% of the sample on measure X_1, he performed worse than 94.84% of the respondents on X_2. If the third respondent performed better than 20.62% of the respondents on X_1, she scored worse than 20.62% of the respondents on X_2.[12] The other limiting value of the covariance between two *z*-distributed variables is −1.00, which indicates a perfect negative relationship, a predictability of one outcome to its opposite.

But what if we were not using *z*-scores? What if we were using values in their observed metric? In that case, we would use the more general formula for the covariance, which is

$$\mathrm{cov}(X_1, X_2) = \frac{\sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n-1}.$$

If we return to the original metrics of X_1 and X_2, our covariance is 7.5. What does that mean? When dealing with various measurement units on different scales, dealing with just the covariance tells us something about how the two variables are related (e.g., whether positive or negative), but the strength of the association is ambiguous, because we lack defined limits for each pair of variables. The advantage of assessing the bivariate distribution of two *z*-distributed variables is that both distributions have been standardized to means of 0 and standard deviations of 1. In other words, through the *z*-transformation, we had incorporated into each observation the information of the first and second moments of each univariate distribution, thereby producing a set of values already standardized on this distributional information. Therefore, it seems only reasonable that if we calculate the covariance of two distributions in their original metrics, we then apply some kind of distributional adjustment to

Table 3.4 *Demonstration data for z-scores, covariance and correlation*

Case no.	X_1	X_2	Z_{X1}	Z_{X2}	$Z_{X1}Z_{X2}$
1	1	5	−1.63	−1.63	2.67
2	2	10	−0.82	−0.82	0.67
3	2	10	−0.82	−0.82	0.67
4	3	15	0	0	0
5	3	15	0	0	0
6	3	15	0	0	0
7	4	20	0.82	0.82	0.67
8	4	20	0.82	0.82	0.67
9	5	25	1.63	1.63	2.67
Sum	27	135	0.00	0.00	8.00
Mean	3	15	0	0	
St. Dev.	1.22	6.12	1	1	
Variance	1.5	37.5	1	1	
Skew	0	0	0	0	
Kurtosis	−0.29	−0.29	−0.29	−0.29	
$\sum(X_1-\bar{X}_1)(X_2-\bar{X}_2)$	60	$\sum Z_1Z_2$	8		
$\text{cov}(X_1, X_2)$	7.5	$\text{cov}(Z_1, Z_2)$	1		
$r_{X_1X_2}$	1	$r_{Z_1Z_2}$	1		

again move us to a standard metric. We need a joint adjustment for the two distributions, and we accomplish that by dividing the covariance by the product of the two standard deviations. So, for example, if we divide the covariance of X_1 and X_2, 7.5, by the product of the two standard deviations, 1.225 × 6.124 = 7.5, we reproduce the value of 1.

This standardized measure of the covariance is, in fact, *Pearson's product moment*[13] *correlation coefficient (r)*, one of the most commonly used measures of linear association for interval/ratio variables, and is defined as the ratio of the covariance to the product of the standard deviations. Pearson's *r* can also be transformed into a *proportional reduction of error* (PRE) measure of association, which returns us to the notion of predictability. PRE measures of association are a special class of measures that indicate how much the error in prediction of one variable can be reduced by knowing the value of the other variable. Since such a proportion must always be positive and because the limiting values of *r* are −1.00 and +1.00, we know that *r* itself cannot be a PRE measure. But r^2 is, with a range of 0 to 1: information on a second variable can reduce your prediction errors not at all, can reduce them to zero (100%), or by any amount between the two.

In addition, some measures of association are symmetrical. Symmetrical measures of association assume no causal direction to the relationship, whereas asymmetrical measures assume that one variable depends on the other. Asymmetrical measures therefore make a distinction between dependent and independent variables, and the mathematical value of the measure incorporates this assumption. Symmetrical measures use the information of both variables in exactly the same way. Reviewing the measures already introduced, we can see that the covariance, *r*, and r^2 are all symmetrical measures.

We can now return to our data extract and explore the bivariate distributions of our interval/ratio variables. Table 3.5 contains information for the interval/ratio variables, including zero-order[14] (bivariate) correlation coefficients, *r*, the covariance, and the pairwise number of cases.[15] Among things to note are that *r* and the covariance always have the same sign; that *r* ranges between −1 and +1; that larger values of the covariance do not imply larger values of *r*. On this latter point, note as examples the bivariate relationships between income and weight, between age at first marriage and weight, and between age at interview and weight. The covariance between income and weight is very large (more than 33 000) and the correlation coefficient is quite small (0.023) – about as small as the correlation coefficient between age at interview and weight, which has a covariance less than 2. The correlation between age at first marriage and weight, 0.11, is notably larger (although still not what we would call 'large' in an absolute sense), with a covariance in the teens. The point is that the size of the

Table 3.5 *Zero-order correlation matrix for variable in data extract*

	weight	tenure	no. job	schooling	age (int)	age (fm)	no. preg.	
tenure	0.06							
	519.1							
	7298							
no. jobs	0.024	–0.468						
	5.01	–547.4						
	8677	7464						
schooling	–0.044	0.062	0.096					
	–4.223	33.13	1.242					
	8679	7465	8877					
age (int)	0.021	0.161	–0.118	0.011				
	1.817	80.533	–1.395	0.063				
	8684	7469	8882	8884				
age (fm)	0.11	0.044	0.075	0.325	0.045			
	17.83	41.648	1.623	3.311	0.422			
	6299	5507	6448	6452	6455			
no. preg.	0.039	–0.138	–0.086	–0.293	0.146	–0.268		
	2.494	–50.12	–0.744	–1.253	0.577	–1.855		
	4227	3502	4409	4408	4411	3371		
income	–0.023	0.192	–0.043	0.37	0.079	0.206	–0.108	*r*
	–33023.2	1573591	–8216.5	33031.7	6443	31613.3	–6795.6	covariance
	6882	5950	6998	6999	7004	5250	3514	no. of cases

covariance, being influenced by the scale on which the variable is measured (e.g., the range of values), tells us only the sign of the relationship. If we want a measure of the strength of association, we must use *r*, since it is a standardized measure.

The closer the value of *r* is to its limits, the stronger the relationship; the closer the value of *r* is to zero, the weaker the relationship. But if we want to discuss strength of association in the text of a report, the preference is to use r^2, since it is a PRE measure. For example, the correlation between age at first marriage and schooling is 0.325, which indicates that more than 10% ($0.325^2 = 0.1056$) of the variance in schooling can be explained (accounted for) by age at first marriage. The relationship is positive; therefore, we can say that respondents who married at older ages achieved higher levels of schooling, on average, than those who married at younger ages. The qualification 'on average' is an important component of the statement. We are *not* claiming that if we compared, one by one, those who married at younger ages with those who married at older ages, we would find no case in which the respondent who married at the younger age had more schooling. We are, however, claiming that if you calculate *mean schooling* for each value of age at first marriage, as age at first marriage increases, so would mean schooling. A cruder way of testing this statement would be to bisect the distribution of age at first marriage into two groups: those who married at age 20 or younger and those who married when they were older than 20. If we do so, we find a mean of 11.9 for those who married 'young' and a mean of 13.52 for those who were 'older' when married. As we can see, the mean for the 'young' group is less than that for those who married later.

Another important caveat is to note that we cannot claim that the level of schooling is *caused* by age at marriage. The causal direction could be the reverse – age at marriage may have led to the level of completed schooling. Or there could be no causal relationship between these two features. Rather, both completed schooling and age at first marriage may be two outcomes of a more complex social process that we have not considered in this simple example. Demonstrating causal relationships requires more than establishing a statistical correlation (see Winship and Sobel, this volume).

Nominal/ordinal variables For variables that are classifications we must rely on a different set of tools to assess relationships. The logic is the same. We are interested in predictability from one set of information to a second, wondering whether having a certain quality makes more or less likely a particular

Table 3.6 *Bivariate distribution of gender and employment status, percentaged by column*

		Gender of respondent		
		Men (0)	Women (1)	Total
Employment status	***Count***	**495**	**925**	**1420**
0	*% within gender*	(11.2)	(20.6)	(16.0)
	Count	**3914**	**3555**	**7469**
1	*% within gender*	(88.8)	(79.4)	(84.0)
Total	***Count***	**4409**	**4480**	**8889**
	% within gender	(100)	(100)	(100)

Source: Author's calculations, using NLSY data

preference, or whether making one choice increases the likelihood of a particular second choice. The bivariate distribution for interval/ratio measures was represented by the covariance. Bivariate distributions of categorical variables are represented through cross-tabulations. Using variables from our data extract, we could ask whether employment rates were different for men and women. Essentially, this is a question of proportions. We know that 49.6% of our sample are men, 50.4% are women, and that 84% are employed and 16% are not. If gender and employment are *not* related, what should we expect? The absence of a relationship suggests uniformity of outcome, that employment among women is no more or less likely than employment among men. In other words, the proportion of respondents who are employed does not depend on (i.e., does not differ by) gender. That suggests that if we limit our attention to women and calculate proportion employed and not employed, we should find the same proportional distribution as for the sample as a whole: 84% and 16%. The same should apply to men.[16] If we produce the joint distribution of gender and employment, we assign each respondent to one of four groups: not employed men, employed men, not employed women, and employed women. Each of these 'groups' is represented by a cell in a 2 × 2 table, as in Table 3.6.

The four shaded cells in the table display the joint distribution of gender by employment. Each cell is associated with a particular pair of categories on the two variables. For example, 495 respondents are both men and not employed; 3555 respondents are both women and employed. Within the body of the table, each respondent is jointly characterized on both variables. The column to the right of the shaded cells and the row below the shaded cells are the marginal distributions of our two original variables. We have 4409 men and 4480 women, 7469 employed persons and 1420 who are not employed. The bottom right cell gives us the total valid cases for these two variables, $n = 8889$.

If the likelihood of employment is not related to respondent's gender, the conditional distribution of employment by gender should be the same as the marginal distribution of employment, ignoring gender. Because we often do not have the same number of observations in each category, we cannot rely on frequencies to tell the story. Instead, we look at the proportional distribution *within categories of gender*. For example, 88.8% (3914/4409) of men are employed compared to 79.4% (3555/4480) of women. In our sample, then, employment is more likely among men than women.

Is there a statistic we can use to quantify that conclusion? There are several. Since this table is 2 × 2, we can use the phi (ϕ) statistic as a measure of association. Calculation of phi depends on a more basic statistic for cross-tabular analyses, chi-square (χ^2).

The chi-square statistic results from a comparison of the observed bivariate distribution with the bivariate distribution we would expect to see if, in fact, gender and employment were not related, that is, under the assumption of independence. Using probability theory, the definition of independence is that the marginal probability equals the conditional probability, as in:

$$\begin{aligned}\Pr(\text{EMP} = 1) &= \Pr(\text{EMP} = 1 \mid \text{Gender} = \text{male})\\ &= \Pr(\text{EMP} = 1 \mid \text{Gender} = \text{female}).\end{aligned}$$

In our example, the empirical probabilities are the same as the proportion of respondents satisfying a particular condition. The probability of employment compared to the probability of employment conditioned on gender is

$$\Pr(\text{EMP} = 1) = 0.840,€€$$

$$\Pr(\text{EMP} = 1 | \text{Gender} = \text{male}) = 0.888,€€$$

$$\Pr(\text{EMP} = 1 | \text{Gender} = \text{female}) = 0.794.$$

Clearly these three values are not equal, but can we quantify the extent to which they are different? We can begin by generating the frequencies we would expect if gender and employment were independent. We know what the conditional probabilities would be: 0.84 for men and 0.84 for women. If we have 4409 men and the probability of employment for men were 0.84, how many men would be employed? The answer is 3704.7.[17] Similarly, if we have 4480 women, 84% of whom should be employed (under independence), we should have 3764.3 employed women. We apply the same logic to determine the expected number of not employed persons, 16% of 4409 = 704.3 and 16% of 4480 = 715.7, and we have a complete set of expected frequencies.[18]

We noted earlier in the chapter that a measure of *variance* was a way to quantify differences among mathematical values, differences between the value observed and the value expected (the *mean*) if we had complete homogeneity. Here, we need a measure that will quantify differences between observed and expected (under the condition of independence) frequencies. We calculate the chi-square statistic as follows:

$$X^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(f_{\text{obs}} - f_{\text{exp}})^2}{f_{\text{exp}}}$$

where f_{obs} are observed frequencies, f_{exp} are expected frequencies, and the summation is performed across all cells, which is denoted by the double summation indicating across all rows and columns. In our example, X^2 equals 146.909, which allows us to calculate phi as:

$$\phi = \sqrt{\frac{146.909}{8889}} = 0.129.$$

In cases where both variables are dichotomous, ϕ is the equivalent of the correlation coefficient, *r*. Known as a *tetrachoric correlation coefficient*, it describes the relationship between two binary variables, which are observed indicators of an underlying latent variable. The latent variable is assumed to be normally distributed, but unobservable.[19]

The procedure is the same when the bivariate distribution requires more than four cells. However, as the number of cells increases the likelihood that some cells may have a zero or very low frequency increases, which creates a problem for the use of chi-square or any other type of cross-tabular analysis. Small expected frequencies can lead to very large values of chi-square, therefore chi-square should not be used if expected cell frequencies are smaller than 5.

Suppose, for example, we continued to be interested in the relationship between schooling and age at first marriage. Since both are discrete variables, we can look at their joint distribution, cell by cell. But schooling has a range of 20 and age at first marriage a range of 24. Therefore, the joint distribution is defined by a 20 × 24 matrix, or 480 cells. However, we always have the option of combining categories (values), or reclassifying according to some other conceptual scheme. Our interest may be in the effect of marrying at a young age, say age 20 or younger, versus marrying at a later age on completing schooling, with the expectation that marrying at a young age would be linked to less schooling. We can reclassify age at first marriage into two groups, and then compare the two groups' schooling distributions. But one could also argue that the primary interest is in obtaining educational credentials, so we can also reclassify years of completed schooling into five categories: 11 or fewer years, 12 years, 13 to 15 years, 16 years, and 17 or more years. The bivariate distribution of these recoded variables produces a 5 × 2 = 10 cell table.

When we compare the conditional distributions (within columns of age at first marriage) to the marginal distribution of schooling, we see that those who married at 20 or younger are overrepresented among high school dropouts and those with a high school diploma. Those who married later are overrepresented among those who went on to college, received a college degree, and continued postgraduate education. The value of chi-square for this table is 614.413. Because the table is larger than 2 × 2 and both variables are ordinal in their collapsed state, appropriate measures of association include Somer's *d* (0.355), gamma (0.499), Kendall's τ_b (0.277), and Spearman's correlation coefficient (0.303). All these measures rely on paired responses.

Gamma, also known as Goodman and Kruskal's gamma (Goodman and Kruskal,

Table 3.7 *Bivariate distribution of age at first marriage (dichotomized) and years of completed schooling, percentaged by column*

		Age at first marriage		
		≤ 20	> 20	Total
Schooling completed				
<12	***Count***	**435**	**392**	**827**
	% within age	(21.0)	(8.9)	12.8%
12	***Count***	**1132**	**1700**	**2832**
	% within age	(54.8)	(38.8)	43.9%
13–15	***Count***	**406**	**1126**	**1532**
	% within age	(19.6)	(25.7)	23.7%
16	***Count***	**63**	**716**	**779**
	% within age	(3.0)	(16.3)	12.1%
17+	***Count***	**31**	**451**	**482**
	% within age	(1.5)	(10.3)	7.5%
Total		2067	4385	6452
$\gamma = 0.499$	Somer's $d = 0.355$	$\tau_b = 0.277$	Spearman's rho = 0.303	

Source: Author's calculations, using NLSY data

1954), is a symmetric measure of association for ordinal variables based on the number of same-ordered pairs (N_s) and the number of different-ordered pairs (N_d). Tied pairs are not considered in the calculation of gamma. The coefficient is defined as

$$\gamma = \frac{0.5\,(N_s + N_d) - \min\,(N_s, N_d)}{0.5\,(N_s + N_d)}\ .$$

Calculation of γ requires a return to the table and an accounting of the different types of pairs. We begin in the upper left corner (435), which describes those who married young and have the least schooling. For same-ranked pairs, we move down and to the right in the table, since respondents in the four cells below right all married later and completed more schooling than our initial 435 respondents. Hence, the first element in our summation of same-ranked pairs is 435 (1700 + 1126 + 716 + 451). Pursuing this same logic, we have three remaining elements in the summation: 1132 (1126 + 716 + 451) + 406 (716 + 451) + 63 × 451. Altogether, then, we have 4 834 846 same-ordered pairs.

To find the number of different-ordered pairs, we move to the upper right cell, with 392 respondents who married later but completed the lowest level of schooling. For pairs that share this difference in ranking (later on marriage but sooner on stopping school), we look to the cells down and to the left, since they are occupied by those who married younger, yet completed more schooling than our 392 respondents. Our first element in the different-ordered pairs is therefore 392 (1132 + 406 + 63 + 31). Remaining elements in the summation are determined by moving down one cell in the right-hand column and multiplying by the combined number of respondents in cells to the lower left. The three remaining elements in the summation are therefore 1700 (406 + 63 + 31) + 1126 (63 + 31) + 716 × 31, which sums to 1 617 784 different-ordered pairs. We then calculate γ by substituting into the formula, using 1 617 784 for the second term in the numerator (since it is smaller than 4 834 846) and generate a value of 0.4986. Ranging between –1.00 and +1.00, γ also allows a PRE interpretation. A second formula for γ follows the same logic, but combines the information in a somewhat simpler way:

$$\gamma = \frac{N_s - N_d}{N_s + N_d},$$

which, in our example, would give us 3 226 315/6 443 377 = 0.4986.[20]

Kendall's τ_b addresses this limitation by amending the formula for γ by including tied pairs in the denominator:

$$\tau_b = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_y)\,(N_s + N_d + T_x)}},$$

where T_y and T_x are the number of pairs tied on y and x, respectively. In our example, schooling is y and age at first marriage is x. We find the number of ties on y by multiplying across columns, within rows, such that

$$435 \times 392 + 1132 \times 1700 + 406 \times 1126 + 63 \times 716 + 31 \times 451 = 2\ 611\ 165 = T_y.$$

To find pairs tied on x, we move across rows but within columns, such that

$$435\ (1132 + 406 + 63 + 31) + 1132(406 + 63 + 31) + 406\ (63 + 31) + 63 \times 31 + 392\ (1700 + 1126 + 716 + 451) + 1700\ (1126 + 716 + 451) + 1126\ (716 + 541) + 716 \times 451 = 8\ 416\ 351 = T_x.$$

The final calculation for Kendall's τ_b is:

$$\tau_b = \frac{3\ 226\ 315}{\sqrt{9\ 054\ 542 \times 14\ 859\ 728}} = 0.278,$$

which is much closer in value to those reported for Somer's d and for Spearman's rho. Since the tied pairs amend the denominator, the value of τ_b will never be greater than γ, although it may be equal to γ in cases where there are no tied pairs. When tied pairs are present, τ_b will always be smaller than γ, with the difference increasing as the number of tied pairs increases. τ_b is also a symmetrical measure of association, ranging from −1 to +1.

Somer's d is a measure of association for ordinal variables, which is also a PRE measure. In this example, it indicates that somewhat more than 7% of the variation in education is accounted for by age at first marriage. Rather than including tied pairs on both x and y, Somer's d adds only pairs tied on y to the denominator, so that the calculation is:

$$d = \frac{3\ 226\ 315}{9\ 054\ 542} = 0.355.$$

Spearman's rho measures the degree of monotonic relationship between two ordinal variables. As the number of categories increases, Spearman's rho becomes a more useful measure, since it relies on a comparison of the rank ordering of respondents within the two distributions. Rank orderings that are quite similar produce high positive values of ρ_S; rank orderings that are opposite produce high negative values of ρ_S; and rank orderings that are unrelated produce values close to zero. It is defined by

$$\rho_S = 1 - \frac{6\sum d^2}{n(n^2 - 1)},$$

where n is the number of pairs of observations in the sample and d is the difference in the ranks of each pair (*not* Somer's d). In this example, the value of 0.303 indicates a positive relationship between the two variables (as age at first marriage increases, average schooling completed increases, as well). Also a PRE measure, the squared value indicates that approximately 9% of the variation in schooling is explained by age at first marriage.[21]

Another useful method for comparing the ordered distribution of two groups is calculating *the index of net difference* (Lieberson, 1976). Although researchers often compare means for different groups, or compare medians when the observed distributions are skewed, the index of net difference makes no assumptions about the distributional form for either group involved in the comparison and is most useful when the researcher is interested in a comparison between entire distributions. The Wilcoxon (1945, 1947) rank-sum statistic and its more general forms, the Mann–Whitney U test (for comparisons between two samples of unequal size) and the Kruskal–Wallis H test (for comparisons of more than two samples) were used frequently in the 1960s and 1970s; however, this set of statistics was less useful in comparing distributions with frequent ties (when the pairs have the same ranking within their groups, which occurs more frequently as the number of ordinal categories decreases).

To calculate the index of net difference, we assume two observed occupational distributions, for example, for groups A and B. We then randomly pair observations from these two groups, noting that sometimes the ranking in A exceeds the ranking in B; sometimes the ranking in B exceeds the ranking in A; and sometimes the rankings are equal (tied). We can express these outcomes in terms of probabilities, which sum to 1.00, since they exhaust the set of possible outcomes. The net difference is $ND_{AB} = \Pr(A > B) - \Pr(B > A)$. Ranking in value from +1 to −1, ND_{AB} will equal zero if the probabilities are the same. The existence of ties is reflected in the maximum ND_{AB}. If, for example, the $\Pr(A = B) = 0.60$, then the maximum value for $ND_{AB} = \pm 0.40$.

Elaborating relationships

Although bivariate relationships are a good starting point when you begin to analyze your data, many of the research questions we develop are more complex and therefore require *multivariate* rather than bivariate analyses. The remaining chapters in this volume will provide readers with a variety of approaches to more complex questions with different types of data. In this section, we will briefly explore what is meant by *partial* relationships, *intervening* variables, and *interaction*.

In assessing partial relationships, we add at least one more variable to the mix. Our intention is to re-examine the relationship of initial interest under a new set of conditions. The term 'partial' is used because we are interested only in that part of the initial relationship that continues to obtain once these new control variables are introduced. The new set of conditions consists in 'controlling' for the effects of additional variables. When we introduced the term 'relationship', we linked it to the covariance between two variables, which represented their bivariate distribution. As we add variables, it becomes more difficult to think in these terms so long as we try to think of all variables at once. If we knew a way to pull them apart, then perhaps our understanding would improve.

Let us return to our example of age at first marriage and schooling, measured as interval variables. We reported a zero-order correlation between the two of 0.325, which is quite close to the statistical estimates of association we calculated after collapsing the two measures into categories. To introduce partials, we will return to the interval metric and examine the process implied by 'partialling'. In fact, the term 'partialling' is descriptive of what we want to accomplish in terms of the covariance, or the semblance of bivariate distributions we have implied. Suppose we are particularly interested in the relationship between age of first marriage and schooling among women in the sample, and we wonder whether the number of pregnancies experienced could be involved in the earlier relationship we observed. It may be that marrying young need not necessarily interfere with schooling, but marrying young could imply a larger number of pregnancies, and it is the pregnancies that make continued schooling impossible. To address that question, what we want is a measure of association between schooling and age at first marriage (in this example, just for women), controlling for the number of pregnancies they have experienced.

What does it mean to say we want to 'control' for number of pregnancies? In the nomenclature of 'partialling out' the effect of pregnancies, we want to rid the covariance between schooling and age at first marriage of the potentially confounding covariance that is shared with number of pregnancies. Although we have already reported the set of zero-order correlations for the sample, in this case we need to reproduce the same correlations for women only. The zero-order correlation coefficient between schooling and number of pregnancies (–0.2609) tells us that 6.8% of the variance in schooling is shared with number of pregnancies, such that more pregnancies are associated with less schooling. Similarly, the correlation coefficient of age at first marriage and number of pregnancies (–0.2681) indicates that 7.2% of the variation in age at first marriage is explained by (overlaps with) the variation in number of pregnancies, such that younger ages of marriage are associated with more pregnancies, on average. What we want to correlate is the remaining 93.2% of the variance in schooling with the remaining 92.8% of the variation in age at first marriage. Then we can see how that correlation coefficient compares to the zero-order coefficient we calculated. This new correlation coefficient is called a first-order partial because we are controlling for one additional variable.[22]

If we denote pregnancies as *P*, schooling as *S*, and age at first marriage by *AFM*, we can calculate the partial correlation coefficient as

$$r_{\mathrm{S,\,AFM.P}} = \frac{r_{\mathrm{S,\,AFM}} - r_{\mathrm{S,P}} r_{\mathrm{AFM,P}}}{\sqrt{(1 - r^2_{\mathrm{S,P}})(1 - r^2_{\mathrm{AFM,P}})}}\,.$$

The notation for the term to the left of the equal sign defines the first-order partial correlation between schooling and age at first marriage, controlling for the number of pregnancies the respondent has experienced. To the right of the equal sign we have a combination of all possible zero-order correlations – those between schooling and age at first marriage, between schooling and pregnancies, and between age at first marriage and pregnancies. Consider the denominator first. When we said earlier that we wanted the remaining 91.4% of the variance in schooling and the 92.8% of the variance that remains in age at first marriage, we were describing the kind of operations performed in the denominator. The expression in the first set of parentheses under the radical sign equals 0.914 (expressed as a proportion rather than a