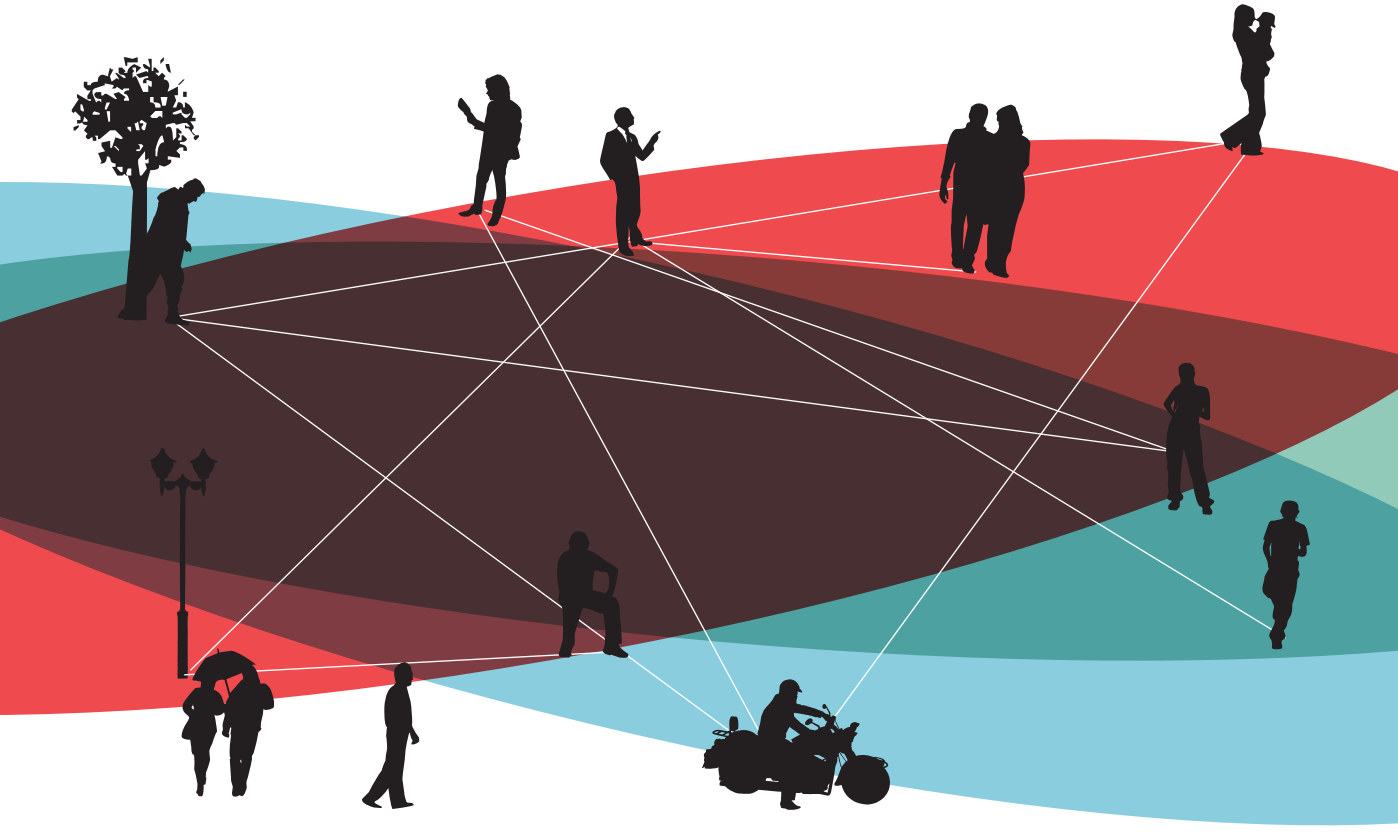
INTERPRETING QUANTITATIVE DATA

WITH IBM SPSS STATISTICS

2ND EDITION



RACHAD ANTONIUS

companion website

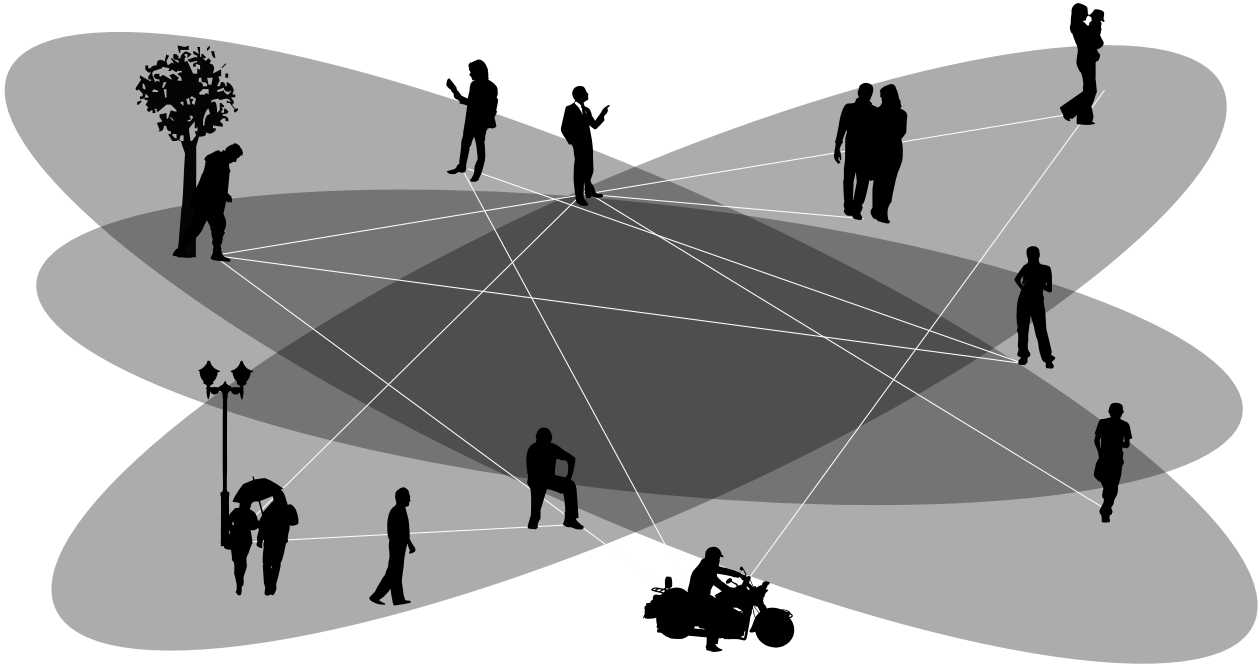
INTERPRETING QUANTITATIVE DATA

WITH IBM SPSS STATISTICS

INTERPRETING QUANTITATIVE DATA

WITH IBM SPSS STATISTICS

2ND EDITION



RACHAD ANTONIUS

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

Los Angeles | London | New Delhi
Singapore | Washington DC

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
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First published 2013

Editor: Katie Metzler
Assistant editor: Anna Horvai
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Proofreader: Jill Birch
Marketing manager: Ben Griffin-Sherwood
Cover design: Wendy Scott
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed by: MPG Books Group, Bodmin, Cornwall

Library of Congress Control Number: 2011945855

British Library Cataloguing in Publication data

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

ISBN 978-1-4462-0742-0
ISBN 978-1-4462-0743-7 (pbk)

BRIEF CONTENTS

CONTENTS

ABOUT THE AUTHOR

Rachad Antonius holds a PhD in Sociology and an MSc in Mathematics. The approach used in this book reflects both the rigour learned in the study of mathematics and the concrete skills learned in the practice of social research and in the teaching of its methods. He presently teaches Sociology at the *Université du Québec à Montréal* (UQAM) at the rank of Full Professor, and is the Deputy-Director of its Research Chair on Immigration, Ethnicity and Citizenship (www.criec.uqam.ca). He has taught quantitative methods at the graduate and undergraduate levels for many years, and the pedagogical approach used here has benefitted from this experience. In addition to the first edition of this book, he has co-authored a 600-page textbook in French titled *Méthodes quantitatives appliquées aux sciences humaines* (Montreal, CEC, 1991). His publications are mostly in the field of ethnic studies, but he has also published on the quantitative measurement of poverty, on the measurement of ethnic diversity, and has worked as a consultant on the quantitative analysis of inequalities for the UN Commission on Human Rights. He has also done extensive work in the Middle East with NGOs, and consultancy work for international agencies such as UNDP, UNICEF, CIDA, IDRC, and the Department of Foreign Affairs – Canada. He is often called in as a resource person by the main media networks in Canada.

He can be reached at: antonius.rachad@uqam.ca

FOREWORD TO THE INSTRUCTOR

This book is the result of many years of teaching in three fields: mathematics, sociology, and, more specifically, quantitative methods applied to the social sciences. Each subject has allowed me to develop a specific knowledge and specific teaching skills, which are reflected in this book. This second edition covers methods that are a little more advanced than the first, and it has been reorganized to make the learning process more efficient. I wish to explain here the logic that guided the choice of material included in the book, its organization, and the pedagogical principles on which my approach is based.

The material included

This book has been written for an introductory course in quantitative methods applied to the social sciences. The instructor can rely on it for presenting the material in class, and will find on the accompanying website (http://uk.sagepub.com/antonius2) additional tools and pedagogical tips. It makes a synthesis between statistics, research methodology, and the use of SPSS, and these three dimensions are integrated and combined to get the students to learn how to conduct a piece of quantitative research, create a data file, analyze the data, and write a satisfactory report. The pedagogical material found on the website will also be handy (some electronic slide presentations, additional data sets, answers to exercises, sample exams, etc.). Students will find in the book a good learning companion, as it is written in a clear and rigorous style. The SPSS tutorials will help them quickly master the basic functions of the software. The book can be read at two levels: the beginner will find it easy to understand, but a more advanced user will read much more into the same text.

The statistical material included is generally basic, but this second edition includes new chapters that are a little more advanced. It covers the elementary notions of descriptive statistics and of inferential statistics, up to statistical associations and their significance levels. No mathematical proofs are given, but the meanings of the formulas, and the concrete ways of interpreting them and of using them are explained in some degree of detail.

As in the first edition, I have insisted on the properties of the normal distribution, and on the fact that sampling distributions of the mean and of proportions are indeed normal, under some minimal assumptions. The logic of estimation and

hypothesis testing is all there: it is because we know where the means of 95% of the samples are likely to fall, when the population mean is given, that we can make an inference about that population mean, when it is not given. I believe that mastering these properties and linking them to graphical representations helps in making precise and rigorous statements when one is interpreting significance levels, or interpreting an estimate. However, I do agree that some students in social sciences may not be prepared to master these technical aspects, and that they could be skipped in an introductory course. I include them in Part III of the book titled 'Methodological Tools', which could be skipped or covered succinctly, depending on the audience to which the course is addressed.

The second edition differs from the first in the following ways. First, it uses version 19 of IBM SPSS.

Second, the material has been reorganized, with the basics of inferential statistics coming now before statistical association. This makes it possible to discuss the significance level of an association immediately after a measure of association has been explained for a sample.

Third, the chapter on statistical association found in the first edition has been expanded to three chapters, covering, respectively, correlation and regression, *t*-tests and one-way analyses of variance, and two-way tables and the chi-squared statistic. In each case, the meaning and measure of the association is explained and then the issue of the significance level is discussed. The distinction is clearly established between observing a statistical association on a sample, and determining whether it is likely to reflect a similar association at the level of the whole population. This distinction is often glossed over in many introductory courses, but we feel it is important to clarify it at this level.

Finally, the SPSS tutorials have been included right after the corresponding chapters, rather than grouped at the end of the book. Instructors will note that the SPSS tutorial that explains how to create a data file is given as an appendix. This way it can be inserted at any moment in the course, or left for the student to work through on his or her own. We suggest having the student do it after the chapter on descriptive statistics. This is because the appropriate statistical procedures to be used depend on the characteristics of the variables, and these characteristics must be factored in the very definition of the variable in the SPSS data file.

IBM SPSS is taught here not as an end in itself, but as a tool to illustrate and accompany statistical methods and procedures. The IBM SPSS tutorials at the end of most chapters should not be thought of as reference material for the software. Rather, they have been written in such a way as to allow the student to perform the calculations explained in the corresponding chapter, and to produce and interpret some of the basic outputs. My experience with teaching this material is that students tend to acquire enough familiarity with SPSS to be able to figure out on their own how to perform procedures that are not taught in the course. It should be pointed out that the book does not explain in detail the rules of the SPSS syntax: rather, I simply show how to paste in the Syntax window commands produced by pointing and clicking the mouse. This is very useful when the same analyses are performed on a modified data set. For instance, if some analyses are done on all

cases in a file, we may want to select a subset of cases (just men, say, or just women) and run the very same commands all at once.

A chapter on accessing online databases, which was included in the first edition, has been omitted in this one. Students are now quite literate web surfers, and search engines are very efficient. For instance, a search with the terms 'National statistics agencies' will give immediate results, including a page that gives dozens of direct links to national statistical agencies all over the world (http://www.statcan.gc.ca/reference/national-eng.htm). But I have included on the website of the book a list of links to statistical databases of interest for social research.

The material included is oriented toward getting the student to be able to write a first-level analysis of the data contained in a data file, including basic descriptive measures, some graphical illustrations, simple estimations, simple hypothesis tests, and some elementary statistical associations, together with their significance levels. The SPSS procedures explained in the tutorials have been determined in accordance with that aim.

An appendix has been included to suggest how a statistical analysis should be reported. Several references on this topic have also been given.

There has been a systematic attempt to situate the use of statistics within a comprehensive view of the research process, which means that the conditions under which a statistical method is used, its limitations, and the interpretation of the results it produces are briefly discussed as the method is presented. However, the material in this book is focused on quantitative methods, and any reference to the research process aims simply to situate the methods shown in their social science context. Ideally, students would be taking another course on research methodology, preferably after completing this course. In the testing stages of this manual, however, some students took the general methodology course either before this one or concurrently with it, and in both cases the results were satisfactory, as students were applying the methods learned here almost concurrently as they were learning them in the general research methods course.

Although the book does not discuss the philosophy of social science at all, I adopt an implicit non-positivist orientation, in the sense that I am careful to state that quantifying an issue is no guarantee of rigor. This approach implies an understanding of the limits of statistical methods, and an understanding of the fact that whatever is measured is not a faithful representation of 'reality', but rather a representation of the reconstruction of what we perceive as being 'reality'. We may hope that such representations are faithful, but that remains in the realm of wishes, and not always in that of science. What follows from this attitude is an increased care in expliciting the assumptions that lie beneath any conceptual construct, and a readiness to question claims about the validity of such constructs.

References

The statistical material included in this book is basic, and it has been part of the folklore of statistics and quantitative methods for a long time. Therefore, it did not make sense to give references for the basic formulas that were mentioned, such as

that of the standard deviation or of the margin of error when estimating a parameter. Instead, I have included in every chapter a section titled 'Suggestions for Further Reading' that could orient the reader to more advanced textbooks. The general bibliography contains all the suggestions for further reading of the various chapters, plus a few books that are either classics, or the source of mathematical proofs, or else that address very specific topics of interest.

Additional material

Some additional material has been posted on the website of the book. It includes additional exercises, some specific pedagogical tips concerning timing and presentation, and solutions to the exercises and tutorials.

In particular, I have included sets of slides with the diagrams that appear in the book, and slide presentations that explain some of the fundamental concepts. I have also included specific exercises that could be performed by the instructor in class, because I consider that they illustrate well some of the concepts, or some of the properties of the statistical measures shown in the book.

The pedagogical approach

What follows is an underpinning of five principles on which the pedagogical approach of the book is based. I have tried to apply them in a comprehensive way, now made possible by the dissemination of computer technology in the classrooms. The method presented here may achieve its full pedagogical potential when students have access on a weekly basis to a computer lab equipped with the SPSS program, or have acquired the academic version of SPSS. It is also quite useful to have access to a multimedia projector in the classroom, linked to an SPSS-equipped computer. Under these conditions, every theoretical discussion can be illustrated by a direct display of SPSS data or of SPSS outputs on the screen. Hundreds of variables become instantly accessible to the instructor, and the class can switch to an interactive mode, allowing the instructor to produce instantly charts or tables in response to a question from a student. This presupposes of course a certain familiarity both with SPSS and with the data files used as examples, but the new versions of SPSS make it so user-friendly that this familiarity is quickly acquired. Moreover, an instructor in a specific discipline can supply data files from his or her own research and fit them into the pedagogical process proposed in this manual. This instant accessibility to many examples in the classroom is what allowed me to write a rather concise text, focusing on the logic of the statistical approach and on its abstract structure. It was no longer necessary to include numerous examples, knowing that many can be accessed directly in the data files, in class, and in an interactive manner.

The pedagogical approach used in this book is partly the result of a long experience in teaching two subjects: mathematics to student audiences which are not inclined toward it, and sociology of foreign societies whose inner workings and logic are not readily accessible to North American student audiences. In both cases,

the questions of induction vs. deduction, of relevance, and of intuitive knowledge and understanding vs. formalization, pose themselves in an acute way. I found that this experience, and the lessons learned from it, became quite useful when teaching quantitative methods to students not inclined toward statistics.

Here, then, are the basic principles on which my approach is based.

Principle 1: Knowledge is constructed by a mixture of *inductive* and *deductive* modes of thinking, which have different roles in the learning process.

Inductive modes of thinking are more effective when the subject is new, or when its relevance has not yet been established, or when the level of maturity needed for learning it deductively has not been achieved. Induction corresponds to the stage of discovery, of changes of paradigms, and of understanding of a new subject.

Deductive modes of thinking are more effective at the stage of establishing a logical order between concepts, ideas, and theories, and of sorting out potentially brilliant – but false – intuitions from correct facts, that is, at the stage of establishing proofs. This is also the stage of effective and efficient organization of knowledge.

I assert that for this course, the inductive modes of reasoning should come first, in order to develop an intuitive understanding of the concepts, and that they should be followed by deductive modes of reasoning to establish proofs or to explain statistical procedures.

Some of the difficulties encountered by students learning quantitative methods and statistics are due precisely to the fact that certain ways of reasoning are presented in a deductive fashion first, making their logical underpinnings very solid, but meaning nothing to a majority of the students. A related principle is the question of relevance explained below.

Principle 2: The establishment of the *relevance* of a statement must *precede* the demonstration of its *truth*.

This is another way of saying that the answer to a question becomes meaningful only after the question itself has been understood, assimilated, integrated, and incorporated into one's view of the subject.

To show the relevance of a question means to show:

- its meaning and significance;
- its importance in comparison to other questions;
- the consequences that different answers to the question might have on the problem we are dealing with.

In many books of statistics, the material covered is organized according to a deductive logic, which draws its internal consistency from the fact that it is the logic of proof, even when proofs of theorems are not given. For the purpose of training future statisticians this is absolutely crucial, of course. But for training users of statistics in the social sciences (and perhaps also in some other domains),

I claim that another order of presentation is required. Such a logical order would involve first an elaboration of the *meaning* of what is to be demonstrated, then of its *relevance*, then of the *practical ways of applying it*, and only last, if time permits, the *proof* of the truth of a statement or of the exactness of a method. At that point, the importance of the proof must be asserted, together with the fact that *the only thing that justifies presenting the method in this way is that a proof of its exactness exists somewhere else*, and that statisticians have actually proven that it works. This does not imply, though, that the proof itself must be presented.

To illustrate my point, let me take the example of confidence statements. We can help students understand, as a first step, the very notion of a confidence statement in a way that does not involve any exact calculation of margin of error or of probability of error. The calculations are introduced only after the concept itself has been understood.

For instance, it can be explained in class that if a representative sample (a notion to be used intuitively at this point) of students in a college is chosen, and if their height was measured and its average found to be 170 cm, we could guess that the average height for the whole population is expected to be *around* 170 cm, may be somewhere between, say, 165 cm and 175 cm unless, by an extraordinary stroke of bad luck, the students we picked at random consisted of the entire basketball team of the college. We can then explain the concept of margin of error without being able to compute such a margin yet, and we can explain that we need to find a way to determine whether the margin of error should be from 165 to 175 cm, or between any other values. Methods for calculating the margin of error should be introduced only after the very concept is understood, and then they can refer to the notion of sampling distribution, which is the basis for calculating the margin of error associated with some probability level. Then we could mention that this method of calculation can be proven to be correct, by mathematical means. A mathematician may object by saying that the very notion of the margin of error can be understood correctly only when you show the proof of why it works (which gives you at the same time the method for computing it). I claim that this is true when a certain level of maturity has been achieved. I claim that the method proposed here contributes to helping the students achieve such a level of maturity. A subsequent, more advanced course in statistics for social science students could certainly include more sophisticated mathematical ideas.

Principle 3: Intuition could be developed through such a course, and not supposed to be already developed.

This goal can be achieved by two methods: the frequent use of *diagrams* and the constant reformulation of conclusions in plain English, that is, the translation of numerical results into full English sentences and vice versa.

A systematic effort has been made to include diagrams that establish relationships between concepts. Thus, students do not learn individual concepts, but networks of concepts that structure the field of quantitative methods. Establishing such networks of concepts has the double function of making them more relevant

and of playing the role of a mnemonic device. Hence, several diagrams have been introduced in the book.

The translation of numerical results into full sentences in plain English is an essential aspect of the exercises students must perform to really understand the material. For instance, exercises that involve *reading* a two-way table are essential. Students should be able to determine whether a sentence like 'the men of this sample are more likely than the women to behave in a manner *X*' follows from a two-way table or not.

Principle 4: The order of succession of the steps of a research project is not necessarily the same as the order in which students can learn them.

For instance, creating a data file comes logically before the analysis of data files, but I assert that many of the methodological issues that need to be addressed when creating a data file can be understood only *after* you had an opportunity to work with a data file and perform some of the statistical analyses. This is particularly true of some of the issues that relate to the codebook and to the operational definition of a concept. This is why I suggest that the tutorial on the creation of a data file in SPSS be shown only after the section on descriptive statistics. I included it in an appendix, so as to give the instructor some leeway on the moment when the students do it.

As a result, the method used here is based on the following succession: *Intuitive* notions of statistics are presented first, and they are stretched to their maximum logical limits. Only when the notions are well understood intuitively can one see the logical limits of such notions. *Formalization and rigorous definitions* are then introduced as a response to the limits of intuitive notions. They come at the end of the discussion of a subject, not at its beginning.

Principle 5: Intellectual maturity should not be considered a prerequisite for this course (although when present, it would allow a much faster progression). It should be consciously built and developed in the mind of students through the learning activities of this course.

Intellectual maturity is the capacity to understand the *relevance* of an argument, and to find *connections* between different arguments and different ideas. It is the capacity to situate the argument as a whole in a larger conceptual entity that confers relevance to it and that gives it its significance. This significance allows us to understand the logical necessity of the various parts of the argument and its logical coherence. This kind of contextual information is usually not included in the argument itself. This means that maturity is achieved when things are understood without having to be said in full detail. *Maturity is also the capacity of filling-in details that may be missing when they are not logically necessary from a deductive point of view but they are necessary form the point of view of the relevance of the argument*. Thus a mathematical proof may be presented in a comprehensive way, and be perfectly understandable to a mature mathematician, but may appear at the same time totally incomprehensible to a beginner.

Maturity is acquired by experience and is founded on analogies between a situation at hand and other situations seen previously and understood, that may present some degree of similarity. The lessons drawn from such analogies are often not explicit, but they require some kind of integration of previous knowledge: an *understanding* of their contextual meaning, then a *retention* of such a meaning followed by a *transfer* to a different context. Thus, at its lowest levels, maturity is specific to a field of knowledge, but it may become transferable to other fields of knowledge.

The implication of the above remarks for this book is that maturity has to be consciously built and developed in the mind of students by a systematic discussion of all the elements that confer relevance to the question studied, and to the logical elements that contribute to establishing the answers. Developing maturity in quantitative methods may be facilitated by the weekly access to a computer lab. Numerous statistical examples can be seen in a very short period of time, and, if properly directed, students can see at the click of a mouse the connections between some features of a distribution and its statistical properties.

The principles outlined above have immediate consequences for the way the material is organized. In the first chapter, for instance, I start describing an SPSS data file, as it first appears when the program is opened. The students learn how to read it and interpret the information in it, with an approach that may appear to be positivist at first. But immediately after, questions are raised about how the data file was produced in the first place. The relevance of many questions that will gradually be answered as the course unfolds is then shown. I end up questioning positivist approaches by showing that there are many ways of defining a concept and of presenting it, depending on how the concept is conceptualized in the theoretical framework that is used.

Finally, I hope that these comments constitute a contribution to a dialog about pedagogy among instructors of quantitative methods, a dialog that the author is eager to engage in. Any comments are welcome, and I will make it a point to respond to any discussion initiated on these issues.

Rachad Antonius
Professor
Department of Sociology
Université du Québec à Montréal
January 2012
antonius.rachad@uqam.ca

FOREWORD TO THE STUDENT

Objectives of the book

This book is addressed to students taking a first course in quantitative methods for the social sciences, and to researchers who are writing a dissertation or a research report, and who have to use quantitative data at an elementary or intermediate level. Researchers who have to use quantitative data in a research that is not essentially quantitative often feel they are not adequately prepared for this task. This book can serve as a useful and accessible reference source.

The book deals with the production, presentation, analysis, and interpretation of quantitative data, all presented and conceptualized as part of a social research process. It makes use of elementary statistical techniques that are explained and put into use with the help of the Statistical Package for Social Sciences software, known today as IBM SPSS. No prior knowledge of IBM SPSS or of statistics is assumed, and you will be guided through the use of this statistical package one chapter at a time. The tutorial found at the end of every chapter (except Chapters 6 and 7, which do not need one) will guide you through the elementary functions of the software, which correspond to the material shown in each chapter. You will find in the book clear and concise explanations of the procedures used to compute statistical results. But more importantly, you will find explanations of how such statistical results are to be read and interpreted.

Indeed, quantitative methods are a component of the research process as a whole. Statistical results become more meaningful when they are linked to a set of explicit research questions, which are themselves anchored into an adequate conceptual and theoretical framework. But to get there, an understanding of the methods and procedures that produce the statistical results is absolutely necessary. This means that a correct interpretation of statistical results relies on two elements: an understanding of the statistical techniques and procedures as such, and the explicit links made between the statistical results and the conceptual and theoretical framework on which they are based.

After studying this book, you should have acquired the practical abilities needed to produce data files, to organize them, to carry out statistical computations with them, to present their results, and to interpret them correctly. In addition to these abilities, you are also expected to acquire some of the theoretical knowledge that will allow you to use quantitative methods in an appropriate manner, and to understand their power and their limits.

More specifically, the book explains:

- what a quantitative data file is and how to organize data into electronic files;
- how to collect data for social research;
- how to read a data file, i.e. how to interpret its immediate meanings;
- how to analyze it, using both descriptive and inferential statistical methods;
- how to interpret the results of the analysis and how to translate them into plain language;
- how to report a quantitative analysis.

Some of these questions have straightforward answers. Others will require a detailed examination of how quantitative methods fit in a research process, which generally involves aspects that are not quantitative. You should keep in mind that it is the *qualitative* aspects of a piece of research that are the crucial element guaranteeing the rigor of the research and the reliability of the results. Indeed, it is very easy, with the availability of powerful software, to carry out complex computations that will produce numerical results. But such results could be meaningless, or even completely erroneous, if they do not rest on solid theoretical foundations, which are usually expressed in qualitative terms. To give an example, consider the notion of intelligence. At the beginning of the twentieth century, some scientists measured with great accuracy the shapes and sizes of people's skulls, and counted the bumps on the skulls of the individuals, noting their size and location. They believed that these measures gave the unmistakable signs of intelligence. But they were wrong. And no statistical treatment of their data, however complex and sophisticated, could compensate for the theoretical flaws that were at the basis of such data.

A number of abilities to be acquired are listed at the beginning of every chapter. A brief version of this list is given below. You could use it as a checklist, to keep track of the main objectives you feel you have achieved. These abilities are grouped into five broad categories: understanding the basics of social research and the place of quantitative methods in it, producing an electronic data file, analyzing a data file using SPSS and producing a descriptive report. Here are the details.

Brief checklist of abilities to be acquired

(See the beginning of every chapter for a more extensive checklist).

I. Understanding the basics of social research

You should be able to explain the steps of social research. You should know:

- the broad steps needed to complete any research (quantitative or qualitative);
- the role and importance of theory in orienting empirical research;

- the main research designs used to produce quantitative data;
- how to select a random sample or a systematic sample;
- the basic steps to be taken in conducting a survey;
- the structure of the basic experimental designs in social research;
- the ethical guidelines that should be followed when doing research on human participants.

II. Understanding quantitative methods and their place in social research

You should know:

- the basic vocabulary of statistics and quantitative methods;
- the type of variables and of measurement scales;
- how concepts are operationalized with the help of indicators;
- the main types of quantitative research (surveys, experiments, and archival research);
- the basic definitions of descriptive and inferential statistics;
- the different uses of the term 'statistics';
- what an electronic data file looks like, and how to identify cases and variables.

III. Producing an electronic data file

You should be able to:

- determine how to organize the variables and how to determine their types;
- specify their characteristics and define them in IBM SPSS;
- enter the data;
- save and print a data file and an output file.

IV. Analyzing a data file using IBM SPSS

You should be able to:

- present data (frequency tables; charts), describe the shape of a distribution (symmetry, skewness) and produce these outputs in IBM SPSS;
- determine which measures and charts are appropriate, depending on the measurement level of the variable;
- produce and interpret the various descriptive measures;

- explain the differences between the mean and the median and their use;
- produce and read a frequency table;
- use the table of areas under a normal curve;
- use that same table for sampling distributions, to determine the proportion of samples having a mean or a proportion that falls in a given range;
- produce and interpret confidence statements (such as the results of a poll);
- reproduce the logical reasoning underlying hypothesis testing;
- perform and interpret simple *t*-tests;
- analyze statistical association using the appropriate procedure, depending on the measurement level of the variables;
- produce and read a two-way table (manually and with SPSS);
- produce and interpret a coefficient of correlation and a scatter plot;
- interpret the comparisons of the mean of a quantitative variable across the various groups of a categorical variable;
- interpret the significance levels for the various tests of association.

V. Producing a descriptive report

You should be able to:

- report a quantitative analysis, highlighting the most important features of the data and providing evidence for the proposed interpretations;
- copy tables and charts into a word processor;
- explain, in plain English, the meanings of the numerical results produced with IBM SPSS;
- write and present the report of a quantitative analysis describing the data file, describing the variables and the scales used, presenting and describing the data, presenting the results of elementary procedures, and their interpretations (distributions, statistical association, inference, etc.).

Study tips

Mastering this material requires several learning activities:

- reading the material carefully and thoroughly;
- doing the exercises;
- doing the SPSS tutorials;
- reviewing the material and integrating the various components of the acquired knowledge.

Here are some suggestions for each one of these four learning activities.

Reading the material

The material in this book is written in a rather concise form. It is very important to read it more than once, and with attention. The first reading allows you to understand the scope of a chapter and its principal aim. Some of the fine points may be missed at that first reading. A second reading allows you to consolidate what you have learned in the first reading and to capture some of the details missed at the first reading. It is a good idea to read with a pencil and paper at hand, and to write down important definitions, or formulas, or some idea that seems to hold the key to understanding subsequent ideas, so as to remember them more easily. Finally, a third reading is recommended toward the end of the course, after you have covered most chapters and gone through the SPSS lab sessions. This will allow you to link the different parts of the book, and to understand its material at a different, deeper level. You will see then that you read into the text elements that you had not seen in the first readings. The glossary at the end of the book allows you to just focus on the definitions, while the index will allow you to find the page where a concept is explained and to revisit it.

Doing the exercises

The exercises at the end of each chapter are important to help you really understand the material explained in the chapter. Some exercises are reading exercises. They force you to look for some specific details in the text and either to mention them as they are, or to reformulate them in your own words. Some are computational exercises. In order to solve these problems, you must have understood the procedure used to perform a certain computation, and you must be able to reproduce it in a specific situation. While doing the exercises, you may have to go to the main text to find a definition or an explanation of the procedure.

Doing the SPSS tutorials

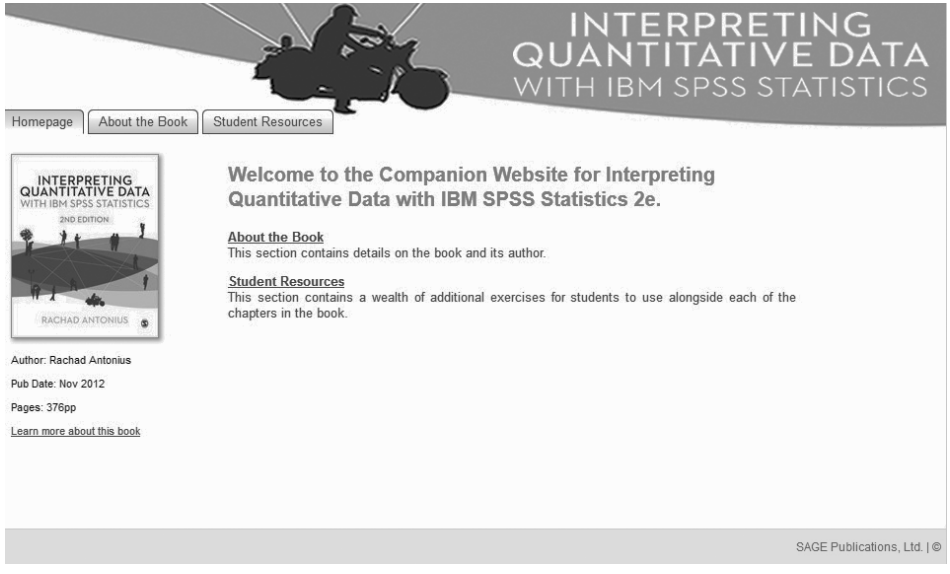
The SPSS tutorials are an integral part of learning the material in this book. The data files you will be working with contain a wealth of concrete examples that illustrate further the theoretical material explained in the book. The tutorials are all structured in a similar way: some procedures are explained in detail and their results shown. You should perform these procedures yourself on your computer as you follow the explanations, making sure that what you get on your screen corresponds to what is explained in the book. This should give you an understanding of how the procedure works. You are then asked to answer a question that requires using the same procedure for a different variable, or perhaps a different data file altogether. After going through all the tutorials, you should be able to perform an elementary statistical analysis of a data file on your own.

Reviewing the material

This is a very important step. Learning does not progress linearly, but rather in a spiral movement. After seeing a concept, you need to see other concepts to which it is applied, or that follow it. You then come back to the first concept and you understand it at a deeper level. The review should allow you to have a broad perspective on the material, and to link ideas and procedures learned in different parts of the book. More importantly, it will allow you to select the appropriate procedure needed in a given context.

Four tools are provided to review the material. The first is given by the set of keywords at the end of every chapter. You can review a chapter by trying to give the definition of each keyword mentioned at its end. The tutorials are a second way of reviewing the material, as you are required to use the concepts introduced in the chapter. A *synthesis of how to analyze a data file* is provided on the website of the book (http://uk.sagepub.com/antonius2). It constitutes a transversal review of the book. This synthesis will help you integrate the material learned, that is, to recapitulate it in a context that is different from the one where it was learned, and to combine several techniques learned in various chapters in a single portion of the analysis. Finally, a set of review questions are also given on the website of the book. They constitute the fourth tool for reviewing the material.

It is hoped that the approach used here will show the relevance of statistical techniques to making meaningful comments about both society and individuals. The book is intended as an introductory first course in quantitative methods. Some students in social sciences will be expected to pursue the learning of quantitative methods through subsequent, more advanced courses. Others will use it as reference material, for writing their dissertation or their research report, without going further in the study of quantitative methods. The book is intended for them as well. I hope that it will motivate both groups in using quantitative methods and that it will make their task both more efficient and more pleasant.



ACKNOWLEDGMENTS

My study of quantitative methods was started at the invitation of my new colleagues in sociology, when I was in the process of switching from a teaching career in mathematics to a research orientation in sociology. They felt that a new program for teaching quantitative methods for the social sciences at the college level would benefit from the input of people who were familiar with both mathematics and sociology. This led me to collaborate with the late Robert Trudel, a statistician, and we co-authored a textbook titled *Méthodes quantitatives appliquées aux sciences humaines*. In the process of writing that book, we had long and elaborated discussions on every aspect of it, and on the formulation of statistical concepts and procedures in a way that would make them relevant to social science students. These discussions helped shape my views on how that subject matter ought to be taught, and made me understand, among other things, that the relevance of a concept had to be established before a formal definition could be explained.

I developed these ideas further when I was asked to teach quantitative methods for the social sciences to a college audience at Champlain Regional College – St-Lambert. My colleagues used the course notes I had developed, and their comments helped me refine my approach. I am indebted to them as well and wish to thank them here, as those notes became the first edition of this book.

I then had to teach quantitative methods courses to a university audience, both at the undergraduate and graduate levels. This allowed me to see some of the issues in a different perspective and to restructure the introductory course in quantitative methods in such a way as to prepare the ground for more advanced methods. This second edition is the result of this teaching experience, and I wish to thank my colleagues and my students at the Department of Sociology of the Université du Québec à Montréal, for their support and their feedback.

PART ONE

INTRODUCTION TO QUANTITATIVE METHODS

ONE

THE BASIC LANGUAGE OF STATISTICS

This chapter is an introduction to statistics and to quantitative methods. It explains the basic language used in statistics, the notion of a data file, the distinction between descriptive and inferential statistics, and the basic concepts of statistics and quantitative methods.

After studying this chapter, you should know:

- **the basic vocabulary of statistics and of quantitative methods;**
- **what an electronic data file looks like, and how to identify cases and variables;**
- **the different uses of the term 'statistics';**
- **the basic definition of descriptive and inferential statistics;**
- **the type of variables and of measurement scales;**
- **how concepts are operationalized with the help of indicators.**

Introduction: Social sciences and quantitative methods

Social sciences aim to study social phenomena – that is, human collective behavior, the culture that sustains it, the relationships between people that make it possible, and the organizations and institutions that regulate it – as rigorously as possible. This involves describing some aspect of social reality, analyzing it to see whether causal or explanatory links can be established between its various parts, and, whenever possible, predicting future outcomes, or at least a range of possible outcomes.

The general objective of such studies is to understand the patterns of individual or collective behavior, the constraints that affect it, the causes and explanations that can help us understand our societies and ourselves better and predict the consequences of certain situations. Such studies are never entirely objective, as they are inevitably based on certain assumptions and beliefs that cannot be demonstrated. Our perceptions of social phenomena are themselves subjective to a large extent, as they depend on the *meanings* we attribute to what we observe. Thus, we *interpret* social and human phenomena much more than we describe them, but we try to make that interpretation as objective as possible.

Some of the phenomena we observe can be *quantified*, which means that we can translate some aspects of our observations into numbers and make use of their properties. For instance, we can quantify population change: we can count how many babies are born every year in a given country, how many people die, and how many people migrate in or out of the country. Such figures allow us to estimate the present size of the population, and maybe even to predict how this size is going to change in a near future. We can quantify psychological phenomena such as the degree of stress or the rapidity of response to a stimulus; demographic phenomena such as population sizes or sex ratios (the ratio of men to women); geographic phenomena such as the average amount of rain over a year or over a month; economic phenomena such as the rate of employment; we can also quantify social phenomena such as the changing patterns of marriage or of unions, and so on.

When a social or human phenomenon is quantified in an appropriate way, we can ground our analysis of it on figures, or statistics. This allows us to describe the phenomenon with some accuracy, to establish whether there are links between some of the variables, and even to predict the evolution of the phenomenon. If the observations have been conducted on a sample (i.e. a group of people smaller than the whole population), we may even be able to generalize to the whole population what we have found on that sample.

When we observe a social or human phenomenon in a systematic, scientific way, the information we gather about it is referred to as *data*. In other words, **data** is information that is collected in a systematic way, and organized and recorded in such a way that it can be interpreted correctly. Data is not collected haphazardly, but in response to some questions that the researchers would like to answer. Sometimes, we collect information (i.e. data) about a character or a quality that has no numerical value, such as the mother tongue of a person. Sometimes, the data is measurable with numbers, such as a person's age. In both cases, we can treat this data numerically: for instance, we can count how many people speak a certain language, or we can find the average age of a group of people. The procedures and techniques used to analyze data numerically are called *quantitative methods*. In other words, **quantitative methods** are procedures and techniques used to analyze data numerically; they include a study of the valid methods used for collecting data in the first place, as well as a discussion of the *limits of validity* of any given procedure (i.e. an understanding of the situations when a given procedure yields valid results), and of the ways the results are to be interpreted.

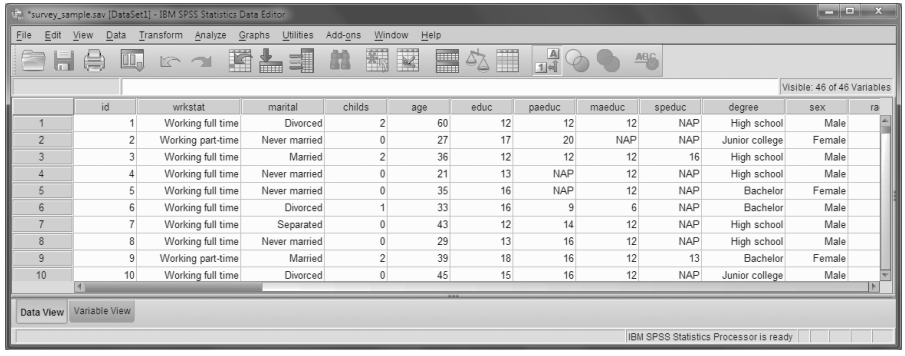
This book constitutes an introduction to quantitative methods for the social sciences. The present chapter covers the basic vocabulary of quantitative methods. This vocabulary should be mastered by the student if the remainder of the book is to be understood properly.

Data files

One of the first objects we deal with, in quantitative methods, is a **data file**. This is an electronic file that contains all the data, organized in a systematic way, often

using numeric codes to refer to the various observations. When conducting research, we must distinguish between **primary data**, that is, data that is produced by the researcher or by a research agency, and **secondary data**, that is, data which is cited in an academic publication but which has been produced by some other researcher, or some agency, or which has been manipulated and summarized. The term **raw data** designates data that has not been subjected to any kind of statistical treatment, such as grouping, recoding, or selecting.

Figure 1.1 illustrates what an electronic data file looks like when we open it with the SPSS program. We can see in this figure the data window and the menu bar that appears on the top of your screen when you open SPSS.



	id	wrkstat	marital	childs	age	educ	paeduc	maeduc	speduc	degree	sex	ra
1	1	Working full time	Divorced	2	60	12	12	12	NAP	High school	Male	
2	2	Working part-time	Never married	0	27	17	20	NAP	NAP	Junior college	Female	
3	3	Working full time	Married	2	36	12	12	12	16	High school	Male	
4	4	Working full time	Never married	0	21	13	NAP	12	NAP	High school	Male	
5	5	Working full time	Never married	0	35	16	NAP	12	NAP	Bachelor	Female	
6	6	Working full time	Divorced	1	33	16	9	6	NAP	Bachelor	Male	
7	7	Working full time	Separated	0	43	12	14	12	NAP	High school	Male	
8	8	Working full time	Never married	0	29	13	16	12	NAP	High school	Male	
9	9	Working part-time	Married	2	39	18	16	12	13	Bachelor	Female	
10	10	Working full time	Divorced	0	45	15	16	12	NAP	Junior college	Male	

Figure 1.1 The Data View window in IBM SPSS 19 and the Menu bar (PC version)

This data file was created by version 19 of the statistical software SPSS (Statistical Package for the Social Sciences), now called IBM SPSS. This software is available for both Windows and Mac. Exercise 1 at the end of this chapter will introduce you to SPSS, but we can look at some of its features right away.

On the top of the window, you can read the name of the data file **survey_sample.sav**.

When we open an SPSS data file, two views can be displayed: the Data View, or the Variable View. Both views are part of the same file, and one can switch from one view to the other by clicking on the tab at the bottom center of the window.

The **Data View** displays the data itself, and the information is organized in rows and columns. Each row refers to a **case**, that is, all the information pertaining to one individual. Each column refers to a **variable**, i.e. a character or quality that was observed and recorded using codes to refer to the various values of the variable. For instance, the second column is a variable called *wrkstat*, and the third is a variable called *marital*. Looking at the contents of each of these two columns, we see that the first variable concerns the work status, and the second the marital status of each individual.

It is important to understand that the data is not stored the way it appears in Figure 1.1. Rather it is stored in codes that take less memory space in a computer, making computations much faster. Thus, instead of recording 'Married', the

program will just store a simple code, such as '1'. There is a way of showing the codes instead of the value labels. This is done by ticking off the command **Value Labels** under the **View** menu, in the data file. The Data View window now looks as shown in Figure 1.2.

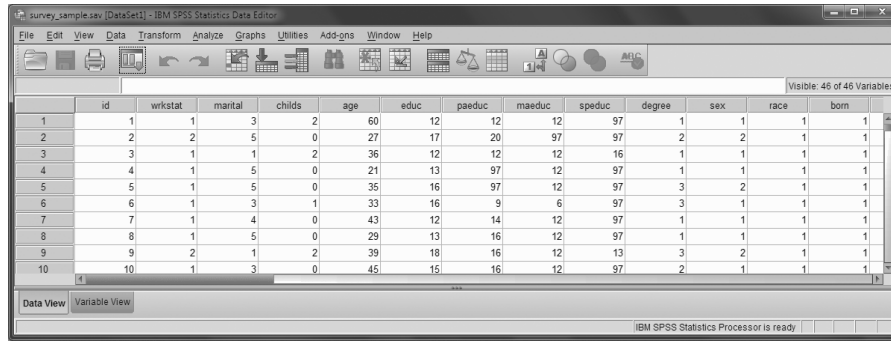


Figure 1.2 The Data View window when the Value Labels command is ticked off the View menu

Here, the codes are displayed rather than the value labels. The meanings of these codes can be seen by clicking on **Variables…** under the **Utilities** menu. The resulting window is shown in Figure 1.3.

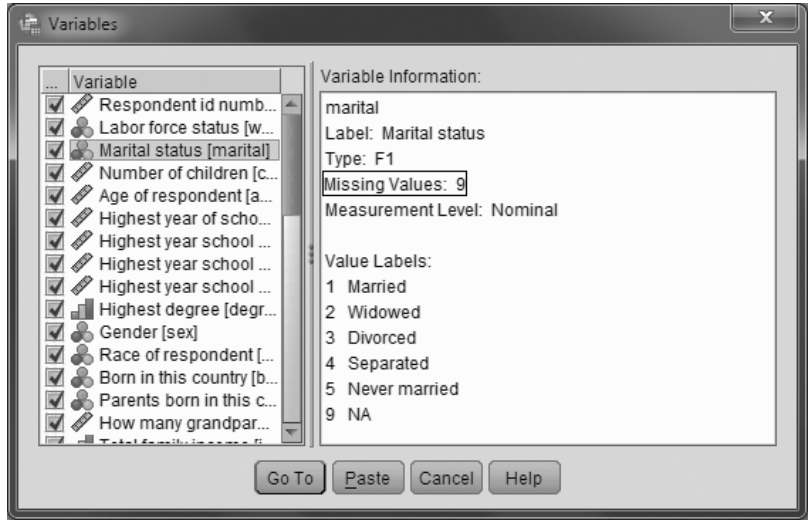


Figure 1.3 The Variable and Value labels for the variable *marital*

You can see here that the variable is designated as 'marital', that its full label is Marital status, and the various codes are also given:

1 stands for Married

2 stands for Widowed

3 stands for Divorced

etc.

This information is part of the **codebook**, which includes a complete listing of all the variables, their labels, their values (the codes) and their value labels. In SPSS, the codebook is referred to as the **data file information**. You can see it by selecting **File → Display Data File Information**. You will get a new window, called the Output window, that contains all the information about the variables, and all the codes and their meanings.

The information about the variables can also be seen by clicking on the **Variable View** tab at the bottom center of the data window. You then get the window shown in Figure 1.4.

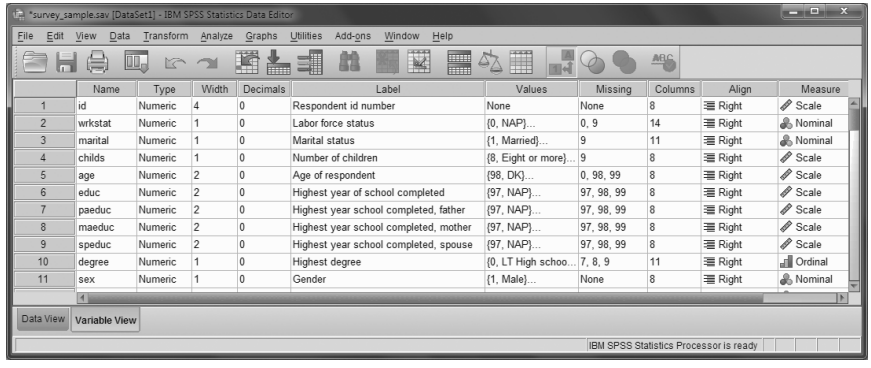


Figure 1.4 The Variable View window

Here every line represents a variable and provides some of its features:

Name This is the brief name of the variable. It must be short, with no spaces. Variable names that are defined by users must begin with a letter. The detailed rules for forming variable names can be found in the SPSS Help menu.

Type This column specifies how the variable is written: either a numeric code, or a string of characters, or a date or a currency, or a special character.

Width This column specifies how much space is devoted to each variable, i.e. the maximum number of characters allowed for writing down the observations relating to the variable. If the variable is Sex and the codes are 1: Male; 2: Female, only one space is needed because only the codes are recorded in the SPSS data file. The codebook tells us what these codes stand for.

Decimals We need to specify whether the numerical values in the data file are decimal numbers or not, and with how many decimals. This is indicated in this column.

Labels These are the long names of the variables. These names will appear in the tables produced in the SPSS output. They must be explicit enough to allow a correct reading of the tables, but preferably succinct. If the data file comes from a survey, the labels will be the questions of the survey, maybe in a shorter form.

Values This column indicates what the codes used in the data file stand for.

Missing Indicates which of the codes must be treated as missing. Missing values are not taken into account when performing computations (such as the mean, or the valid frequencies, as will be seen in subsequent chapters).

Columns The numbers shown in this column refer to the visual aspect of the data file, i.e. its appearance: they indicate the width of the columns in the Data View.

Align This also refers to the appearance of the data file, indicating the alignment of the data that is displayed: left, center or right.

Measure This column indicates whether the codes used for the variable are to be taken as codes that do not indicate size or magnitude (nominal, such as the codes 1 and 2 for the variable Sex), or numbers that indicate a rank (ordinal) or else as numerical values that indicate a magnitude (scale variables, such as age, length, duration, weight, or a score on some numerical scale).

Now that we have seen what a data file looks like and how the data is stored, we can raise a number of questions: How did we come up with this data? What are the rules for obtaining reliable data that can be interpreted easily? How can we analyze this data? Table 1.1 includes a systematic list of such questions. The answers to these questions will be found in the various chapters and sections of this manual.

Table 1.1 Questions arising from the use of quantitative methods

Questions	Chapters
How did we come up with a given data set? What are the questions we are trying to answer? What is the place of quantitative analysis in social research, and how does it link up with the qualitative questions we may want to ask? What is the scientific way of defining concepts and operationalizing them? What are independent and dependent variables?	1. The Basic Language of Statistics
How do we conduct social research in a scientific way? What procedures should we follow to ensure that results are scientific? What are the basic types of research designs? How do we go about collecting the data?	2. The Research Process
Once collected, the data must be organized and described. How do we do that? When we summarize the data, what are the characteristics that we focus on? What kind of information is lost?	3. Univariate Descriptive Statistics
What are the most common types of shapes and distributions we encounter? How do we select the appropriate graphical representation of the data pertaining to a variable?	4. Graphical Representations

Questions	Chapters
Once variables are entered in a data file, can we recode them in a way that is more adequate for our analysis? How do we create new ones? How do we regroup the categories? Can we sort them? etc.	5. Creating New Variables with SPSS
What are the fundamental properties of samples that allow us to build inference procedures? This chapter is a technical one, and it is crucial for understanding the logic of inference.	6. Normal Distributions and Sampling Distributions
What are the procedures for selecting a sample? Are some of them better than others? How do we ensure that our sample is representative and that we can draw general conclusions from it?	7. Sampling Designs
When the data comes from a sample, under what conditions can we generalize our conclusions to the whole population? How can this be done? Is it precise? What are the risks that our conclusions are wrong?	8. Estimation 9. Hypothesis Testing
Sometimes we notice coincidences in the data: for instance, those who have a higher income tend to behave differently on some social variables than those who do not. Is there a way of describing such relationships between variables, and drawing their significance?	10. Correlation and the Regression Line 11. Two-way Tables and the Chi-squared Test 12. *t*-tests and ANOVA
Finally, when we have finished our statistical analyses, how do we report them? What constitutes good practice?	Appendix: Reporting a Quantitative Analysis

The discipline of statistics

The term *statistics* is used in two different senses: it can refer to the *discipline* of statistics, or it can refer to the *actual data* that has been collected.

As a scientific discipline, the object of **statistics** is the numerical treatment of data pertaining to a large quantity of individuals or a large quantity of objects. It includes a general, theoretical aspect which is based on the mathematical study of probability, but it can also include the study of the concrete problems that are raised when we apply the theoretical methods to specific disciplines. The term **quantitative methods** is used to refer to methods and techniques of statistics which are applied to concrete problems. Thus, the difference between statistics and quantitative methods is that the latter include practical concerns such as finding solutions to the problems arising from the collection of real data, and interpreting the numerical results as they relate to concrete situations. For instance, proving that the mean (or average) of a set of values has certain mathematical properties is part of statistics. Deciding that the mean, rather than, say, the median, is an appropriate measure to use in a given situation is part of quantitative methods. But the line between statistics and quantitative methods is fuzzy, and the two terms are sometimes used interchangeably. In practice, the term 'statistics' is often used to mean quantitative methods, and we will use it in that way too.

The term **statistics** also has a different meaning, and it is used to refer to the actual data that has been obtained by statistical methods. Thus, we will say, for instance, that the latest statistics published by the Ministry of Labor indicate a decrease in unemployment. In that last sentence, the word *statistics* was used to refer to data published by the Ministry.