

Wim Janssens
Katrien Wijnen
Patrick De Pelsmacker
Patrick Van Kenhove

Marketing Research with SPSS

 Pearson

MARKETING RESEARCH WITH SPSS



We work with leading authors to develop the strongest educational materials in marketing, bringing cutting-edge thinking and best learning practice to a global market.

Under a range of well-known imprints, including FT Prentice Hall, we craft high quality print and electronic publications which help readers to understand and apply their content, whether studying or at work.

To find out more about the complete range of our publishing, please visit us on the World Wide Web at: www.pearsoned.co.uk

MARKETING RESEARCH WITH SPSS

Wim Janssens

Katrien Wijnen

Patrick De Pelsmacker

Patrick Van Kenhove



Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

Pearson Education Limited

Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsoned.co.uk

First published 2008

© Pearson Education Limited 2008

The rights of Wim Janssens, Katrien Wijnen, Patrick De Pelsmacker and Patrick Van Kenhove to be identified as authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

ISBN: 978-0-273-70383-9

British Library Cataloguing-in-Publication Data

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

Library of Congress Cataloging-in-Publication Data

Marketing research with SPSS / Wim Janssens . . . [et al.].
p. cm.

Includes bibliographical references and index.

ISBN 978-0-273-70383-9 (pbk. : alk. paper) 1. Marketing research. 2. SPSS for Windows. I. Janssens, Wim.

HF5415.2.M35842 2008

658.8'3028555—dc22

2007045264

10 9 8 7 6 5 4 3 2 1

11 10 09 08 07

Typeset in 10/12.5pt GraphicSabon Roman by 73

Printed and bound in Great Britain by Ashford Colour Press, Gosport, Hants

The publisher's policy is to use paper manufactured from sustainable forests.

Contents

Preface

0 Statistical analyses for marketing research: when and how to use them

Descriptive statistics

Univariate statistics

Multivariate statistics

1 Working with SPSS

Chapter objectives

General

Data input

Typing data directly into SPSS

Inputting data from other application programs

Data editing

Creating labels

Working with missing values

Creating/calculating a new variable

Research on a subset of observations

Recoding variables

Further reading

2 Descriptive statistics

Chapter objectives

Introduction

Frequency tables and graphs

Multiple response tables

Mean and dispersion

Further reading

3 Univariate tests

Chapter objectives

General

One sample

Nominal variables: Binomial test (z-test for proportion)

Nominal variables: χ^2 test

Ordinal variables: Kolmogorov-Smirnov test

Interval scaled variables: Z-test or t-test for the mean

Two dependent samples

Nominal variables: McNemar test

Ordinal variables: Wilcoxon test

Interval scaled variables: t-test for paired observations

ix Two independent samples

Nominal variables: χ^2 test of independence
(cross-table analysis)

Ordinal variables: Mann-Whitney U test

Interval scaled variables: t-test for independent
samples

1 K independent samples

Nominal variables: χ^2 test of independence

Ordinal variables: Kruskal-Wallis test

Interval scaled variables: Analysis of variance

7 K dependent samples

Nominal variables: Cochran Q

Ordinal variables: Friedman test

Interval scaled variables: Repeated measures
analysis of variance

11 Further reading

4 Analysis of variance

Chapter objectives

Technique

Example 1: Analysis of variance as a test of difference or one-way ANOVA

Managerial problem

Problem

Solution

SPSS commands

Interpretation of the SPSS output

Example 2: Analysis of variance with a covariate (ANCOVA)

Technique: supplement

Managerial problem

Problem

Solution

SPSS commands

Interpretation of the SPSS output

Example 3: Analysis of variance for a complete $2 \times 2 \times 2$ factorial design

Managerial problem

Problem

Solution

SPSS commands

Interpretation of the SPSS output

Example 4: Multivariate analysis of variance (MANOVA)	108	6 Logistic regression analysis	184
Technique: supplement	108	Chapter objectives	184
Managerial problem	108	Technique	184
Problem	109	Example 1: Interval-scaled and categorical independent variables, without interaction term	187
Solution	110	Managerial problem	187
SPSS commands	110	Problem	187
Interpretation of the SPSS output	113	Solution	188
Example 5: Analysis of variance with repeated measures	120	SPSS commands	188
Managerial problem	120	Interpretation of the SPSS output	192
Problem	122	Example 2: Interval-scaled and categorical independent variables, with interaction term	206
Solution	122	Managerial problem	206
SPSS commands	122	Problem	207
Interpretation of the SPSS output	125	Solution	208
Example 6: Analysis of variance with repeated measures and between-subjects factor	129	SPSS commands	210
Managerial problem	129	Interpretation of the SPSS output	220
Problem	129	Important guidelines	229
Solution	129	One last remark	229
SPSS commands	129	Example 3: The 'stepwise' method, in addition to the 'enter' method, and more than one 'block'	230
Interpretation of the SPSS output	131	Managerial problem	230
Further reading	136	Problem	230
Endnote	136	Solution	230
5 Linear regression analysis	137	SPSS commands	230
Chapter objectives	137	Interpretation of the SPSS output	233
Technique	137	Example 4: Categorical independent variables with more than two categories	237
Example 1: A cross-section analysis	141	Managerial problem	237
Managerial problem	141	Problem	237
Problem	142	Solution	238
Solution	142	SPSS commands	238
SPSS commands	142	Interpretation of the SPSS output	241
Interpretation of the SPSS output	150	Further reading	243
Example 2: The 'Stepwise' method, in addition to the 'Enter' method	174	Endnotes	244
Problem	174	7 Exploratory factor analysis	245
Solution	175	Chapter objectives	245
SPSS commands	175	Technique	245
Interpretation of the SPSS output	175	Example: Exploratory factor analysis	249
Example 3: The presence of a nominal variable in the regression model	179	Managerial problem	249
Problem	179	Problem	250
Solution	179	Solution	251
SPSS commands	179	SPSS commands	251
Interpretation of the SPSS output	181	Interpretation of the SPSS output	255
Further reading	183	Further reading	278
Endnotes	183	Endnote	278

8 Confirmatory factor analysis and path analysis using SEM

Chapter objectives	279
Technique	279
Example 1: Confirmatory factor analysis	281
Managerial problem	281
Problem	282
Solution	282
AMOS commands	282
Interpretation of the AMOS output	294
Example 2: Path analysis	311
Problem	311
Solution	311
AMOS commands	311
Interpretation of the AMOS output	312
Further reading	316

9 Cluster analysis

Chapter objectives	317
Technique	317
Example 1: Cluster analysis with binary attributes – hierarchical clustering	319
Managerial problem	319
Problem	320
Solution	320
SPSS Commands	320
Interpretation of the SPSS output	324
Example 2: Cluster analysis with continuous attributes – hierarchical clustering as input for K-means clustering	342
Managerial problem	342
Problem	342
Solution	343
SPSS commands: Hierarchical clustering	344
Interpretation of the SPSS output: Hierarchical clustering	347
SPSS commands: K-means clustering	353
Interpretation of the SPSS output: K-means clustering	355
Further reading	362
Endnotes	362

10 Multidimensional scaling techniques 363

Chapter objectives	363
Technique	363
The form of the data matrix: the number of ways and the number of modes	363
The technique: the measurement level of the input and output and the representation of the data	366
Data collection method: direct or indirect measurement	368
Example 1: 'Two-way, two-mode'	
MDS – correspondence analysis	370
Technique: supplement	370
Managerial problem	370
Problem	373
Solution	373
SPSS Commands	373
Interpretation of the SPSS output	384
Example 2: 'Three-way, two-mode'	
MDS – 'two-way, one-mode' MDS using replications in PROXSCAL	398
Managerial problem	398
Technique: supplement	400
Problem	401
Solution	402
SPSS commands: data specification	402
SPSS commands: dimensionality of the solution	404
Interpretation of the SPSS output: dimensionality of the solution	407
Further reading	415
Website reference	415
Endnotes	416

11 Conjoint analysis 417

Chapter objectives	417
Technique	417
Example: Conjoint analysis	418
Managerial problem	418
Problem	419
Solution	419
SPSS commands	419
Interpretation of the SPSS output	428
Further reading	433

Index 435



Preface

Statistical procedures are a ‘sore point’ in every day marketing research. Usually there is very little knowledge about how the proper statistical procedures should be used and even less about how they should be interpreted. In many marketing research reports, the necessary statistical reporting is often lacking. Statistics are often left out of the reports so as to avoid scaring off the user. Of course this means that the user is no longer capable of judging whether or not the right procedures have been used and whether or not the procedures have been used properly. This book has been written for different target audiences. First of all, it is suitable for all marketing researchers who would like to use these statistical procedures in practice. It is also useful for those commissioning and using marketing research. It allows the procedures used to be followed, understood and most importantly, interpreted. In addition, this book can prove beneficial for students in an undergraduate or postgraduate educational programme in marketing, sociology, communication sciences and psychology, as a supplement to courses such as marketing research and research methods. Finally, it is useful for anyone who would like to process completed surveys or questionnaires statistically.

This book picks up where the traditional marketing research handbooks leave off. Its primary goal is to encourage the use of statistical procedures in marketing research. On the basis of a concrete marketing research problem, the book teaches you step by step which statistical procedure to use, identifies the options available, and most importantly, teaches you how to interpret the results. In doing so, the book goes far beyond what the minimum standard options available in the software packages have to offer. It opts for the processing of data using the SPSS package. At present, SPSS is one of the most frequently used

statistical packages in the marketing research world. It is also available at most universities and colleges of higher education. Additionally, it uses a simple menu system (programming is not necessary) and is thus very easy to learn how to use. The book is based on version 15 of this software package.

Information is drawn from concrete datasets which may be found on the website (www.pearsoned.co.uk/depelsmacker). The reader simply has to open the dataset in SPSS (not included) and may then – with the book opened to the appropriate page – practice the techniques, step by step. Most of the datasets originate from actual marketing research projects. Each of the datasets was compiled during the course of interviews performed on consumers or students, and were then input into SPSS. The website also contains a number of syntaxes (procedures in program form).

This book is not however a basic manual for SPSS. The topic is marketing research with the aid of SPSS. This means that a basic knowledge of SPSS is assumed. For the inexperienced reader, the first chapter contains a short introduction to SPSS. This book is also not a basic manual for marketing research or statistics. The reader should not expect an elaborate theoretical explanation on marketing research and/or statistical procedures. The reader will find this type of information in the relevant literature which is referred to in each chapter. The technique used is described briefly and explained at the beginning of every chapter under the heading ‘Technique.’ The book’s primary purpose is to demonstrate the practical implementation of statistics in marketing research, which does more than simply display SPSS input screens and SPSS outputs to show how the analysis should proceed, but also provides an indication of the problems which may crop up and error messages which may appear.

The book starts with a brief introduction to the use of SPSS. The most current data processing techniques are then addressed. The book begins with the simpler analyses. First, descriptive statistics are discussed such as creating visual displays and calculating central tendency and measures of dispersion. After that, we discuss hypothesis testing. The Chi-square test and t-tests are the primary focus, in addition to the most current measures of association. Also, multivariate statistical procedures are discussed at length. The more explorative procedures (factor analysis, cluster analysis, multidimensional scaling techniques and conjoint measurement) as well as the confirmative techniques (analysis of variance, linear regression analysis, logistic regression analysis and linear structural models) are also explained. Some of these techniques require that the reader has more than just the standard modules available within SPSS at his or her disposal. The chapter 'Confirmative factor analysis and path analysis with the aid of SEM' for example requires the separate module 'Amos,' and the chapter 'Multidimensional scaling techniques' makes use of the 'Categories' module.

Each chapter may essentially be read independently from the other chapters. The reader does not have to examine everything down to the very last detail. The 'digging deeper' sections indicate that the text following involves an in-depth exploration that the reader may skip if desired. These areas of text may involve commands in

SPSS windows as well as interpretations of SPSS outputs. Grey frames alongside text and figures contain steps which may be immediately relevant within the scope of the technique being discussed, but which may not necessarily be tied to this label under SPSS (see for example the calculation of Cronbach's Alpha values in a chapter on factor analysis). They are labelled as supporting techniques.

The realization of this book would not have been possible without the assistance of and critical commentary from a number of colleagues. A special word of thanks goes to Tammo H.A. Bijmolt, Frank M.T.A. Busing, Ben Decock, Maggie Geuens, Marc Swyngedouw, Willem A. van der Kloot and Yves Van Handenhove for making datasets available and for providing useful tips and advice.

The authors also wish to thank Lien Standaert, Kirsten Timmermans and Ellen Sterckx for their assistance in creating the screenshots. Finally, it would be appropriate to state here that the first two authors mentioned have made an equal contribution toward the creation of this book.

Wim Janssens
Katrien Wijnen
Patrick De Pelsmacker
Patrick Van Kenhove
January 2008

Chapter 0

Statistical analyses for marketing research: when and how to use them

In quantitative marketing research, be it survey or observation based, pieces of information are collected in a sample of relevant respondents. This information is then transformed into variables containing verbal or numerical labels (scores) per respondent. To make sense of this data set, a variety of statistical analytical methods can be used. Statistical analysis normally takes place in a number of steps or stages. The first set of techniques, called **descriptive statistics**, is used to obtain a descriptive overview of the data at hand, and to summarize the data by means of a limited number of statistical indicators. Next, each variable can be studied separately, for instance to compare average scores of a variable for different groups or subsamples of respondents, or to judge the difference between rankings or frequency distributions. These analyses are called **univariate statistics** or **statistical tests**. Finally, in **multivariate statistics**, several variables can be jointly analysed, to assess which variables explain or predict other variables, or how variables are related to one another. Both in univariate and multivariate statistics, not only description is important, but also statistical validation. In other words, results do not only have to be described and to be assessed on what this description means for the marketing problem at hand; it is at least as important to assess how statistically meaningful or significant the results are, in other words how confident the researcher can be that the descriptive conclusions are statistically reliable and valid.

Descriptive statistics

Univariate statistical description usually contains three types of indicators: frequency distributions, central tendency measures and dispersion measures. **Frequency distributions** indicate how scores of individual respondents are distributed over meaningful categories, for instance, how many male and female respondents, or respondents in three pre-defined age groups there are in the sample. **Central tendency measures** summarize the characteristics of a variable in one statistical indicator, for instance the average consumption of coffee per month in kilograms, the average satisfaction score of a sample of customers of a company on a five-point scale (mean), the gender group in which there are the most respondents (mode), or the middle score of a set of scores ranked from low to high (median). **Dispersion measures** provide an indication of the variability in a set of scores on a variable. Respondents can largely agree on certain issues, in which case dispersion will be low, or the scores on a certain variable can substantially vary between them, in which case dispersion will be high. For instance, everyone can consume about the same amount of coffee, or the satisfaction score of a sample of customers can strongly vary, with large numbers of respondents scoring 1 and 2 as well as 4 and 5 on a five-point scale. Descriptive statistics allow summarizing large data sets in a smaller number of meaningful statistical indicators.

Multivariate description can take many forms, depending of the multivariate technique used. They are normally an integral part of the outcome of each analysis, together with the statistical validation measures, that can also be different for each technique.

Univariate statistics

In univariate statistics or statistical tests, a set of observations in one variable is analysed across different groups of respondents, and the statistical meaningfulness of the difference between these groups is assessed, for instance what is the difference in the average consumption of coffee per month in kilograms between men and women, and is this difference statistically meaningful. The choice of the appropriate statistical test is based on three characteristics of the variables in the samples: the measurement level, the number of samples to be compared, and the (in)dependence of these samples. Variables can be measured on a nominal, ordinal or interval/ratio level. Nominal variables are category labels without meaningful order or metric distance characteristics (for instance men and women). Ordinal variables have a meaningful order, but no metric distance characteristics (for instance, preference rank order indications for a given number of brands). In the case of interval/ratio variables, scores have a metrical meaning, for instance the number of kilograms of coffee purchased by a certain person (one person buys one kilogram, the other buys three, and the distance between the two observations is a metrically meaningful 2 kilograms).

Univariate analysis can be carried out on one sample (for instance, is the average satisfaction score of the whole sample of respondents statistically significantly different from the midpoint score 3?), on two samples (for instance, is the average rank order of brand A significantly different between men and women), or on more than two samples (is the average consumption of coffee significantly different between the three age groups in a sample?).

Finally, in the case of two or more samples, these samples can be dependent or independent. In the case of independent samples, the respondents in one subsample are not linked to the respondents in another subsample, for instance men and women, or three age groups that are not in any way related. In dependent samples, the respondents in one subsample are related to those in other subsamples, for instance husbands and wives, sons and daughters, or the same respondents that are measured at different points in time.

Based on these three characteristics, a selection grid for univariate statistical tests can be constructed:

Measurement level	One sample	Two samples		k Samples	
		Independent	Dependent	Independent	Dependent
Nominal	Binomial test (Z-test on proportion) χ^2	χ^2	McNemar	χ^2	Cochran Q
Ordinal	Kolmogorov-Smirnov	Mann-Whitney U	Wilcoxon	Kruskal-Wallis	Friedman
Interval or ratio	t-test Z-test	t-test Z-test	t-test for differences	Analysis of Variance	Repeated measures Analysis of Variance

In each cell, the appropriate statistical test(s) can be found. In Exhibit 1, for each of these cells, a number of examples of marketing research questions are given.

Exhibit 1 Marketing research applications of univariate statistical tests

- Is the percentage of people interested in museums, as measured in a sample of UK citizens, significantly different from the percentage of museum-lovers as measured in an earlier French study?
- Is the average satisfaction score of a sample of customers of a company, measured on a 5-point scale, significantly different from midpoint (3)?
- Is the average number of pairs of shoes bought per family in The Netherlands significantly larger than 6?
- Is the average percentage recall score of radio ads different between men and women in a sample?
- Is there a difference between the preference for different car models between three age groups in France and Germany?
- Is the average consumption of beer per capita per year in Germany significantly different from Belgium?
- Is there a significant difference between the purchase intention (will/will not buy) for a brand of wine in a sample of potential consumers, before and after an advertising campaign for the product?
- Is there a significant difference between the scores on two examinations of a sample of students?
- Is there a difference between the brand attitude scores measured at different points in time (tracking), in a sample of potential customers?
- Is there a difference between sales figures in three samples of shops in which a different sales promotion campaign has been implemented?

Multivariate statistics

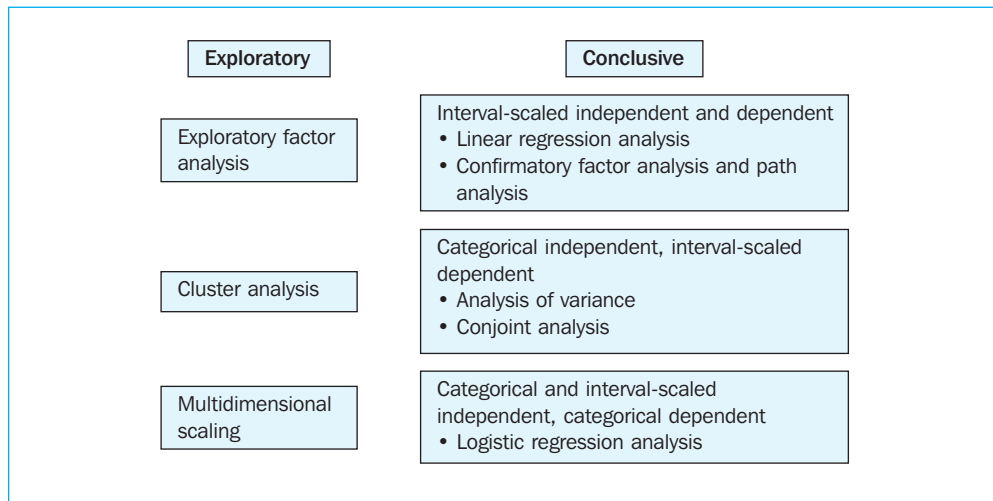
Multivariate analytical methods are research methods in which different variables are analysed at the same time. Each of these techniques requires specific types of data, and has its own fields of application to marketing research. Knowing which type of data a certain analytical technique requires is essential for taking the right decisions about data collection methods and techniques, given certain marketing and marketing research problems at hand.

Which multivariate analytical techniques to use depends on a number of criteria. A first important issue is whether a distinction should be made between independent and dependent variables. **Dependent variables** are factors that the researcher wants to explain or predict by means of one or more **independent variables**, factors of which he/she believes can contribute to the explanation in the variation or evolution of the dependent variables. For instance, a brewery may want to study to what extent price, advertising, distribution and sales promotions (independent variables) explain and predict the evolution of beer consumption over a certain period of time (dependent variable). This type of techniques is called **analysis of dependence**. In case the research problem at hand does not require this distinction to be made, another set of techniques, **analysis of interdependence**, is called for. For instance, a bank may ask itself how many fundamentally different customer segments it can define on the basis of multiple customer characteristics. In this example, no distinction between dependent and independent variables is made; the objective is to

assess the relationship between variables or observations. Interdependence techniques are also called exploratory, while dependence techniques are called confirmatory. Indeed, the purpose of the former is to look for patterns, for structure in variables and observations, while the objective of the latter is to find proof for a pre-defined model that predicts a criterion using predictors. Therefore, interdependence techniques will be mostly used in the exploratory, descriptive stages of a research project, when looking for patterns and structures. Confirmatory techniques will be mainly used in the conclusive stages of a project, in which conclusive answers are sought about which phenomena and factors explain and predict others.

The second important criterion that is important to select a multivariate analytical technique is only relevant for dependence techniques, namely the measurement level of both the dependent and the independent variables. More particularly, the distinction has to be made between nominal or categorical variables on the one hand, and interval/ratio variables on the other. Multivariate analytical techniques that use ordinal data also exist, but they are beyond the scope of this book, and they will not be discussed further. The figure [Multivariate statistical techniques](#) provides an overview of the multivariate techniques discussed in this book.

Multivariate statistical techniques



The objective of exploratory factor analysis is a meaningful reduction of the number of variables in a dataset, based on associations between those variables. In the process, meaningful dimensions in a set of variables are found, and the number of factors to use in further analysis is reduced. In cluster analysis the objective is to reduce the number of observations by assigning them to meaningful clusters on the basis of recurrent patterns in a set of variables. The end result of a cluster analysis is a relatively limited number of clusters or groups of respondents or observations, to be used in further analysis. In multidimensional scaling, perceptions and preferences of consumers are mapped, based on the opinion of consumers about products, brands and their characteristics. Again, the result is a more structured insight in the perception and preference of respondents than based on their detailed preference or perception scores.

In linear regression analysis a mathematical relation is defined that expresses the linear relationship between an interval-scaled dependent variable and a number of independent interval-scaled variables. The objective is to find out to what extent the independent variables can explain or predict the dependent variable, and what the contribution of each independent variable is to explaining variations in the dependent one. The data used to apply this technique can be longitudinal (i.e. measured at different points in time), cross-sectional (measures on different respondents or points of observation at one point in time), or both. Logistic regression analysis is a similar technique, but in this case the dependent variable is categorical, and the independent variables can be both categorical and interval-scaled. The objective of analysis of variance and of conjoint analysis is similar, but the measurement level of the variables is different. In both techniques the relative impact of a number of categorical independent variables on an interval-scaled dependent variable is measured. Finally, in confirmatory factor analysis a predefined measurement model (a number of pre-defined factors), and the relation (path) between a number of independent, mediating and dependent interval-scaled variables are statistically tested. In Exhibit 2, for each of these multivariate methods, a number of examples are given of marketing research problems for which they can be used.

Exhibit 2 Marketing research applications of multivariate statistical methods

1. Exploratory factor analysis

- A car manufacturer measures the reaction of a group of customers to 50 criteria of car quality and tries to find what the basic dimensions of quality are that underlie this measurement
- A bank measures satisfaction scores of a group of customers on 40 satisfaction criteria and explores the basic dimensions of satisfaction judgments
- A supermarket asks its customers how they assess the importance of 20 different shopping motives to try to discover a more limited number of basic shopping motivations

2. Cluster analysis

- A bank tries to identify market segments of similar potential customers on the basis of the similarities in their socio-demographic characteristics (age, level of education . . .) and their preference for certain investments
- A supermarket chain tries to define different segments of customers on the basis of the similarities in the type of goods they buy, the amount they buy, and the brands they prefer
- A radio station defines different type of ads based on the characteristics of the ads, the formats and emotional and informative techniques used (image-orientedness, level of informative content, degree of humour, feelings . . .)

3. Multidimensional scaling

- A car manufacturer wants to find out to what extent potential customers perceive his models and those of competitors similar or dissimilar, and for which models the customer has the greatest preference
- A fashion boutique wants to find out how it is positioned on various image attributes in comparison with its competitors
- A furniture supermarket wants to know which type of customers are attracted to what type of characteristics of his shop

4. Linear regression analysis

- A manufacturer of branded ice cream wants to find out to what extent his price level and advertising efforts have contributed to sales over a period of 36 months
- An insurance company has collected scores on six components of customer satisfaction and wants to assess to what extent each of them contributes to overall satisfaction

5. Confirmatory factor analysis and path analysis

- An Internet shop has identified five factors that contribute to 'shop liking', and on the basis of measurements in a sample of potential customers wants to test to what extent these five factors are compatible with the data he collected, to what extent they determine 'shop liking', and to what extent shop liking, in turn, determines purchase intention
- An advertiser has identified three factors of the attitude of consumers towards advertisements. He wants to find out if these three factors are reflected in the perception of a test sample of customers, and if these factors, together with a brand loyalty measure, determine brand attitudes and buying behaviour

6. Analysis of variance

- A manufacturer of yoghurt has tested three types of promotions and two types of packaging in a number of shops. He wants to find out to what extent each of these variables have influenced sales and what their joined effect is
- A manufacturer of shoes wants to find out if the age of his customers (three categories) and the size of the customers' families (single, married or couple with children) has an impact on annual shoe sales

7. Conjoint analysis

- An airline wants to find out what the impact is of free drinks or not, free newspapers or not, and the availability of mobile phone services on the plane on the customers' preference for a flight
- A jeweller wants to launch a new type of diamond jewel and tries to find out to what extent colour, clarity, cut and carat have an impact on the propensity to spend a certain amount of money for the new jewel

8. Logistic regression analysis

- A telecom provider wants to find out to what extent the age of a person, his education level, and the place he lives in determines whether he is a customer or not
- A hotel wants to know if the country of origin of a traveller, his age, and the number of children he has determines whether he will select his hotel or not for a summer holiday.

Chapter 1

Working with SPSS

Chapter objectives

This chapter will help you to:

- Understand how to construct an SPSS data file
- Create and define variables and labels
- Deal with missing data
- Manipulate data and variables

General

SPSS is a widely distributed software program which allows data to be analysed. This may involve simple descriptive analyses as well as more advanced techniques, such as multivariate analysis. SPSS consists of different modules. This means that in addition to the basic module (Base System), there are also other modules. These are normally destined for more advanced and specialized analyses (for example, the AMOS module is used in Chapter 8, and in Chapter 10, the Categories module is used).

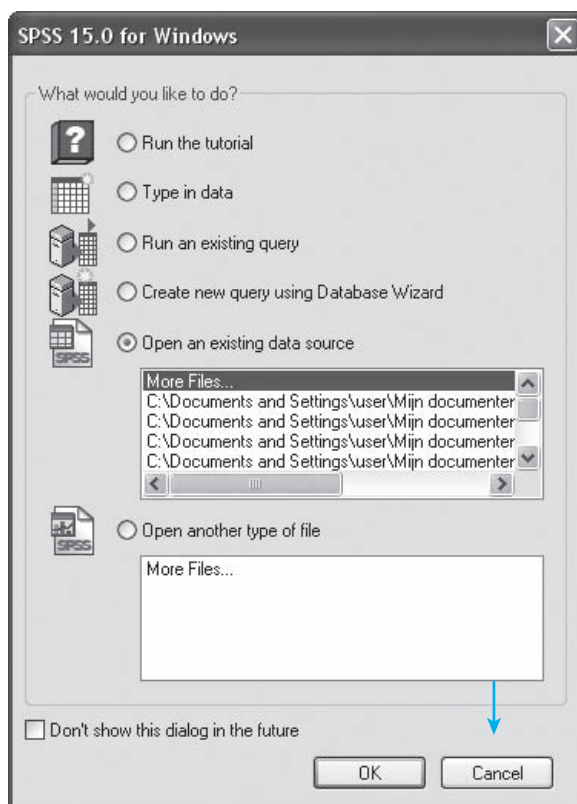
SPSS works with different screens for each type of action (for example data input, output, programming, etc.). This first chapter deals with the Data Editor screen (data input), and several basic topics involving data input and processing will be discussed so that we can quickly begin with the analysis afterwards. Data files are indicated by the extension *.sav*. Starting in Chapter 2, we will also discuss other relevant screens such as the output screen. This is the screen in which all of the results are displayed; this is denoted with the extension *.spo*. For the sake of clarity, it may be said that there are also several other types of screens. For example, there is the 'Chart Editor' which may be used to edit graphs. There is also the syntax screen which will have to be used if the user would like to program the commands instead of clicking on them. This last type of file is indicated with the extension *.sps*. The major advantage of this system is the possibility to move about quickly between input and output.

Additional references may be found at the end of this chapter.

Data input

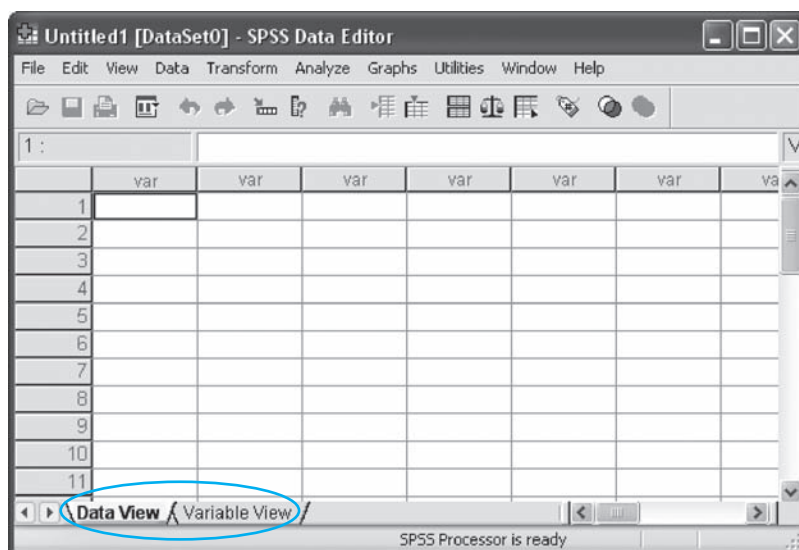
When SPSS starts up, the user will first see a dialogue window (Figure 1.1) which will ask the user what he would like to do.

Figure 1.1



When the user checks 'Type in data' here, and then clicks 'OK', he will enter the data input screen (Data Editor, see Figure 1.2). The same result may be achieved by clicking 'Cancel'.

Figure 1.2



The data input screen in Figure 1.2 consists of two tabs, 'Data View' and 'Variable View'. The user may input the data in the first tab and the characteristics relating to the different variables in the second, such as the name of the variable, the description of the variable, the meaning of each value of the variable, type of variable (numeric, string, etc.), etc.

The user will automatically enter the 'Data View' tab. The tab which is active is indicated with a white tab label (Figure 1.2). To move from one tab to the other, the user just has to click on the tab label.

In order to discuss the different items which are important during the input of data, the following simple example is used here. Suppose the user would like to input the following table into SPSS:

Table 1.1

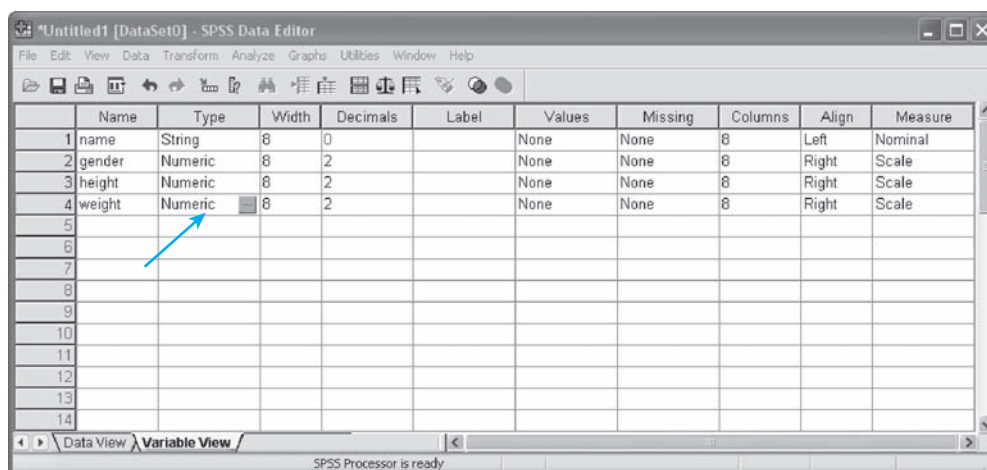
<i>Name</i>	<i>Gender</i>	<i>Height (cm)</i>	<i>Weight (kg)</i>
Joseph	1	180	75
Caitlin	0	165	67
Charles	1	175	80
Catherine	0	170	70
Peter	1	185	75

There are two methods which may be used to input this data into SPSS: they may be either typed in directly or imported from another application program.

Typing data directly into SPSS

A first step is to go to the 'Variable View' tab (Figure 1.3).

Figure 1.3



In the first column (Name), you may type the relevant variable name, and the format in the second column (Type). Click on the relevant cell and then on the '.' field that appears in the relevant cell.

In the example, a string format (= text format) has been chosen for 'name', and a numerical format has been chosen for the other variables (this allows the software

Inputting data from other application programs

If the data are located in application programs other than SPSS (Excel, etc.), these may be imported into SPSS using the path: **File/Open/Data**. The user then has a choice from among a whole series of possible file types which may be clicked on and loaded. In the event that the user encounters problems with this, the following tips may be helpful. Try to save the original dataset in an older version format (e.g. save files in Excel 4.0 format) and then read them into SPSS. The user must also be aware of headings (variable names) which are sometimes not imported or are imported as a missing value. The latter also applies when a simple Copy-Paste command is performed from another application program.

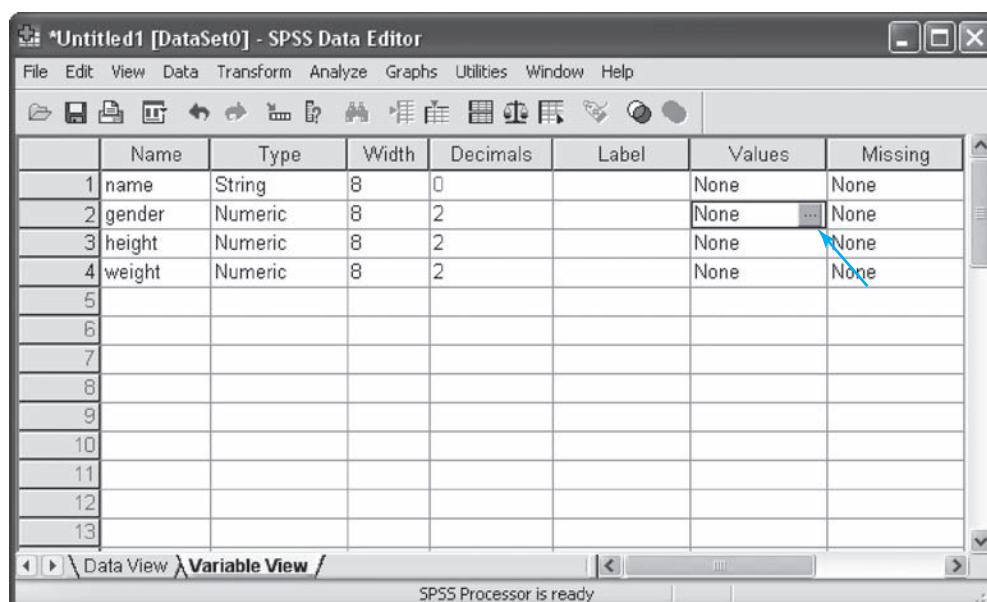
Data editing

In this section, we will discuss several techniques for performing different data editing activities in SPSS.

Creating labels

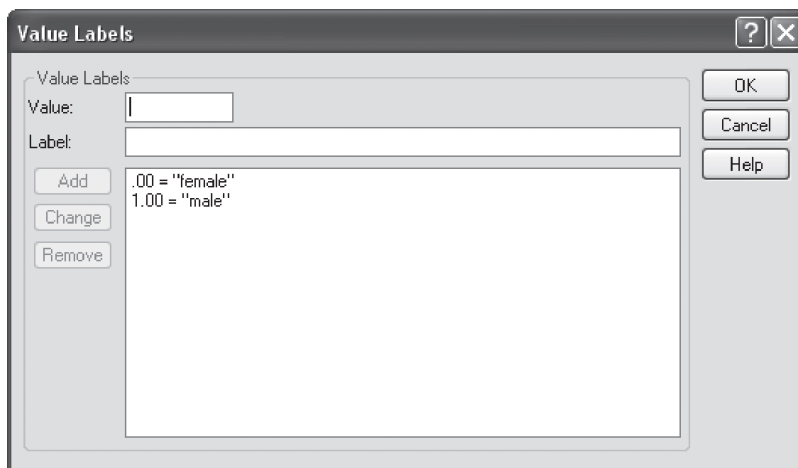
In the example, 'gender' is still defined as a 0/1 variable. Let's say that instead of the '0/1', the researcher would prefer to see the 'female/male' coding appear in the Data View screen. This would also allow the labels 'male' and 'female' to appear in the output, which is easier to interpret than '0' and '1'. This is certainly the case when the researcher is working with many different variables.

Figure 1.5



In the 'Variable View' screen (Figure 1.5), go to the line for the variable to be edited, and then to the 'Values' field. Click on this cell and then on the '.' which appears.

Figure 1.6

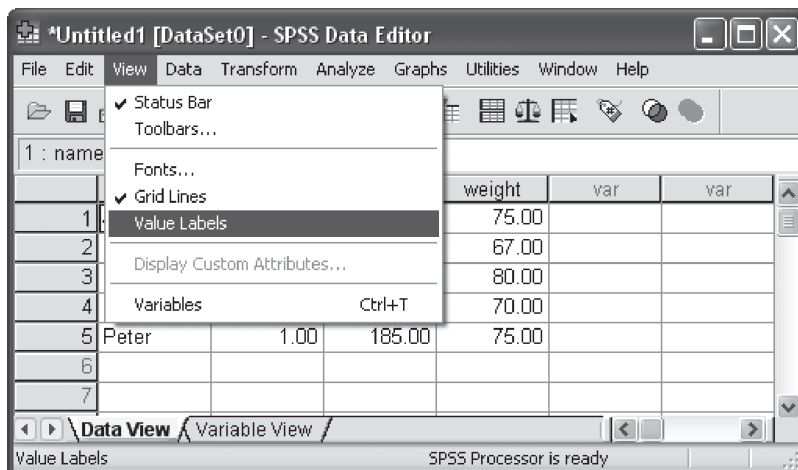


For 'Value' type in '0' and 'female' for 'Value Label' and then click 'Add'. Use the same method for '1' and 'male' (do not forget to click 'Add' each time). This will produce the image that is displayed in Figure 1.6. Now click 'OK'.

If the researcher also prefers to use identical value labels for other variables as for the value labels created for a certain variable, this may be done by simply copying the relevant Values cell in the Variable View window and then pasting this into the Values column for the desired other variables. This is particularly useful in the case of a labeled 7-point scale (1 = totally disagree, 2 = ... up to 7 = totally agree). Instead of entering this for every variable separately, this may be typed in once and then copied and pasted for all of the other variables.

In order to be able to view the changes made to the data set, first go back to the 'Data View' tab, then choose [View/Value Labels](#) from the top (Figure 1.7).

Figure 1.7



This way, you will activate this function and the label values will be displayed in the data set instead of the numerical values (see Figure 1.12 under ‘gender’). In order to turn this function off, you must repeat these steps one more time.

Working with missing values

It occurs regularly that some respondents do not answer all of the questions in a survey. In this case, the researcher would not fill in a value in the ‘Data View’ screen of SPSS and this would remain an empty cell (SPSS will automatically insert a full stop here and this will be processed as ‘System Missing’). If however the user must work with a large amount of data, is unable to fill in the data in one session, or when there are different people who must work with the same data set, it is recommended that a clear indication is provided of whether this involves a value that has not yet been filled in or whether it is a real observation for which no answer was obtained. In this last case, the user can indicate this by using the value ‘99’ for example, or another value that does not occur among the possible answers (this is then called ‘User Missing’). The user must however indicate this explicitly in SPSS; failure to do so will result in SPSS treating the value ‘99’ as a normal input. Imagine that the researcher wishes to calculate an average value (mean) later on of a series of values in which ‘99’ occurs a number of times, then SPSS will see this ‘99’ as a real value and include it in the calculations for the average, instead of just neglecting to include these observations in the analyses.

Let’s say that in the example, the last respondent, ‘Peter’, did not provide an answer to the question about his weight; this may be input in one of two ways. First, the cell may simply be left blank, but then it is not 100% clear whether or not the value must be input later or that the value truly is missing. It is better to opt for the second possibility, which would require that, for example, the value ‘-1’ be filled in in the cell. This way there is then a clear indication that it is a missing value. The user must still indicate in SPSS that the value ‘-1’ used is actually a code for missing values.

Figure 1.8

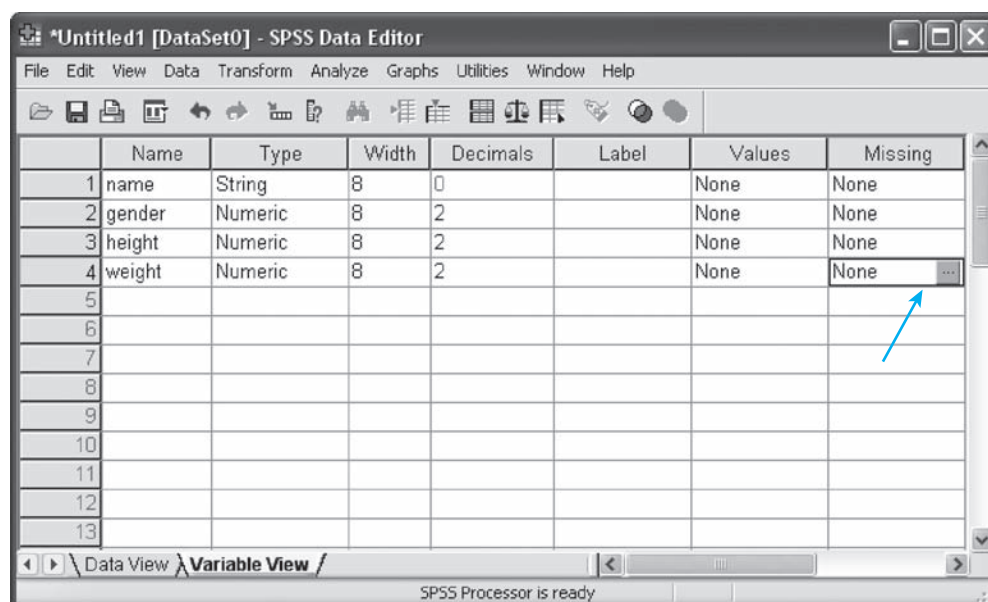
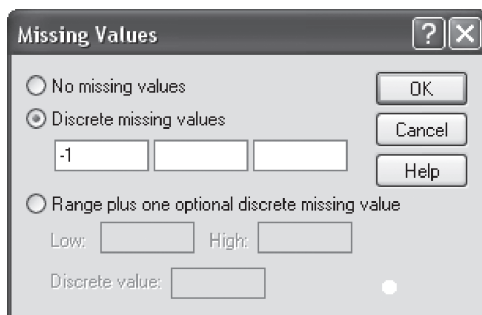


Figure 1.9



Go to the tab 'Variable View' and then choose the cell which is the result of the combination of the 'weight' row and the 'Missing' column. When you click this cell once, a grey box with three dots will appear (see Figure 1.8). Click on this box so that a dialogue window such as that shown in Figure 1.9 will appear.

Click the option 'Discrete missing values' and fill in one of the three boxes with '-1'. As you might notice, it is possible to indicate three different discrete values as a

code, as well as a range of values (plus one discrete value). Now click 'OK' and from now on, SPSS will recognize the value '-1' as a 'missing' value for 'weight.' This setting may be copied to the other variables if desired using a simple Copy-Paste command (in the Variable View tab).

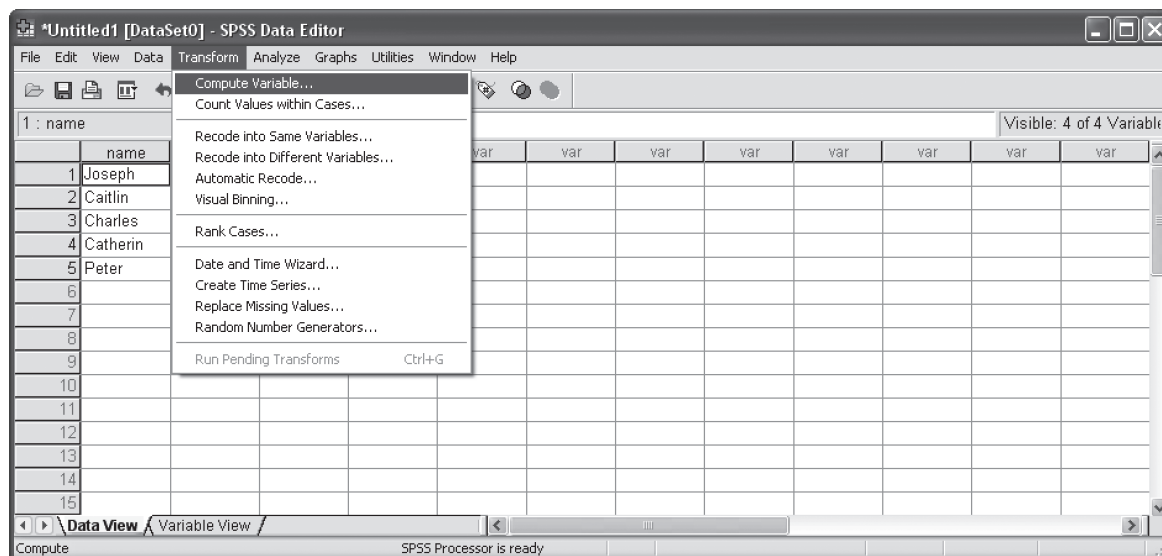
For the further analyses in this chapter, the '-1' will be replaced in the dataset by the original value 75 (Peter's weight).

Creating/calculating a new variable

Suppose that the researcher would like to include an extra column in the example which indicates the 'body-mass index (BMI)'. The BMI is defined as the body weight in kilograms divided by the square of the height in metres.

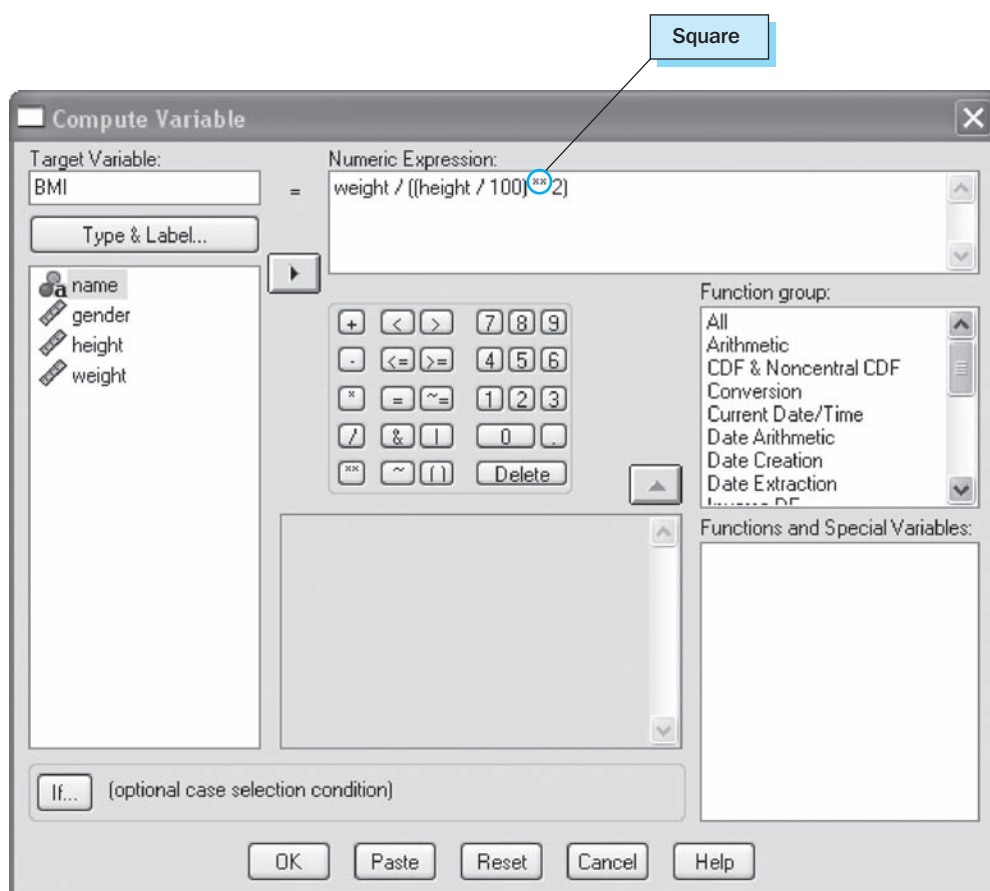
The path to be followed to calculate an additional variable is [Transform/Compute Variable](#) (Figure 1.10).


Figure 1.10



A dialogue window will be displayed such as the one seen in Figure 1.11.

Figure 1.11



In the 'Target Variable' box, type the name of the new variable you would like to calculate (BMI in this case). In the 'Numeric Expression' field, type the formula which the new variable is equal to (instead of typing in the variable names, you may also select the variable names from the left box and click the  button). Figure 1.11 also demonstrates that, if necessary, the possibility also exists to choose from a number of pre-defined functions. Then click 'OK'.

The new variable will now be shown in the 'Data View' screen (Figure 1.12).

Figure 1.12

*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : name Joseph Visible: 5 of 5

	name	gender	height	weight	BMI	var	var	var
1	Joseph	1.00	180.00	75.00	23.15			
2	Caitlin	.00	165.00	67.00	24.61			
3	Charles	1.00	175.00	80.00	26.12			
4	Catherin	.00	170.00	70.00	24.22			
5	Peter	1.00	185.00	75.00	21.91			
6								
7								
8								
9								
10								
11								
12								

Data View Variable View

SPSS Processor is ready

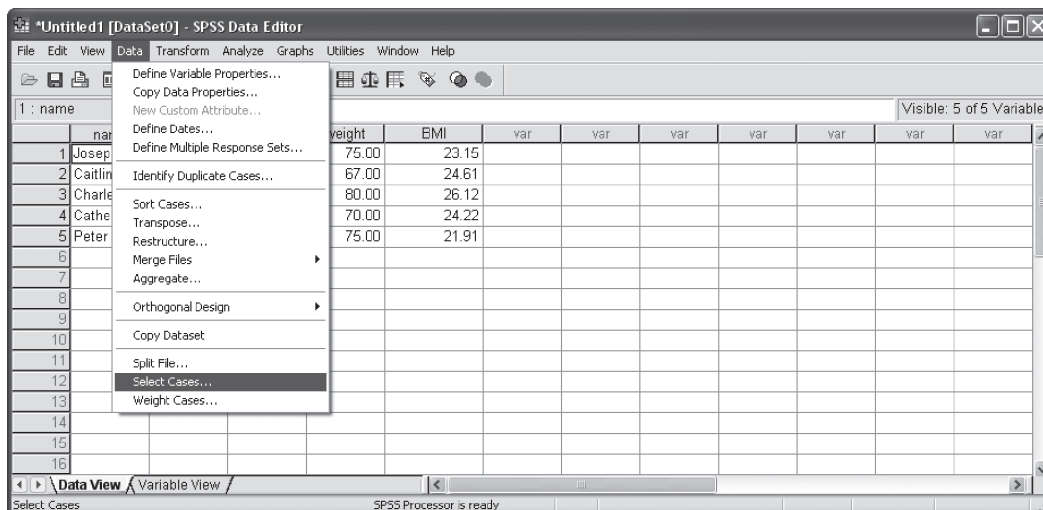
Research on a subset of observations

Selecting cases

Sometimes a certain subanalysis requires that the analysis to be performed may only be done using a number of specific observations (cases). It is then possible to create separate files by deleting the non-relevant observations in the total data file each time, however this method is not efficient. There is a procedure in SPSS which may be used to temporarily turn off the observations which the user does not wish to include in the sub-study (thereby not deleting them permanently). Suppose the researcher in the example would like to select only the male cases (e.g. for a subanalysis), but at the same time, does not wish to permanently delete the other observations (the females).

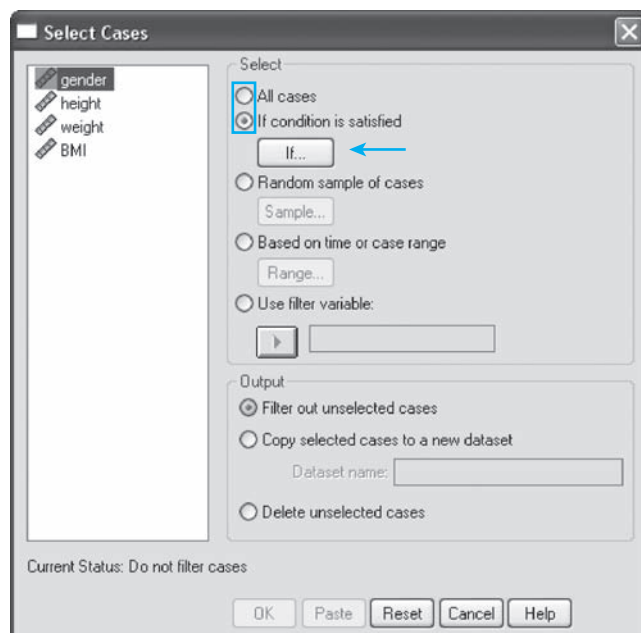
The path that then must be followed is [Data/Select Cases](#) (Figure 1.13).

Figure 1.13



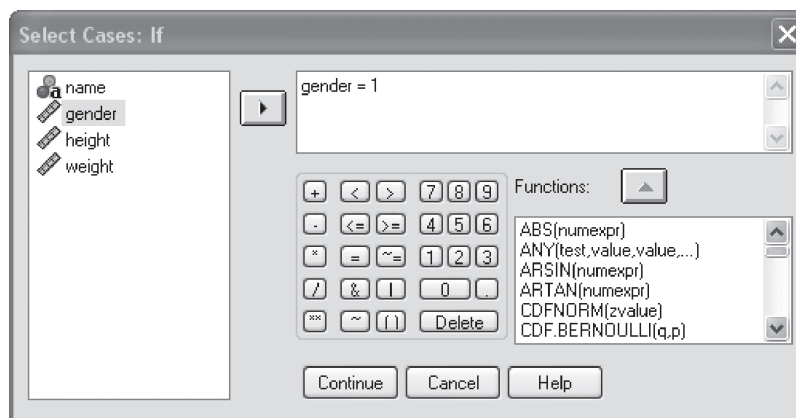
The default setting 'All Cases' must be changed by checking the option 'If condition is satisfied' and then clicking the 'If' button (Figure 1.14).

Figure 1.14



This will cause the screen in Figure 1.15 to be displayed.

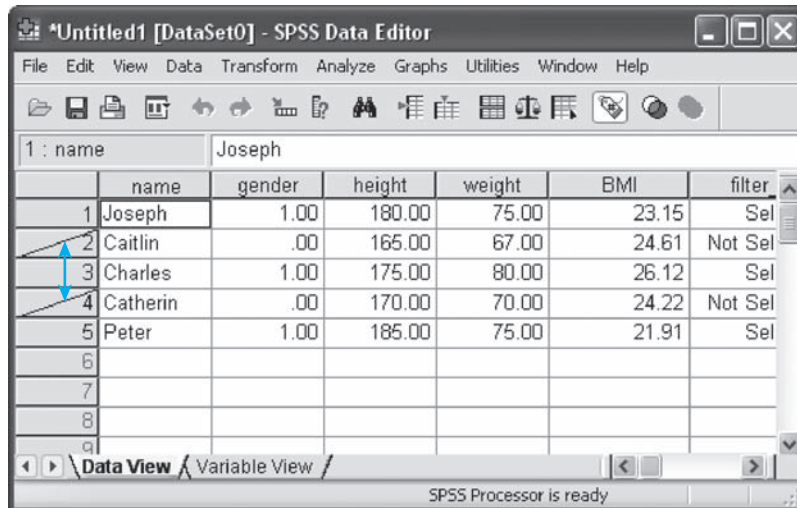
Figure 1.15



Click on 'gender' and make this equal to 1 (1 is the code for the male gender). Then click on 'Continue' and then on 'OK'.

In the 'Data View' window (Figure 1.16), one will see a slanted line through certain respondent numbers indicating that the observations for the females have been turned off. An extra variable has also been created (filter_\$) which indicates whether or not the observation has been selected.

Figure 1.16

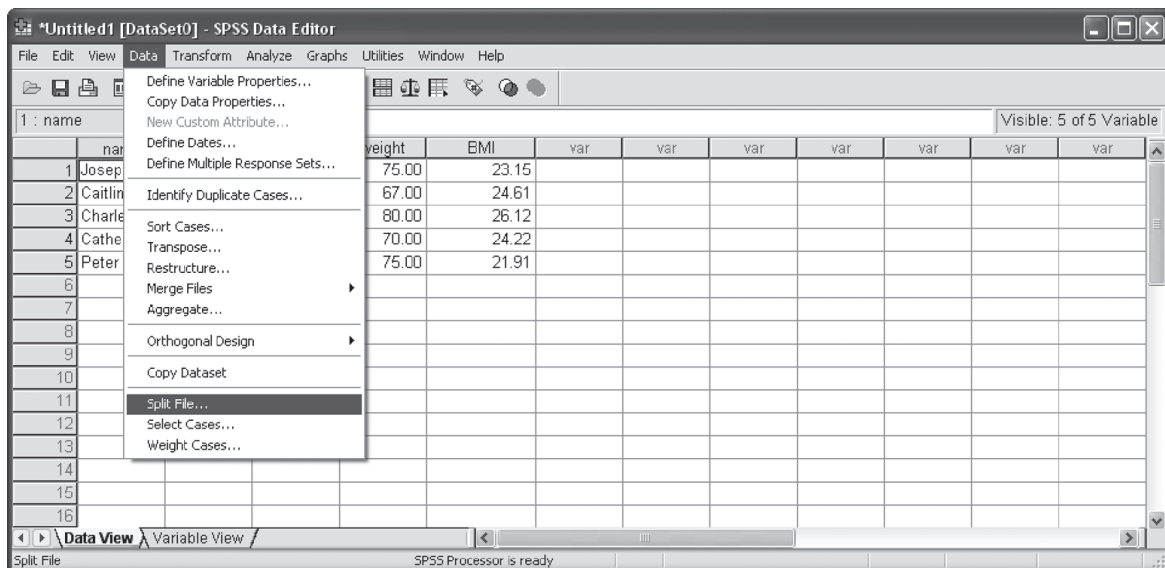


When the researcher would like to go back and work on all of the observations, he will once again follow the path [Data/Select Cases](#) and recheck the default setting 'All Cases'. The extra variable created earlier (filter_\$) remains. If the user wants, he can use this variable again later on for further analyses. He may also remove it by clicking on the grey variable heading with the right mouse button (filter_\$) and then selecting 'Clear'.

Splitting the data file (split file)

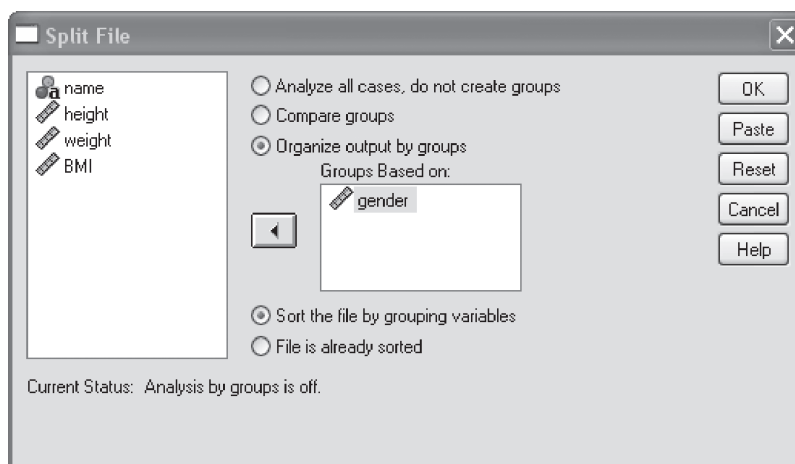
Another option is to split the data file. This means that when an analysis is performed, the user will obtain the results for the different groups for the variable for which the file has been split. Suppose that the researcher wishes to perform separate analyses for the women as well as the men.

Figure 1.17



The path which then must be followed is [Data/Split File](#) (Figure 1.17).

Figure 1.18



Change the default setting ‘Analyze all cases, do not create groups’ in ‘Organize output by groups’. Next, move ‘gender’ to the ‘Groups Based on:’ subscreen. Then click on ‘OK’ (Figure 1.18).

You can now see that the observations have been ranked by ‘gender’ in the Data View tab. Now when the researcher performs an analysis (starting from the next chapter), the output for the indicated analysis will be grouped separately for men and women.

Recoding variables

Let’s say that these five people must complete a questionnaire. For the sake of simplicity, we assume that this questionnaire consists of three questions (statements) in which their preferences regarding candy are being studied. The three statements must be evaluated on a 7-point scale, ranging from ‘totally disagree (1)’ to ‘totally agree (7)’.

- *Question 1:* When I watch television in the evening, I eat candy on a regular basis.
- *Question 2:* If I’m hungry between meals, I will eat fruit more often than candy.
- *Question 3:* I always like to add extra sugar to my dessert.

Their answers are shown in Table 1.2:

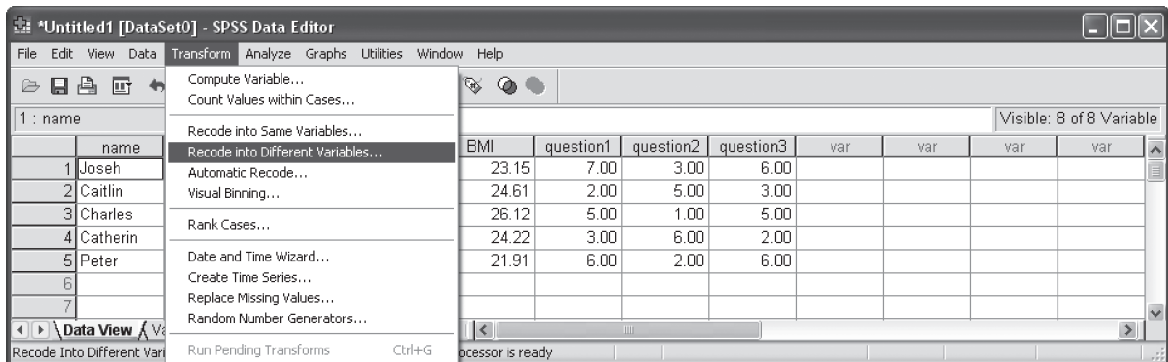
Table 1.2

Name	Question 1	Question 2	Question 3
Joseph	7	3	6
Caitlin	2	5	3
Charles	5	1	5
Catherine	3	6	2
Peter	6	2	6

The data are input in the manner described above. Variable names may not contain spaces in SPSS, therefore type ‘question1’ for ‘Question 1’.

If the researcher wishes to perform an analysis of this data (e.g. calculate an ‘average for candy preference’), he must first determine whether the questions were all scaled ‘in the same direction’. Take question 2 for example. A high score indicates that these people are not so quick to reach for candy, while a high score for questions 1 and 3 indicates that there is a great preference for candy. In other words, question 2 is not scaled in the same direction as questions 1 and 3 and for this reason needs to be recoded.

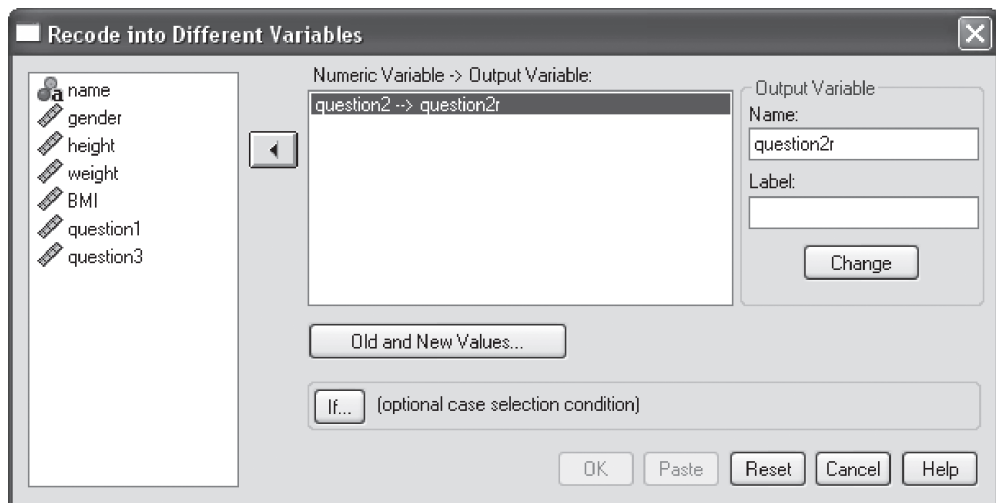
Figure 1.19



For the purpose of recoding there are two options, namely ‘into Different Variables’ and ‘into Same Variables’. If this last option is chosen, the recoded values are placed in the same variable (column) which means that the original variables are overwritten. If an incorrect recoding takes place by accident, the original data will be lost. To prevent this from happening, it is recommended to convert the recoded values into another variable.

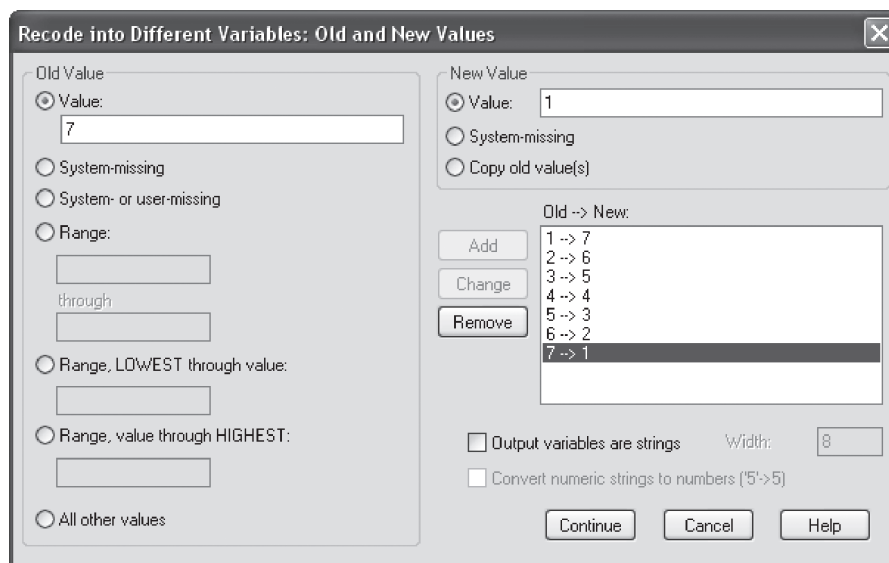
Go to [Transform/Recode into Different Variables](#) (Figure 1.19) which will bring up the subscreen Figure 1.20.

Figure 1.20



Click on the ‘Old and New Values’ button so that you see a dialogue window such as that shown in Figure 1.21. For each value to be recoded, the researcher must input the old and the new value.

Figure 1.21



For ‘Old Value’, fill in the value to be changed (e.g. 7) and under ‘New Value’, type the new value (1). Next, click on the ‘Add’ button and in the Old → New window you will now see the recoding. Repeat this for each of the values to be recoded (the 4 to 4 recoding is also necessary since otherwise SPSS will not incorporate this value in the new variable).

Next, click ‘Continue’ and then ‘OK’ and you will notice that an extra variable with the recoded values has been created in the ‘Data View’ tab (see Figure 1.22). The data file as it is now, can also be found on the cd-rom under the name *introduction.sav*.

Figure 1.22

[illegible]

There is one more way to perform the recoding discussed above. A new variable may be calculated for this type of recoding (see above) as 8 minus the original score, such that the score 1 now becomes $8 - 1 = 7$, etc. Now, a useful average of the variables ‘question1’, ‘question2’ and ‘question3’ may be calculated if desired.

Further reading

Field, A. (2005), *Discovering Statistics Using SPSS*. London: Sage Publications.

Green, S.B., Salkind, N.J. and Akey, T.M. (2000), *Using SPSS for Windows – Analyzing and understanding data*. 2nd ed. Englewood Cliffs, N.J.: Prentice Hall.

SPSS Base 13.0 Users Guide (2004), Chicago, Illinois: SPSS, Inc.

Chapter 2

Descriptive statistics

Chapter objectives

This chapter will help you to:

- Create descriptive tables and graphs
- Compose multiple response tables
- Calculate means and standard deviations of a distribution of observations

Introduction

The objective of this chapter is to illustrate several simple procedures which may serve as the basis to describe a dataset. Further reading in this regard may be found at the end of this chapter. In this chapter, we use the dataset *seniors.sav*. In this file, several buying behaviour concepts have been measured for 310 people (aged 20–34, 50–59, and 60–69), as were their preferences for several types of leisure activities. Finally, inquiries were also made about several socio-demographic variables.

The buying behaviour concepts are shown in Table 2.1. Each concept is a mean of a series of statements relevant to that particular concept. These statements were measured on a 7-point Likert scale (1 = totally disagree, 7 = totally agree).

Table 2.1

Name	Variable	Description
Value consciousness	value	Degree to which people strive for an optimal value-for-money relationship
Price consciousness	price	Degree to which consumers focus on finding and paying low prices
Coupon proneness	coup	Tendency to respond to a sale, because the discount coupon has a positive influence on the purchase evaluation
Sale proneness	sale	Tendency to respond to a sale, because a discount off the original price has a positive influence on the purchase evaluation
Price mavenism	primav	Tendency to be a source of information for many products, services and places where lower prices may be found; consumers are eager to transfer this information to other consumers

Table 2.1 *Continued*

<i>Name</i>	<i>Variable</i>	<i>Description</i>
Price-quality schema	priqua	tendency to consider prices as an indicator of quality
Prestige sensitivity	prest	degree to which higher prices are perceived to be a status symbol
Brand consciousness	brand	degree to which the consumer focuses on brands
Importance of convenience	conv	degree to which consumers feel that ease or convenience are important
Impulsiveness	impuls	degree to which consumers are impulse-driven
Risk-aversion (-)	risk	degree to which consumers have a risk preference
Innovativeness	innov	degree to which consumers would like to be innovative or are open to innovation

The leisure activity variables, not tabled, were coded on a 7-point Likert scale (1 = do not like at all, 7 = like very much) and are: drawing-painting [free1], reading [free2], music [free3], sport [free4], studying [free5], television [free6], going out [free7], cultural activities [free8], and walking [free9].

The coding for the socio-demographic variables is shown in Table 2.2:

Table 2.2

<i>Variable name</i>	<i>Description/Coding</i>
Gender	male (0), female (1)
Mrhp	main person responsible for household purchases: no (0), yes (1)
location	I live in the city (1), in the suburbs (2), in the countryside (3)
numfamily	number of persons in the family: 1(1), 2(2), 3(3), 4(4), >5(5)
age	25–34 y (1), 50–59 y (2), 60–69 y (3)
education	elementary school (1), high school (2), higher education (3)
income	< 1250 EUR (1), 1250–1875 EUR (2), 1876–2500 EUR (3), 2501–3750 EUR (4), > 3750 EUR (5), I prefer not to answer (6)

The two extra variables ‘rank A’ and ‘rank AA’ will be used in Chapter 3.

Figure 2.1 shows the ‘SPSS Data Editor’ with the scores for the variables in the *seniors.sav* dataset.

Figure 2.1

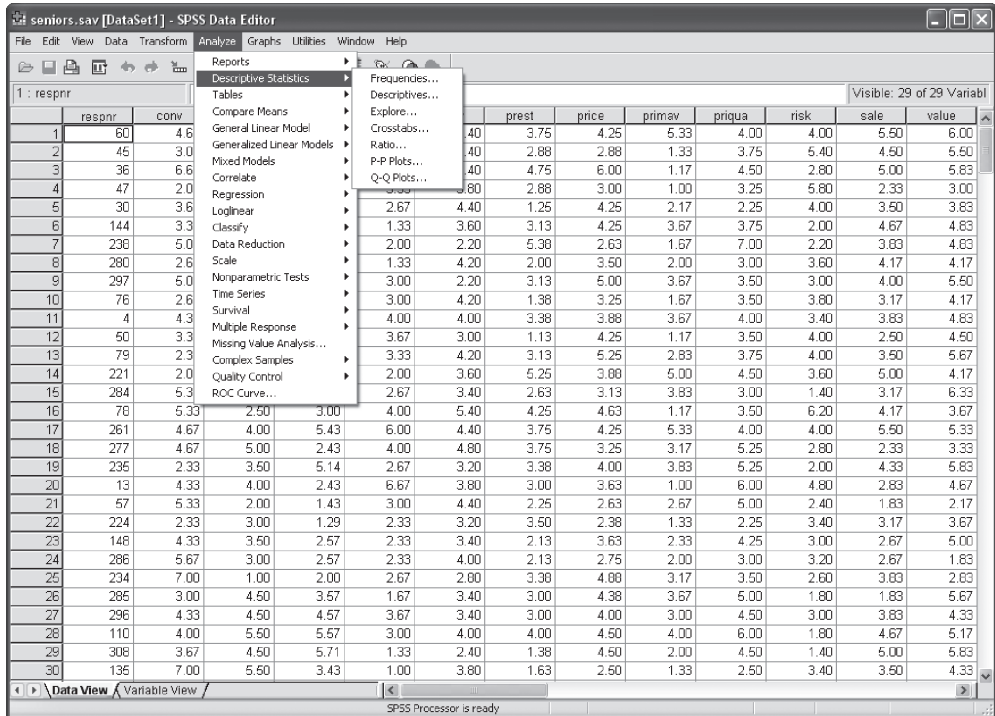
	respnr	corv	brand	coup	impuls	innov	prest	price	primav	priqua	risk	sale	value
1	60	4.67	4.00	5.57	5.00	4.40	3.75	4.25	5.33	4.00	4.00	5.50	6.00
2	45	3.00	3.00	3.29	4.00	5.40	2.88	2.88	1.33	3.75	5.40	4.50	5.50
3	36	6.67	2.50	4.29	3.67	5.40	4.75	6.00	1.17	4.50	2.80	5.00	5.83
4	47	2.00	4.00	3.86	3.33	5.80	2.88	3.00	1.00	3.25	5.80	2.33	3.00
5	30	3.67	2.00	3.57	2.67	4.40	1.25	4.25	2.17	2.25	4.00	3.50	3.83
6	144	3.33	4.00	3.71	1.33	3.60	3.13	4.25	3.67	3.75	2.00	4.67	4.83
7	238	5.00	5.50	2.29	2.00	2.20	5.38	2.63	1.67	7.00	2.20	3.83	4.83
8	280	2.67	1.00	5.14	1.33	4.20	2.00	3.50	2.00	3.00	3.60	4.17	4.17
9	297	5.00	4.00	5.29	3.00	2.20	3.13	5.00	3.67	3.50	3.00	4.00	5.50
10	76	2.67	2.50	2.86	3.00	4.20	1.38	3.25	1.67	3.50	3.60	3.17	4.17
11	4	4.33	4.00	3.71	4.00	4.00	3.38	3.88	3.67	4.00	3.40	3.83	4.83
12	50	3.33	2.00	2.00	3.67	3.00	1.13	4.25	1.17	3.50	4.00	2.50	4.50
13	79	2.33	3.50	2.86	3.33	4.20	3.13	5.25	2.83	3.75	4.00	3.50	5.67
14	221	2.00	4.50	2.86	2.00	3.60	5.25	3.88	5.00	4.50	3.60	5.00	4.17
15	264	5.33	4.00	4.00	2.67	3.40	2.63	3.13	3.83	3.00	1.40	3.17	6.33
16	78	5.33	2.50	3.00	4.00	5.40	4.25	4.63	1.17	3.50	6.20	4.17	3.67
17	261	4.67	4.00	5.43	6.00	4.40	3.75	4.25	5.33	4.00	4.00	5.50	5.33
18	277	4.67	5.00	2.43	4.00	4.80	3.75	3.25	3.17	5.25	2.80	2.33	3.33
19	235	2.33	3.50	5.14	2.67	3.20	3.38	4.00	3.83	5.25	2.00	4.33	5.83
20	13	4.33	4.00	2.43	6.67	3.60	3.00	3.63	1.00	5.00	4.80	2.83	4.67
21	57	5.33	2.00	1.43	3.00	4.40	2.25	2.63	2.67	5.00	2.40	1.83	2.17
22	224	2.33	3.00	1.29	2.33	3.20	3.50	2.38	1.33	2.25	3.40	3.17	3.67
23	148	4.33	3.50	2.57	2.33	3.40	2.13	3.63	2.33	4.25	3.00	2.67	5.00
24	266	5.67	3.00	2.57	2.33	4.00	2.13	2.75	2.00	3.00	3.20	2.67	1.83
25	234	7.00	1.00	2.00	2.67	2.80	3.38	4.88	3.17	3.50	2.60	3.83	2.83
26	265	3.00	4.50	3.57	1.67	3.40	3.00	4.38	3.67	5.00	1.60	1.83	5.67
27	296	4.33	4.50	4.57	3.67	3.40	3.00	4.00	3.00	4.50	3.00	3.83	4.33
28	110	4.00	5.50	5.57	3.00	4.00	4.00	4.50	4.00	5.00	1.80	4.67	5.17
29	308	3.67	4.50	5.71	1.33	2.40	1.38	4.50	2.00	4.50	1.40	5.00	5.83
30	135	7.00	5.50	3.43	1.00	3.80	1.63	2.50	1.33	2.50	3.40	3.50	4.33

Frequency tables and graphs

The calculation of frequency tables is on one hand useful in order to quickly be able to obtain a descriptive idea of the dataset you are working with, and on the other hand to determine, for example, whether or not the distribution male/female in the sample corresponds proportionally to the population data. It is also an excellent tool for performing 'data cleaning'. This essentially means that the user must find out if any erroneous (impossible) data have been entered. Sometimes when the user types in scores on a 7-point scale for example, instead of pressing a number once, this is accidentally done twice, and for example '33' would be entered instead of '3'. It goes without saying that further analysis (for example, the calculation of the mean) is then performed on erroneous data and this can distort the entire analysis. For this reason, we cannot emphasize the importance of the process of data cleaning strongly enough. It is therefore always advisable to create a frequency table for each variable and to check this for the presence of 'unexpected' values.

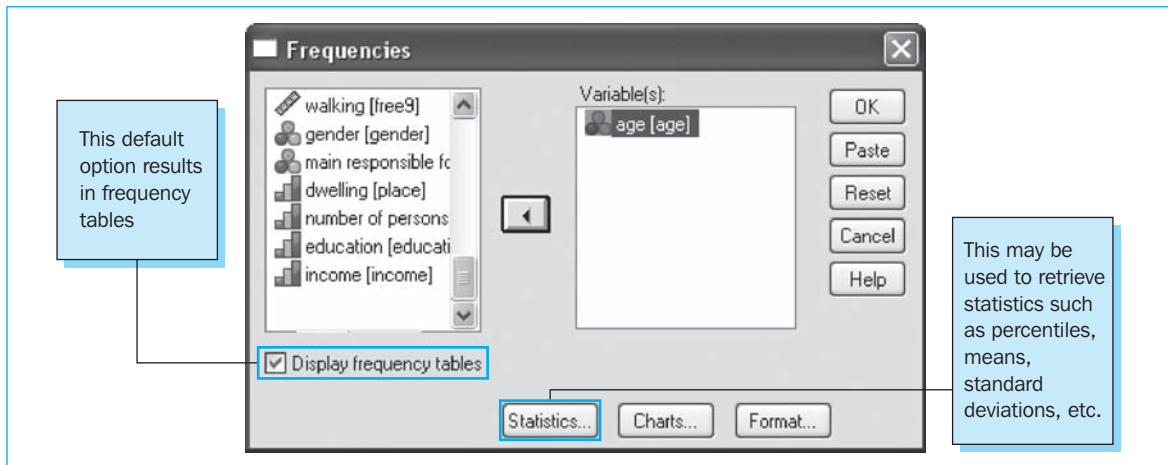
As was mentioned in the description of the dataset, three age groups were surveyed in the example. Suppose the researcher would like to know how many people were surveyed in each of these groups. In order to be able to answer this, a frequency table must be created.


Figure 2.2



Go to: [Analyze/Descriptive Statistics/Frequencies](#) (Figure 2.2).

Figure 2.3

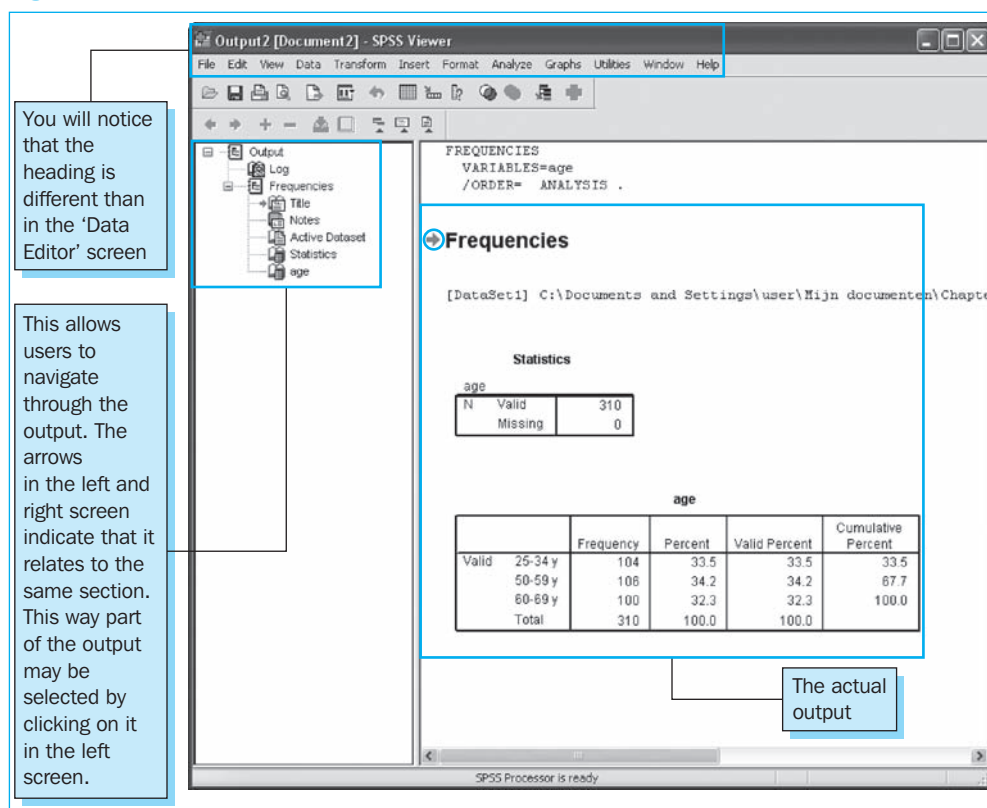


Click on 'age' and then on , and then click 'OK'. The researcher can select multiple variables at the same time, for example by holding down the 'CTRL' key (for

non-sequential variables) or the 'Shift' key (for sequential variables), while indicating the variables. Sequential variables may also be clicked and dragged using the mouse.

The output is obtained in the output window (Figure 2.4).

Figure 2.4



In the further output discussions, only the output in the right subscreen will be shown. If you want to alternate between the 'Output' and the 'Data Editor' windows, you can do this using the Windows toolbar.

The output in the right screen in Figure 2.4 is recaptured in Figure 2.5.

Figure 2.5

Frequencies

Statistics

Age

N	Valid	310
	Missing	0

Age

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 25-34 y	104	33.5	33.5	33.5
50-59 y	106	34.2	34.2	67.7
60-69 y	100	32.3	32.3	100.0
Total	310	100.0	100.0	

As you can see, SPSS has considered 310 observations to be valid, because there were no ‘missing values’ found in any of the observations. Furthermore, it may be determined that 104 ‘25–34 year olds,’ 106 ‘50–59 year olds,’ and 100 ‘60–69 year olds’ were surveyed.

Several percentages have also been calculated. The difference between ‘Percent’ and ‘Valid Percent’ is that with the former, missing values are also viewed as being part of the total while the percentages which are shown in the column ‘Valid Percent’ are calculated for all of the observations which do not contain missing values. In order to illustrate this difference, the frequency table in Figure 2.6 is shown for the variable ‘income’ from the same study.

Figure 2.6

Income					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<1250 EUR	55	17.7	19.7	19.7
	1250–1875 EUR	60	19.4	21.5	41.2
	1876–2500 EUR	54	17.4	19.4	60.6
	2501–3750 EUR	47	15.2	16.8	77.4
	>3750 EUR	10	3.2	3.6	81.0
	I prefer not to answer	53	17.1	19.0	100.0
	Total	279	90.0	100.0	
Missing	99.00	31	10.0		
Total		310	100.0		

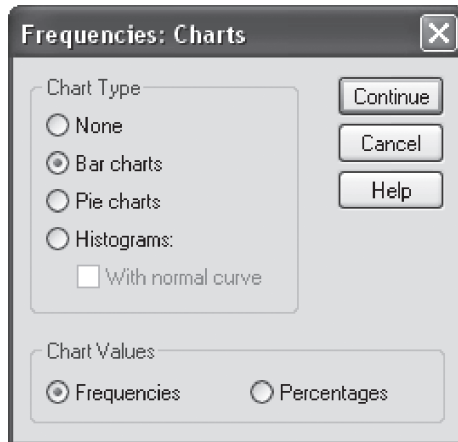
Given the fact that this is fairly personal information that people are not generally quick to disclose, a number of missing values may be expected here (even in the event that the additional option ‘I do not wish to answer this’ is offered in the questionnaire). In fact, it appears that there were 31 respondents in the total dataset (= 10%) who did not fill in an answer (these ‘missings’ were coded as ‘99’).

This means that 279 people did provide a response. In the ‘Percent’ column, we see that the total of 100% is made up of 90% respondents who answered and 10% who did not. The 55 people in the class ‘<1250 EUR’ agrees with the 17.7% in the Percent column which is equal to 55 divided by 310. If however we make an abstraction from the missing observations, these 55 people will agree with the 19.7% in the Valid Percent column (which is 55 divided by 279). The last column shows the cumulative (valid) percentage. This column shows the sum of 21.5 and 19.7, or 41.2.

It is often useful to portray the results obtained in the form of a graph. SPSS offers several simple possibilities for doing this. Imagine the researcher would like to display the results from Figure 2.5 in a graph as well.

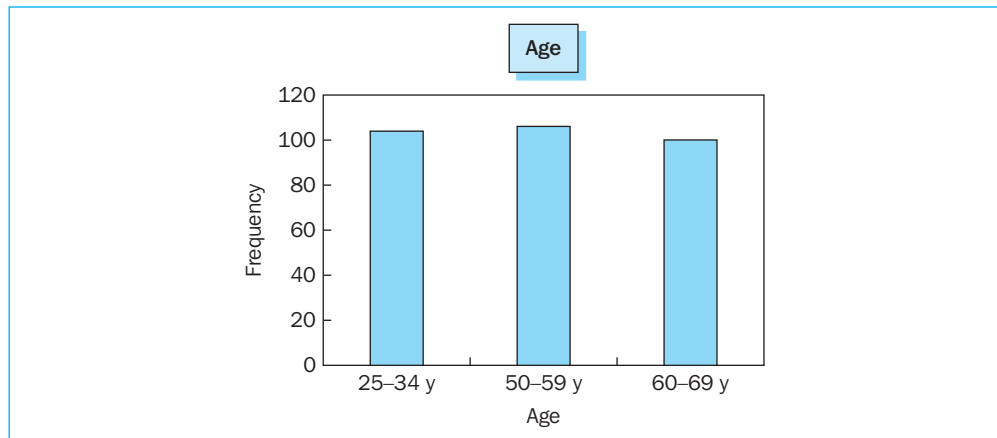
Go to [Analyze/Descriptive Statistics/Frequencies](#) and select age (see Figure 2.3). Then click on Charts at the bottom. The researcher will then see the screen as shown in Figure 2.7.

Figure 2.7



Change the default setting under 'Chart Type' from 'None (no graphs)' into 'Bar Charts'. Now click on 'Continue' and then 'OK' (in the main window). The researcher will then see a bar chart like the one shown in Figure 2.8.

Figure 2.8



These figures may be illustrated in other ways as well. This is demonstrated here using a pie chart (Figure 2.9) for the education variable. Follow the steps again as described for the creation of a bar chart, but this time indicate 'Pie charts' in the window in Figure 2.7. Once this graph has been created (Figure 2.9) in the output window, you may adjust and revise it by double-clicking on the graph. You will now see an image such as that displayed in Figure 2.10.

Figure 2.9

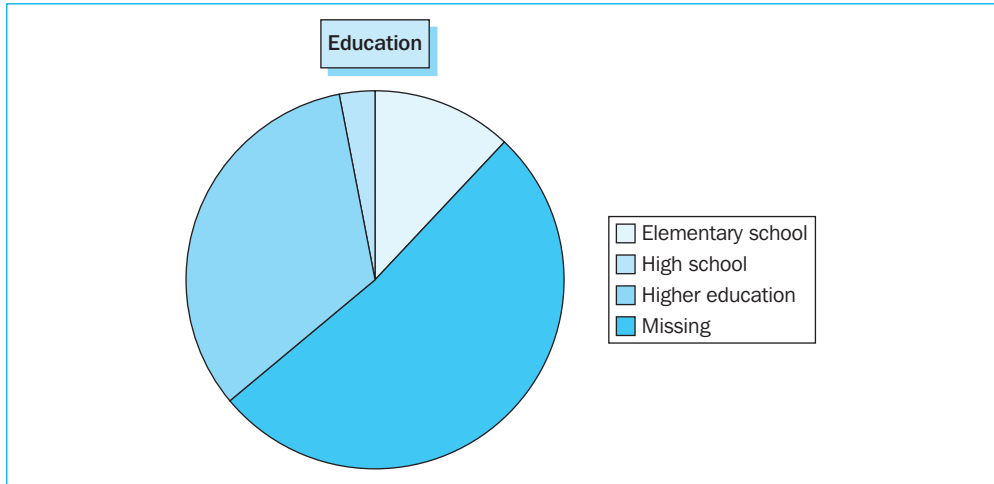
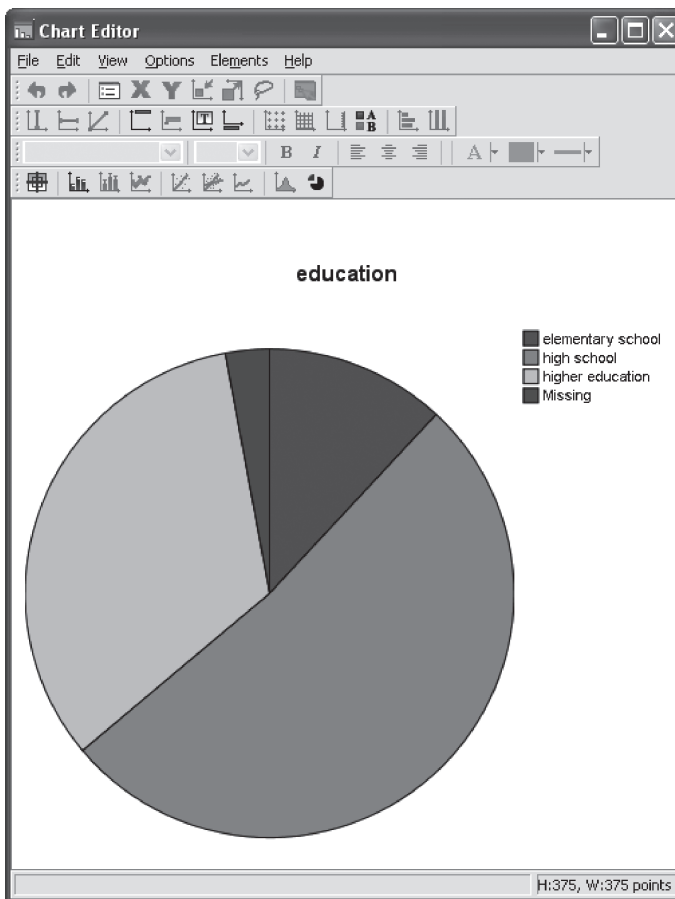
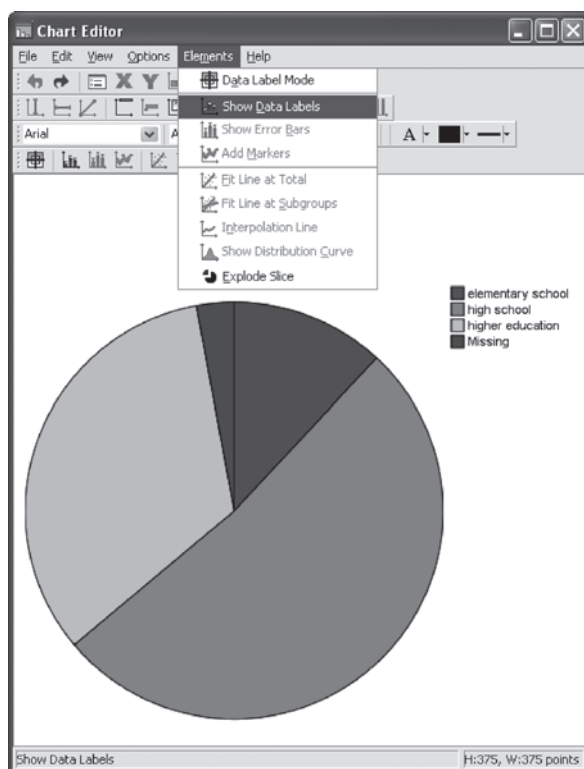


Figure 2.10



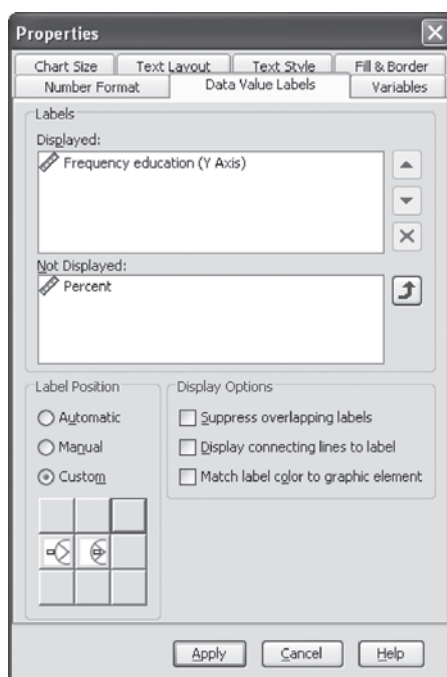
Continue now with [Elements/Show Data Labels](#) (Figure 2.11).

Figure 2.11



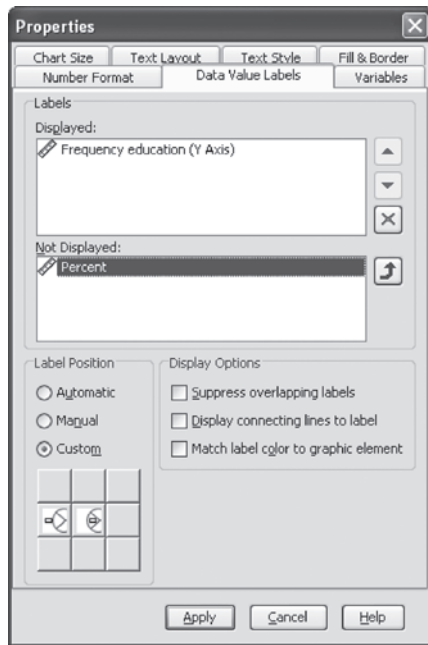
You will now see a window such as that shown in Figure 2.12.

Figure 2.12



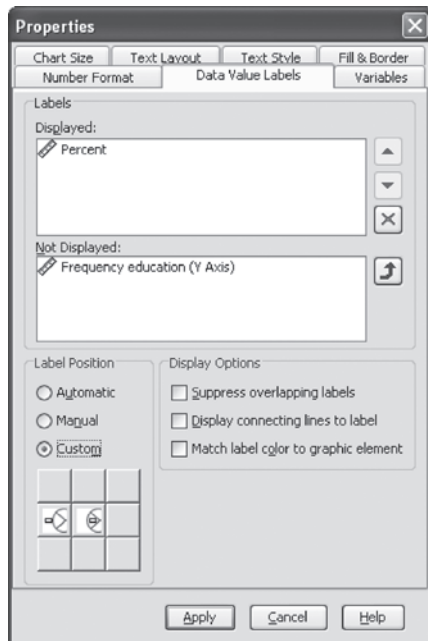
If you want to see percentages in the graph as well as a reference to the different types of education, you must select 'Percent' and move this to the 'Displayed' box by clicking on the green arrow (Figure 2.13).

Figure 2.13



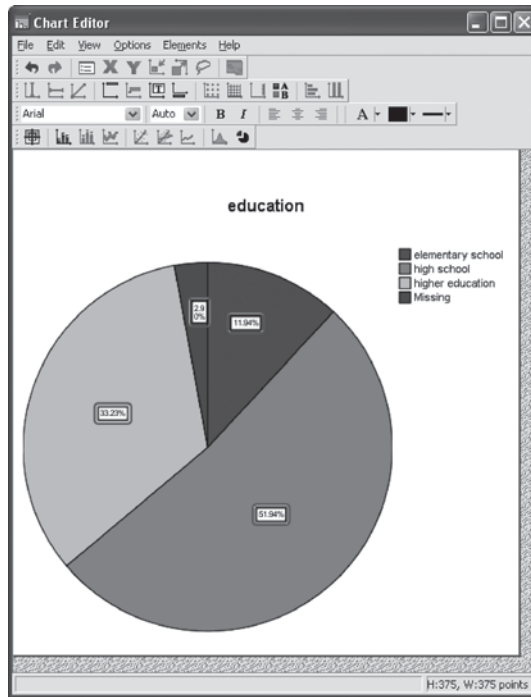
It is not desirable to see a display of the frequency of the responses here. For this reason, the option 'Frequency education' must be selected and moved to the 'Not Displayed' box using the red cross (Figure 2.14).

Figure 2.14



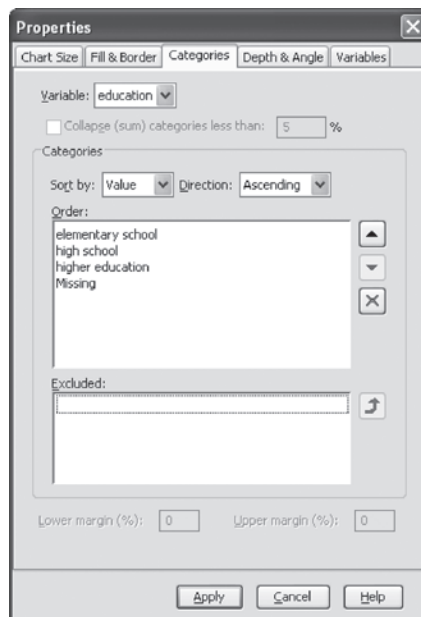
You will now see a pie chart with four pie pieces, which are actually the three types of education plus the portion with the missing values (Figure 2.15).

Figure 2.15



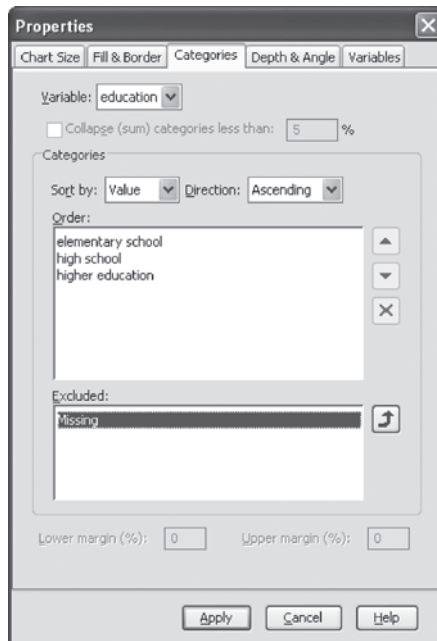
If the researcher prefers to remove this last part from the graph, he must do the following. Double click on the Chart which will cause the 'Properties' window to appear (Figure 2.16).

Figure 2.16



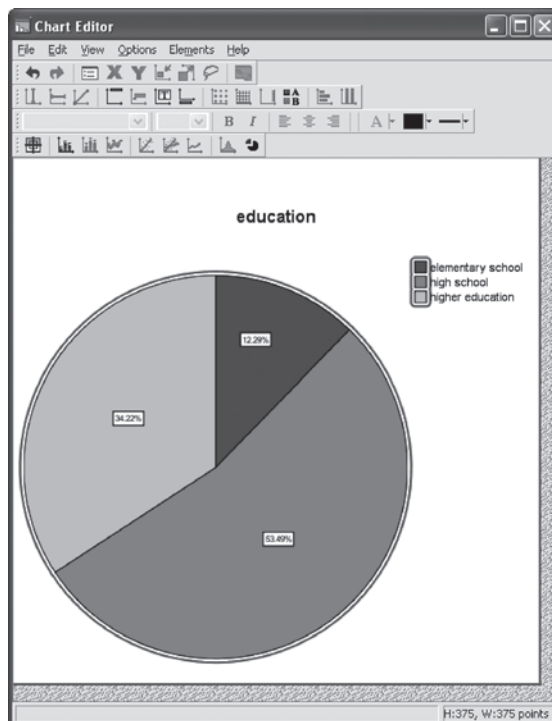
Go to the subscreen ‘Categories’, select ‘Missing’ and move this with the red cross to the ‘Excluded’ box (Figure 2.17).

Figure 2.17



Click on ‘Apply’ and Figure 2.18 will appear.

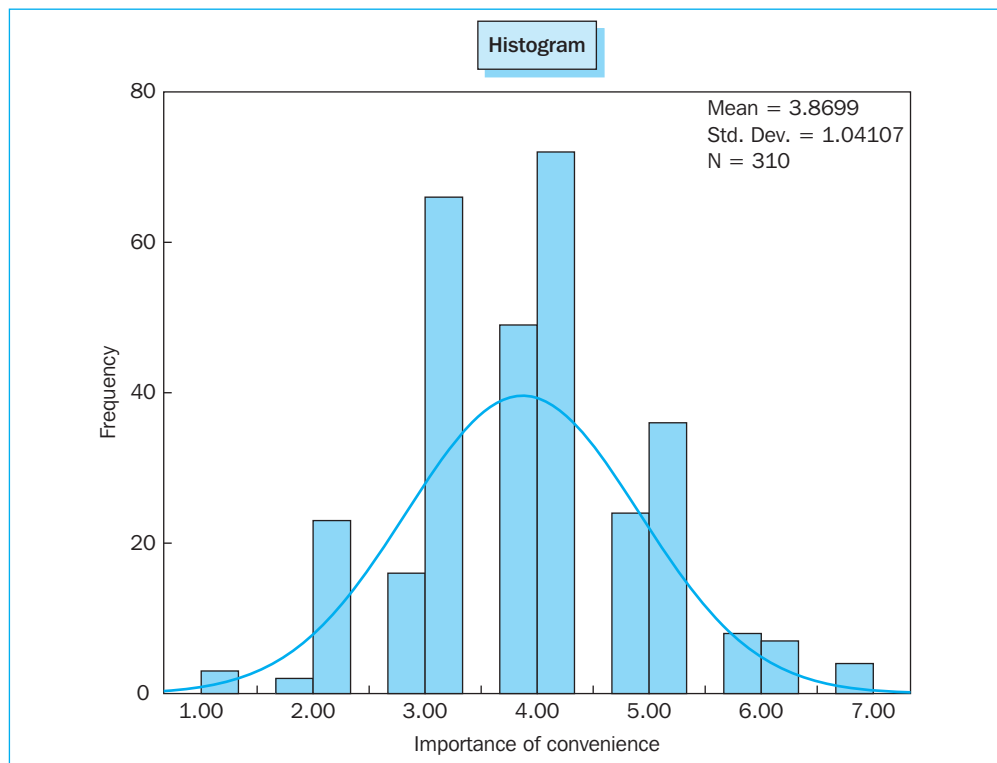
Figure 2.18



You will notice that only the three educational levels are shown, from which it appears that people with a secondary education comprise just over half of the respondents who were prepared to indicate their level of education on the survey.

Figure 2.7 also includes the option 'Histograms' with the option 'With normal curve'. Just like a 'bar chart', a histogram is a way to display frequencies in graphic form. A 'Bar chart' will however show the number of observations for every observed possible response. This can lead to a potentially confusing situation. For example, when the 'importance of convenience' variable is examined, it becomes clear that nearly all of the observations have a different value. When an organised display and an idea of the distribution is the objective here, this may best be obtained by creating a histogram. The researcher must then check 'Histograms' for 'Chart type' (Figure 2.8). If he would also like to compare the distribution obtained with the normal distribution, the option 'with normal curve' must be indicated. For the variables (Figure 2.3), the user may choose 'importance of convenience' only. In Figure 2.19, the relevant histogram with normal distribution may be found.

Figure 2.19



SPSS has created a number of groups itself and at first sight the distribution is not really normal (the frequencies which should be expected from the normal distribution (the curve) do not differ very much from the actual observed frequencies). If you would like to change the number of groups, just double-click on the X-axis.

Statistical tests which formally test this graphic assumption are for example the Kolmogorov-Smirnov test (with Lilliefors correction) and the Shapiro-Wilk W test. Although univariate tests form the subject of the subsequent chapters, the normality test will be treated here.