



Introduction to **Statistics and SPSS** in Psychology

Andrew Mayers

Introduction to **Statistics and SPSS** in Psychology

PEARSON

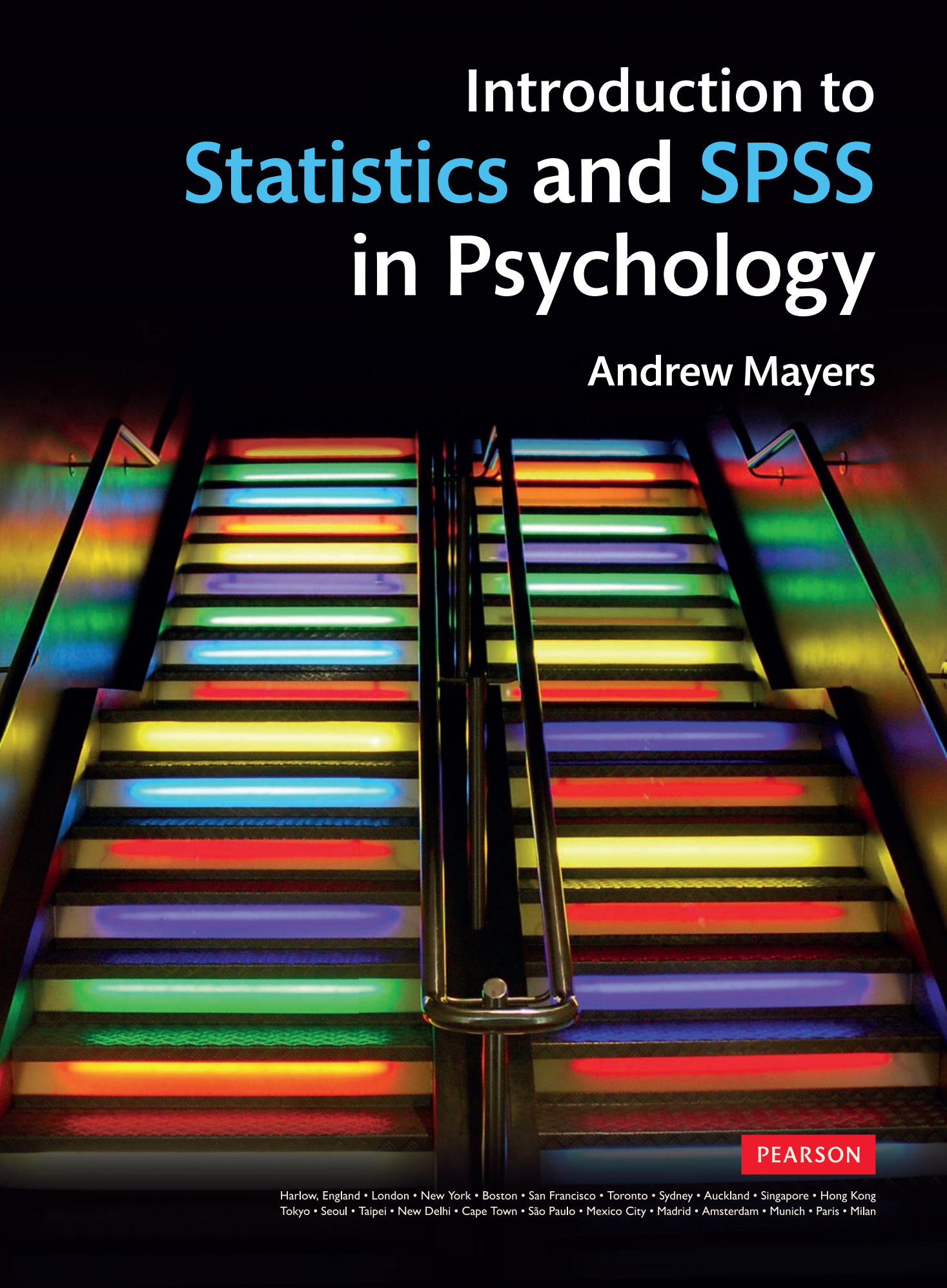
At Pearson, we take learning personally. Our courses and resources are available as books, online and via multi-lingual packages, helping people learn whatever, wherever and however they choose.

We work with leading authors to develop the strongest learning experiences, bringing cutting-edge thinking and best learning practice to a global market. We craft our print and digital resources to do more to help learners not only understand their content, but to see it in action and apply what they learn, whether studying or at work.

Pearson is the world's leading learning company. Our portfolio includes Penguin, Dorling Kindersley, the Financial Times and our educational business, Pearson International. We are also a leading provider of electronic learning programmes and of test development, processing and scoring services to educational institutions, corporations and professional bodies around the world.

Every day our work helps learning flourish, and wherever learning flourishes, so do people.

To learn more please visit us at: www.pearson.com/uk



Introduction to Statistics and SPSS in Psychology

Andrew Mayers

PEARSON

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Auckland • 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 CM20 2JE

Tel: +44 (0)1279 623623

Fax: +44 (0)1279 431059

Website: www.pearson.com/uk

First published 2013 (print and electronic)

© Pearson Education Limited 2013 (print and electronic) [2012 onwards]

The right of Dr Andrew Mayers to be identified as author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

The print publication is protected by copyright. Prior to any prohibited reproduction, storage in a retrieval system, distribution or transmission in any form or by any means, electronic, mechanical, recording or otherwise, permission should be obtained from the publisher or, where applicable, a licence permitting restricted copying in the United Kingdom should be obtained from the Copyright Licensing Agency Ltd, Saffron House, 6-10 Kirby Street, London EC1N 8TS.

The ePublication is protected by copyright and must not be copied, reproduced, transferred, distributed, leased, licensed or publicly performed or used in any way except as specifically permitted in writing by the publishers, as allowed under the terms and conditions under which it was purchased, or as strictly permitted by applicable copyright law. Any unauthorised distribution or use of this text may be a direct infringement of the author's and the publishers' rights and those responsible may be liable in law accordingly.

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.

Contains public sector information licensed under the Open Government Licence (OGL) v1.0.
<http://www.nationalarchives.gov.uk/doc/open-government-licence>.

The screenshots in this book are reprinted by permission of Microsoft Corporation.

Pearson Education is not responsible for the content of third-party internet sites.

ISBN: 978-0-273-73101-6 (print)
978-0-273-73102-3 (PDF)
978-0-273-78689-4 (eText)

British Library Cataloguing-in-Publication Data

A catalogue record for the print edition is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for the print edition is available from the Library of Congress

10 9 8 7 6 5 4 3 2 1
16 15 14 13 12

Getty images

Print edition typeset in 9/12 and GiovanniStd-Book

Print edition printed and bound by Rotolito Lombarda, Italy

NOTE THAT ANY PAGE CROSS REFERENCES REFER TO THE PRINT EDITION

Contents

About the author	x
Acknowledgements	xi
Publisher's acknowledgments	xi
Guided tour	xii
1 Introduction	1
Why I wrote this book – what's in it for you?	2
Why do psychologists need to know about statistics?	2
How this book is laid out – what you can expect	3
Online resources	9
2 SPSS – the basics	10
Learning objectives	10
Introduction	11
Viewing options in SPSS	11
Defining variable parameters	12
Entering data	18
SPSS menus (and icons)	19
Syntax	35
Chapter summary	35
Extended learning task	35
3 Normal distribution	37
Learning objectives	37
What is normal distribution?	38
Measuring normal distribution	43
Statistical assessment of normal distribution	49
Adjusting non-normal data	57
Homogeneity of between-group variance	61
Sphericity of within-group variance	61
Chapter summary	62
Extended learning task	62
4 Significance, effect size and power	63
Learning objectives	63
Introduction	64
Statistical significance	64
Significance and hypotheses	67
Measuring statistical significance	72
Effect size	81
Statistical power	83
Measuring effect size and power using G*Power	83
Chapter summary	87
Extended learning task	88
5 Experimental methods – how to choose the correct statistical test	89
Learning objectives	89
Introduction	90

	Conducting 'experiments' in psychology	90
	Factors that determine the appropriate statistical test	91
	Exploring differences	96
	Examining relationships	99
	Validity and reliability	100
	Chapter summary	101
	Extended learning task	102
6	Correlation	103
	Learning objectives	103
	What is correlation?	104
	Theory and rationale	104
	Pearson's correlation	108
	Spearman's rank correlation	118
	Kendall's Tau-b	121
	Biserial (and point-biserial) correlation	122
	Partial correlation	125
	Semi-partial correlation	131
	Chapter summary	134
	Research example	135
	Extended learning task	136
7	Independent t-test	137
	Learning objectives	137
	What is a t-test?	138
	Theory and rationale	139
	How SPSS performs an independent t-test	144
	Interpretation of output	147
	Effect size and power	148
	Writing up results	149
	Presenting data graphically	150
	Chapter summary	152
	Research example	153
	Extended learning task	153
8	Related t-test	155
	Learning objectives	155
	What is the related t-test?	156
	Theory and rationale	156
	How SPSS performs the related t-test	161
	Interpretation of output	163
	Effect size and power	164
	Writing up results	165
	Presenting data graphically	165
	Chapter summary	168
	Research example	168
	Extended learning task	169
9	Independent one-way ANOVA	170
	Learning objectives	170
	Setting the scene: what is ANOVA?	171
	Theory and rationale	173
	How SPSS performs independent one-way ANOVA	181
	Interpretation of output	185
	Effect size and power	189
	Writing up results	190
	Presenting data graphically	190

Chapter summary	191
Research example	191
Extended learning task	192
10 Repeated-measures one-way ANOVA	194
Learning objectives	194
What is repeated-measures one-way ANOVA?	195
Theory and rationale	195
How SPSS performs repeated-measures one-way ANOVA	203
Interpretation of output	206
Effect size and power	211
Writing up results	212
Presenting data graphically	213
Chapter summary	215
Research example	216
Extended learning task	217
11 Independent multi-factorial ANOVA	218
Learning objectives	218
What is independent multi-factorial ANOVA?	219
Theory and rationale	219
How SPSS performs independent multi-factorial ANOVA	231
Interpretation of output	234
Effect size and power	238
Writing up results	239
Chapter summary	240
Research example	241
Extended learning task	241
Appendix to Chapter 11: Exploring simple effects	243
12 Repeated-measures multi-factorial ANOVA	247
Learning objectives	247
What is repeated-measures multi-factorial ANOVA?	248
Theory and rationale	249
How SPSS performs repeated-measures multi-factorial ANOVA	255
Effect size and power	268
Writing up results	269
Chapter summary	276
Research example	277
Extended learning task	278
13 Mixed multi-factorial ANOVA	279
Learning objectives	279
What is mixed multi-factorial ANOVA?	280
Theory and rationale	280
How SPSS performs mixed multi-factorial ANOVA	291
Effect size and power	301
Writing up results	303
Chapter summary	314
Research example	314
Extended learning task	315
14 Multivariate analyses	317
Learning objectives	317
What are multivariate analyses?	318
What is MANOVA?	318
Theory and rationale	319

How SPSS performs MANOVA	323
Interpretation of output	328
Effect size and power	332
Writing up results	333
Presenting data graphically	333
Repeated-measures MANOVA	334
Theory and rationale	335
How SPSS performs repeated-measures MANOVA	337
Interpretation of output	342
Effect size and power	349
Writing up results	351
Chapter summary	352
Research example (MANOVA)	353
Research example (repeated-measures MANOVA)	354
Extended learning tasks	355
Appendix to Chapter 14: Manual calculations for MANOVA	356
15 Analyses of covariance	362
Learning objectives	362
What are analyses of covariance?	363
What is ANCOVA?	363
Theory and rationale	365
How SPSS performs ANCOVA	370
Effect size and power	378
Writing up results	379
MANCOVA: multivariate analysis of covariance	380
How SPSS performs MANCOVA	382
Effect size and power	390
Writing up results	390
Chapter summary	390
Research examples	391
Extended learning tasks	393
Appendix to Chapter 15: Mathematics behind (univariate) ANCOVA	394
16 Linear and multiple linear regression	397
Learning objectives	397
What is linear regression?	398
Theory and rationale	399
Simple linear regression	399
Effect size and power	408
Writing up results	409
Multiple linear regression	409
How SPSS performs multiple linear regression	418
Chapter summary	431
Research example	432
Extended learning task	432
Appendix to Chapter 16: Calculating multiple linear regression manually	434
17 Logistic regression	440
Learning objectives	440
What is (binary) logistic regression?	441
Theory and rationale	441
How SPSS performs logistic regression	448
Writing up results	458
Chapter summary	458
Research example	459
Extended learning task	460

18 Non-parametric tests	461
Learning objectives	461
Introduction	462
Common issues in non-parametric tests	462
Mann-Whitney U test	464
How SPSS performs the Mann-Whitney U	467
Wilcoxon signed-rank test	473
How SPSS performs the Wilcoxon signed-rank test	476
Kruskal-Wallis test	482
How SPSS performs Kruskal-Wallis	485
Friedman's ANOVA	492
How SPSS performs Friedman's ANOVA	495
Chapter summary	501
19 Tests for categorical variables	503
Learning objectives	503
What are tests for categorical variables?	504
Theory and rationale	505
Measuring outcomes statistically	510
Categorical tests with more than two variables	520
Loglinear analysis when saturated model is rejected	530
Chapter summary	533
Research example	534
Extended learning task	534
20 Factor analysis	536
Learning objectives	536
What is factor analysis?	537
Theory and rationale	537
How SPSS performs principal components analysis	547
Writing up results	557
Chapter summary	558
Research example	559
Extended learning task	560
21 Reliability analysis	561
Learning objectives	561
What is reliability analysis?	562
Theory and rationale	562
How SPSS performs reliability analysis	567
Writing up results	572
Chapter summary	573
Research example	573
Extended learning task	574
Appendix 1: Normal distribution (z-score) table	575
Appendix 2: t-distribution table	578
Appendix 3: r-distribution table	580
Appendix 4: F-distribution table	582
Appendix 5: U-distribution table	585
Appendix 6: Chi-square (χ^2) distribution table	586
References	588
Glossary	590
Index	604

About the author

Dr Andrew Mayers gained his PhD at Southampton Solent University, and has held a number of academic, teaching and research positions. Previously at London Metropolitan University and the University of Southampton, he is now a Senior Lecturer in psychology at Bournemouth University. He teaches statistics and clinical psychology to undergraduate and postgraduate students, and has received a number of teaching awards. His research focuses on mental health, particularly in children and families. Currently, that work focuses on postnatal depression, children's sleep, community mental health, care farming, and behavioural, emotional and emotional problems in children. He has frequently appeared on national television and radio about children's sleep problems, and has published widely on factors relating to sleep and mood. He is passionate about reducing mental health stigma and supports a number of groups, such as the national Time to Change campaign. He is a board member of Barnardo's Bournemouth Children's Centres and is Patron for Bournemouth and District Samaritans.

Companion Website

For open-access **student resources** specifically written to complement this textbook and support your learning, please visit **www.pearsoned.co.uk/mayers**



Lecturer Resources

For password-protected online resources tailored to support the use of this textbook in teaching, please visit **www.pearsoned.co.uk/mayers**

Acknowledgements

I would like to thank my friends and family for sticking with me during this mammoth project. I particularly dedicate this book to my wife Sue, to whom I will always be grateful for her support. I also thank our Sammy, Holly, Katy and Simon for all their encouragement over the years. Finally, I would like to thank all of my students who have used the draft versions of this book and have given me such valuable feedback.

Publisher's acknowledgments

We are grateful to the following for permission to reproduce copyright material:

Figures

Figure 1.2 from screenshots from IBM SPSS Statistics, copyright © IBM SPSS Statistics Software; Figure 2.11 from Microsoft Excel screenshots and icons, Microsoft product screenshots reprinted with permission from Microsoft Corporation; Figure 4.5 from Normal Distribution Calculator, <http://stattrek.com/Tables/Normal.aspx>, copyright © 2012 StatTrek.com. All Rights Reserved; Figure 4.7 from G*Power opening screen, <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>, Letzte Änderung: 20.08.2012. Reproduced with kind permission from Professor Dr Axel Buchner; Figure 4.8 from G*Power outcome, <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>, Letzte Änderung: 20.08.2012. Reproduced with kind permission from Professor Dr Axel Buchner; Figure 4.9 from G*Power outcome for calculating required sample size, <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>, Letzte Änderung: 20.08.2012. Reproduced with kind permission from Professor Dr Axel Buchner; Figure 10.9 from G*Power data input screen for repeated-measures one-way ANOVA <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>, Letzte Änderung: 20.08.2012. Reproduced with kind permission from Professor Dr Axel Buchner

In some instances we have been unable to trace the owners of copyright material, and we would appreciate any information that would enable us to do so.

Guided Tour

Using the SPSS file *Sleep quality*

Select **Analyze** → **Regression** → **Linear...** as shown in Figure 6.24.

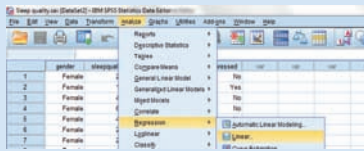


Figure 6.24 Semi-partial correlation (via regression) – step 1

In new window (see Figure 6.25) transfer *Mood* to **Dependent** window → transfer *Sleep quality perceptions* and *Age* to **Independent(s)** window → click **Statistics...**

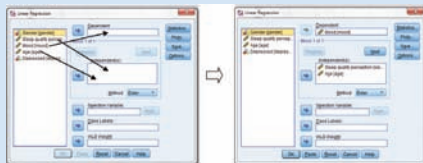


Figure 6.25 Semi-partial correlation – step 2

In new window (see Figure 6.26), tick boxes for **Estimates**, **Model fit** and **Part and partial correlations** → click **Continue** → click **OK**

6.12 Nuts and bolts

Partial correlation terminology

The following terms indicate the extent to which an additional variable might explain the original relationship.

Explanation: An explanation occurs when the strength of the original relationship has been altered by the effect of additional variables. That explanation may be 'full' or 'partial'. If factoring out variables causes the original correlation to be reduced to zero, we can say that we have 'full explanation'. The additional variable(s) explained all of the relationship we originally observed; there was no relationship in the first place. This could be an example of a 'spurious correlation' (see next section).

Partial explanation: If the introduction of additional variables has some effect on original correlation we can say that we have 'partial explanation'. This effect might be very small or it could be substantial. In some cases, the relationship might be strengthened.

Spurious correlation

If the action of a partial correlation results in 'full explanation' (whereby the original correlation is 'wiped out'), it begs the question of whether there was really correlation in the first place. We might call that 'spurious'. Say we find a strong correlation between the number of driving errors and university exam results. It seems illogical to imagine that there might be a relationship, but a correlation analysis indicates otherwise. However, if we then controlled for alcohol intake, we might find that the correlation disappears! The correlation was spurious because the relationship between driving errors and exam scores was actually explained by the amount of alcohol consumed.

6.13 Nuts and bolts

Examples of spurious correlation

In Table 6.8 there are some examples of apparent correlations. However, all is not what it seems: the relationship is actually due to something else altogether!

Table 6.8 When is a correlation not a correlation?

Apparent correlation	Actual explanation
A positive correlation between the number of fire engines attending a fire and the damage that ensues suggests fire engines cause the damage.	The size of the fire is related to the amount of damage – larger fires simply need more fire engines.
In a psychology class, students with longer hair got better exam results than those with shorter hair. It could be concluded that longer hair is related to better academic performance.	Since the girls in the class had longer hair than the boys, it is more likely that the effect was due to gender, not hair length.

SPSS screenshots and accompanying step-by-step instructions guide you through the processes you need to carry out, using datasets provided on the companion website.

Nuts and Bolts boxes help you to understand the conceptual issues and to go beyond the basics.

7.4 Mini exercise

Between-groups or within-groups?

If you are new to statistics, you may still be a little confused about how to determine whether a design is between-groups or within-groups. The following exercise might help to clarify that for you. We explored some of these points in Chapter 5, so you might want to read that again. Look at the following short scenarios and decide whether they are an example of a between-group or within-group study.

1. A group of UK students are compared with those from the USA on how many hours they watch television.
2. One group of depressed patients are given two different types of drug, at different times, to assess how well their symptoms improve.
3. Children are compared with adults in respect of how many green vegetables they eat.
4. Several questionnaires are given to one group of people to see how they differ on several outcome measures, but in respect of their nationality, ethnicity and religious belief.
5. A group of students are given two tests: before one of these tests they are given some tips on revision skills. Their test scores are compared.

Look at the answers below. How did you do?

1. Between;
2. Within;
3. Between;
4. Between;
5. Within.

You may have had some trouble with Question 4. It is quite common to believe that this constitutes a within-group design (because several questionnaires were given to one group) but it is not. It would be within-groups only if the same questionnaire was repeated. For example, we could give a stress questionnaire to a single group, then we could manipulate that stress (such as make them watch a scary movie), and then we would give them the same stress questionnaire again. In short, a between-group study explores differences in the characteristics of the sample, using different groups; a within-group study examines different conditions performed across a single group. In the case of Question 4, the independent variables are nationality, ethnicity and religious belief, not the number of questionnaires used.

Assumptions and restrictions

There are a number of criteria that we must satisfy before we can consider using an independent t-test to explore outcomes. The independent variable must be categorical and must be

Mini exercises are practical things you can do to improve your understanding of new concepts.

9.7 Take a closer look

Planned contrasts (a summary)

Planned contrasts are used to confirm predictions that have been made about the relationship between three or more groups of an independent variable about an outcome on a dependent variable. There are two types of planned comparison – orthogonal and non-orthogonal.

Orthogonal:	Used where the experimental conditions are compared with a control group, followed by a comparison between the experimental groups. Adjustments for multiple comparisons are not needed.
Non-orthogonal:	Used where there is no control group, but where all of the groups are independent and can be compared with each other. Adjustments must be made to account for multiple comparisons.

Post hoc tests

If no specific prediction has been made about differences between the groups, *post hoc* tests must be used to determine the source of difference. We can also choose to use *post hoc* tests in preference to non-orthogonal planned contrasts. However, there must be a significant ANOVA outcome in order for *post hoc* tests to be employed. If we try to run these tests on a non-significant ANOVA outcome it might be regarded as 'fishing'. Also, we run *post hoc* tests only if there are three or more groups. If there are two groups we can use the mean scores to indicate the source of difference. *Post hoc* tests explore each pair of groups to assess whether there is a significant difference between them (such as Group 1 vs. 2, Group 2 vs. 3 and Group 1 vs. 3). Most *post hoc* tests account for multiple comparisons automatically (so long as the appropriate type of test has been selected – see later).

The mathematics behind *post hoc* tests is relatively complex, so we will focus on how we run tests in SPSS. As we will see later, SPSS has something like 18 *post hoc* tests to choose from, but only a few are routinely used in practice. Each test employs a different method of calculating the result, depending on how it accounts for multiple comparisons, equality of variance and equal group sizes. An overview of the types of test is shown in Box 9.8. Many researchers employ a *Tukey* analysis, since it is relatively conservative (without losing too much power). However, that test should probably not be used when there are unequal group sizes, or if equality of variances has been violated. We will probably know whether we have equal group sizes prior to analysis. However, we will not know the outcome of tests for homogeneity of variance until we look at the SPSS output. If we know that we have unequal group sizes we should request *Gabriel's* or *Hochberg's* *GF2* *post hoc* tests (instead of *Tukey*) when we set the parameters to run independent one-way

Take a closer look boxes explore particular aspects of the topics in more detail.

10.2 Calculating outcomes manually

Repeated-measures ANOVA calculation



To illustrate how we can calculate outcomes for repeated-measures one-way ANOVA, we will use some data that relate to the research question posed by CALM earlier. You will find a Microsoft Excel spreadsheet associated with these calculations on the web page for this book.

Table 10.1 Number of words recalled in each condition

Participant	Word condition			Case mean	Case variance
	W	WP	WPS		
1	62	70	82	71.33	101.33
2	63	68	68	66.33	8.33
3	65	61	72	66.00	31.00
4	68	75	88	77.00	103.00
5	69	72	80	73.67	32.33
6	71	77	80	76.00	21.00
7	78	82	87	82.33	20.33
8	75	73	79	75.67	9.33
9	70	77	82	76.33	36.33
10	71	76	84	77.00	43.00
11	60	70	77	69.00	73.00
Condition mean	68.36	72.82	79.91		Σ 479.00
Grand mean	73.70		Grand variance	53.72	

Key: W (word); WP (word and picture); WPS (word, picture and sound)

Chapter summary

In this chapter we have explored reliability analysis. At this point, it would be good to revisit the learning objectives that we set at the beginning of the chapter.

You should now be able to:

- Recognise that we use reliability analysis to examine the consistency of responses to a group of items or questions. It is the next logical step from factor analysis, where the validity of themes and sub-themes has been established.
- Comprehend that reliability is an important factor in research. It confirms the consistency and repeatability of the methods used and the data gained from that research. In establishing reliability, we are adding to the validity of the constructs that we seek to measure.
- Understand different types of reliability. Repeatability of measures can be examined using test-retest reliability. Consistency of observational ratings between researchers can be explored using inter-rater reliability. Stability of observations from a single researcher can be investigated with intra-rater reliability. The internal consistency of responses to a group of items can be examined with split half reliability, but it is better analysed with Cronbach's alpha (and other measures associated with reliability analysis).
- Appreciate that there are very few assumptions and restrictions associated with reliability analysis. It is important that we account for reverse scoring and adjust if need be.
- Perform analyses using SPSS.
- Understand how to present the data and report the findings.

Research example



It might help you to see how principal components analysis has been applied in a research context. You could read the following paper (an overview is provided below).

Sapin, C., Simeoni, M.C., El Khammar, M., Antonietti, S. and Augier, P. (2005). Reliability and validity of the VSP-A, a health-related quality of life instrument for ill and healthy adolescents. *Journal of Adolescent Health*, 36 (4) 327-336. DOI: <http://dx.doi.org/10.1016/j.jadohealth.2004.01.016>

If you would like to read the entire paper you can use the DOI reference provided to locate that (see Chapter 1 for instructions).

We last saw this paper in Chapter 20, when we explored how the authors used principal components analysis to examine the factor structure of the VSP-A (Vécu et Santé Perçue de l'Adolescent – or, translated, the life and health perceptions of adolescents). From 37 questions, 10 factors were identified: Vitality (five items), psychological well-being (five items), relationships with friends (five items), leisure activities (four items), relationships with parents (four items), physical well-being (four items), relationships with teachers (three items), school performance (two items), body image (two items), and relationships with medical staff (three items). This paper also examines the internal consistency of those factors.

The results showed that all items possessed a minimum item-total correlation of 0.40 (so were at least moderate). The Cronbach's α for all factors exceeded 0.74, and no factor would benefit

Calculating outcomes manually boxes show you how to do the calculations by hand so that you understand how they work.

Research examples put the statistical tests in the context of real-world research, while the chapter summaries bring everything together and recap what you've read.

Extended learning tasks



You will find the data sets associated with these tasks on the website that accompanies this book, (available in SPSS and Excel format). You will also find the answers there.

ANCOVA learning task

Following what we have learned about ANCOVA, answer the following questions and conduct the analyses in SPSS and G*Power. (If you do not have SPSS, do as much as you can with the Excel spreadsheet.) For this exercise, we will look at a fictitious example of treatment options for a group of patients. We will explore the effect of drug treatment, counselling or both on a measure of mood (which is measured on a scale from 0 [poor] to 100 [good], and is taken before and after the intervention). There are 72 participants in this study, with 24 randomly assigned to each of the treatment groups.

Open the data set **Mood and treatment**

1. Which is the independent variable (and describe the groups)?
2. What is the dependent variable?
3. What is the covariate?
4. What assumptions should we test for?
5. Conduct the ANCOVA test.
 - a. Describe how you have accounted for the assumptions.
 - b. Describe what the SPSS output shows, including pre- and post-treatment analyses.
 - c. Describe the effect on estimated marginal means.
 - d. Describe whether you needed to conduct post hoc analyses.
 - e. Run them if they were needed.
6. Also show the effect size and conduct a power calculation, using G*Power.
7. Report the outcome as you would in the results section of a report.

MANCOVA learning task

Following what we have learned about MANCOVA, answer the following questions and conduct the analyses in SPSS and G*Power (you will not be able to perform this test manually). For this exercise, we will look at some data that explore the impact of two forms of treatment on anxiety and mood outcomes. The treatments are cognitive behavioural therapy (CBT) and medication. A group of 20 anxious patients are randomised into those treatment groups. Ratings of anxiety and depression are made by the clinician eight weeks after treatment. Both scales are scored in the range of 0-100, with higher scores representing poorer outcomes. To ensure that these outcomes are not related to prior anxiety, the anxiety ratings are also taken at baseline.

Open the SPSS data **CBT vs. drug**

1. Which is the independent variable (and describe the groups)?
2. What are the dependent variables?
3. What is the covariate?
4. What assumptions should we test for?
5. Conduct the MANCOVA test.
 - a. Describe how you have accounted for the assumptions.
 - b. Describe what the SPSS output shows, including pre- and post-treatment analyses.
 - c. Describe the effect on estimated marginal means.

Extended learning tasks help you to go further, using the datasets provided on the website to carry out extra data analysis.

Visit www.pearsoned.co.uk/mayers for datasets to use for the exercises in the text, answers to the all learning exercises, revision questions and much more.

This page intentionally left blank

1

INTRODUCTION



Why I wrote this book – what's in it for you?

There are a lot of statistics books around, so why choose this one? I have been teaching research methods and statistics in psychology for many years, in several universities. When I recently set about writing my lecture notes, I had to choose a course book to recommend. When I looked at what was available I noticed a number of things. Some books explain when to use a statistical test, and give a broad overview of the theory and concepts, but don't show you how to run it using statistical analysis software. Others show you just how to run the test in that software, but don't explain how and when to use the test, nor do they tell you very much about the theory behind the test. There are several that are very complicated, with loads of maths and formulae – and take themselves far too seriously. Others still are less serious in their approach. I wanted to find something in between all of that; I hope this is it.

In this book you should find sufficient theory and rationale to tell you when you should use a test, why you should use it and how to do so. I will also explain when it is probably not so good to use the test, if certain assumptions are not met (and what to do instead). Then there's the maths thing. I know that most people hate maths, but there is good reason for learning this. When I started studying psychology and statistics, computers and statistical analysis software were all pretty new. It took so long for the valves on the computer to warm up that, by the time it was ready, the data were too old to use. So we had to use maths. Once using a computer was viable, statistical analysis software became the thing to use and it was all very exciting. There seemed little need to ever go back to doing it by hand, I thought. Press a few buttons and off you go. However, when I started teaching statistics, I had another go at doing it all manually and was surprised how much it taught me about the rationale for the test. Therefore, I have decided to include some sections on maths in this book. I really do recommend that you try out these examples (I have attempted to make it all quite simple) – you may learn a lot more than you imagined.

For many, statistics is their very idea of hell. It need not be that way. As you read this book, you will be gently led and guided through whole series of techniques that will lay the foundations for you to become a confident and competent data analyst. How can I make such a bold claim? Well, you only need to ask my students, who have read various iterations of this book. Their feedback has been one of the most motivating aspects of writing it. Over the past few years, several hundred psychology students have used draft versions of this book as part of their studies. They have frequently reported on how the book's clarity and humour really helped them. Many have told me that the friendly style has helped them engage with a subject that had always troubled them before. They also like the unique features of the book that combine theory, rationale, step-by-step guides to performing analyses, relevant real-world research examples, and useful learning exercises and revision.

Above all, I want to make this fun. There will be occasional (hopefully appropriate) moments of humour to lighten the mood, where points may be illustrated with some fun examples. I hope you enjoy reading this book as much as I enjoyed writing it. If you like what you see, tell your friends; if you hate it, don't tell them anything.

Why do psychologists need to know about statistics?

Much of what we explore in psychological research involves people. That much may seem obvious. But because we are dealing with people, our investigations are different to other scientific methods. All the same, psychology remains very much a science. In physical science, 'true experiments' manipulate and control variables; in psychology, we can do that only to a certain extent. For example, we cannot induce trauma in a group of people, but we can compare people

who have experienced trauma with those who have not. Sometimes, we can introduce an intervention, perhaps a new classroom method, and explore the effect of that. All of this is still scientific, but there will always be some doubt regarding how much trust we can put in our observations.

A great deal of the time a psychology researcher will make predictions and then design studies to test their theory. We may observe children in a classroom, or investigate attitudes between two groups of people, or explore the risk factors for depression. When we design our experiments and research studies, we will be pleased when our predicted outcomes have been demonstrated. However, we need to be confident that what we have observed is due to the factors that we predicted to be ‘responsible’ for that outcome (or that might illustrate a relationship) and not because of something else. The observed outcome could just possibly have occurred because of chance or random factors. We are dealing with people, after all. Try as we might, we cannot control for all human factors or those simply down to chance. That's where statistics come in.

Throughout this book you will encounter a whole series of different statistical techniques. Some will be used to explore differences between groups, others examine changes across time, while some tests may simply look at relationships between outcomes. Whatever the focus of that investigation, we need to find some way to measure the likelihood that what we observed did not happen by chance, thus increasing our confidence that it probably occurred because of the factors that we were examining. The statistical analyses in this book have one thing in common: they express the likelihood that the outcome occurred by chance. We will see how to apply that to the many contexts that we are likely to encounter in our studies.

1.1 Take a closer look

Who should use this book?



- This book is aimed at anyone who needs some direction on how to perform statistical analyses.
- The main target audience is probably psychology students and academics, but I hope this book will be equally useful for those working in medicine, social sciences, or even natural sciences.
- Most students are likely to be undergraduates, but this book should also be a valuable resource to postgraduates, doctoral students, lecturers and researchers.
- You may be new to all of this statistics stuff, or an old lag in need of a refresher.
- Whatever your reason for picking up this book, you are most welcome.

How this book is laid out – what you can expect

Introductory chapters: the basics

Chapter 2 will introduce you to some of the basic functions of **SPSS** (a software package designed for analysing research data). In this book, we are using SPSS version 19. You will be shown how to create data sets, how to define the variables that measure the outcome, and how to input those data. You will learn how to understand the main functions of SPSS and to navigate the menus. You will see how to investigate, manipulate, code and transform data. The statistical chapters will explain how to use SPSS to perform analyses and interpret the outcome.

1.2 Nuts and bolts

I don't have SPSS! Is that a problem?



One of the central features of this book is the way in which it will guide you through using SPSS. The web page resources for this book include SPSS data sets for all of the worked examples and learning exercises. If you are a psychology student at university, it is quite likely that you will have access to the latest version of SPSS during the course of your studies. The licence is renewed each year, so once you leave, the program may stop working. If that happens, you may feel a little stuck. Alternatively, you may not have access to SPSS at all. Either way, it is extraordinarily expensive to buy a single-user copy of SPSS. To address that, all of the data sets are also provided in spreadsheet format, which can be opened in more commonly available programs such as Microsoft Excel.

Chapter 3 explores the concept of normal distribution. This describes how the scores are 'distributed' across a data set, and how that might influence the way in which you can examine those data. We will explore why that is important, and we will learn how to measure and report normal distribution. If the outcome data are not 'normally distributed' we may not be able to rely on them to represent findings. We will also see what we can do if there is a problem with normal distribution.

Chapter 4 examines three ways in which we can measure the impact of our results: statistical significance, effect size and power. We will not explore what those concepts mean here, as that would involve exposing you to factors that you have not learned yet. Most importantly, we will learn about how probability is used in statistics to express the likelihood that an observed outcome happened due to chance factors. We will discuss effect size and power briefly a little later in this chapter.

Chapter 5 provides an overview of experimental methods and guidance on how to choose the correct statistical test. We will learn how to understand and interpret the key factors that determine which procedure we can perform. Using that information, we will explore an overview of the statistical tests included in this book, so that we can put all of it into context.

1.3 Take a closer look

Icons



A common feature throughout the chapters in this book relates to the use of 'boxes'. This 'Take a closer look' box will be employed to explore aspects of what you have just learned in a little more detail, or will summarise the main points that have just been made.

Nuts and bolts

Within the chapter text, you should find all you need to know to perform a test. However, it is important that you also learn about conceptual issues. You can do the tests without knowing such things, but it is recommended that you read these 'Nuts and bolts' sections. The aim is to take you beyond the basic stuff and develop points a little further.

Calculating outcomes manually

In all of the statistical chapters, you will be shown how to run a test in SPSS. For most readers, this will be sufficient. However, some of you may want to see how the calculations are performed manually. In some cases your tutor will expect you to be able to do this. To account for those situations, most of the statistical tests performed in SPSS will also be run manually. These mathematical sections will be indicated by this calculator icon. While these sections are optional, I urge you to give them a go – you can learn so much more about a test by taking it apart with maths. Microsoft Excel spreadsheets are provided to help with this.

Statistical chapters (6–21)

Each of the statistical chapters presents the purpose of that procedure, the theory and rationale for the test, the assumptions made about its use, and the restrictions of using the measure. In many cases we will explore how to calculate the test manually, using mathematical examples. Before learning how to perform the test in SPSS, you will see how to set up the variables in the data set and how to enter the data. You will then be guided gently through data entry and analysis with a series of screenshots and clear instructions. You will learn about what the output means and how to interpret the statistics (often with use of colour to highlight the important bits). You will be shown how to report the outcome appropriately, including graphical displays and correct presentation of statistical terminology. You will also be able to read about some examples of how those tests have been reported in published studies, to give you a feel for their application in the real world (and sometimes how not to do it). Finally, you have the opportunity to practise running the tests for yourself with a series of extended learning exercises.

Statistical chapter features

The format of the statistical chapters has been standardised to help give you a better understanding of each test. Certain features will be common across the chapters.

Learning objectives

At the start of each chapter you will be given an overview of what you can expect to learn.

Research question

Throughout each chapter, a single research theme will be used to illustrate each statistical test. This will help maintain some consistency and you will get a better feel for what that procedure is intended to measure.

Theory and rationale

In order to use a test effectively, it is important that you understand why it is appropriate for the given context. You will learn about the theoretical assumptions about the test and the key factors that we need to address. Much of this will focus on the arguments we explore in Chapter 5, relating to the nature of the variables that you are exploring. Sometimes you will be shown how the test compares to other statistical procedures. This will help you put the current test into context, and will give you a better understanding of what it does differently to the others.

Assumptions and restrictions

Related to the last section, each test will come with a set of assumptions that determines when it can be legitimately used. Often this will relate to factors that we explore in Chapters 3 and 5 regarding whether the data are normally distributed and the nature of the data being measured. We will explore the importance of those assumptions and what to do if they are violated.

Performing manual calculations

Although a main feature of this book focuses on the use of SPSS, wherever possible there will be instructions about how to calculate the outcomes manually. There are several reasons for doing this. As we saw earlier, witnessing how to explore the outcome using maths and formulae can reinforce our understanding of the analyses. Also, some of you simply may not have SPSS. Most of these calculations are provided just prior to the SPSS instructions. However, some are a little more complex, so they are safely tucked away at the end of the chapter to protect the

faint-hearted, or those of a more nervous disposition. Where appropriate, those calculations are supported by a Microsoft Excel spreadsheet that is provided on the web page for this book. These could also be used as a template to analyse other Excel-formatted data sets (such as those provided for the learning exercises). In some cases, those data can also be used to perform the complete statistical test in Microsoft Excel.

Creating the SPSS data set

Many statistical books show you how to perform a test in SPSS; this book is quite unique in the way that it shows you how to set up the data set in the first place. Data analysis can be so much easier if we create data sets that are appropriate for the type of analysis that we need to conduct. Using procedures that we learned in Chapter 2, we will explore the best way to create a data set, suitable for your analysis.

Conducting tests in SPSS

Each statistical chapter includes full instructions about how to perform the test using SPSS. These include easy-to-follow boxes that will guide you on how to undertake each stage of the statistical analyses. An example is shown in Figure 1.1.

Open the SPSS data set **Sleep 2**

Select **Analyze** → **Compare Means** → **Independent-Samples T Test...** (in new window) transfer **Sleep Quality** to **Test Variable List** → transfer **HADS cut-off depression** to **Grouping Variable** → click **Define Groups** button → (in the window) enter **1** in box for **Group 1** → enter **2** in box for **Group 2** → click **Continue** → click **OK**

Figure 1.1 An example of SPSS procedure instructions

You will also be shown screenshots of the SPSS displays that you will encounter during the process. You can refer to these to ensure that you are using the recommended method. An example of this is shown in Figure 1.2.

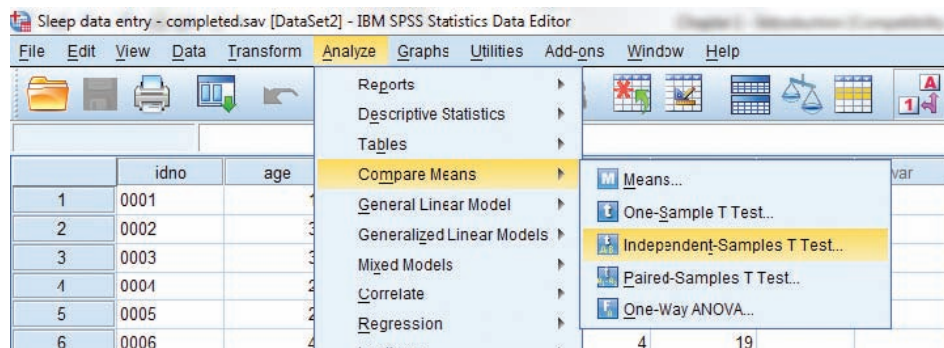


Figure 1.2 An example of SPSS screenshot

Interpretation of output

Once each test has been run, you will be taken through the SPSS output more thoroughly, so that you understand what each table of results shows and what the implications are. In some

cases, this output is relatively easy to follow – there may be just one line of data to read. In other cases, there may be several lines of data, some of which are not actually that important. Where there may be some doubt about what part of the output to read, colour and font will be used to illustrate where you should be focusing your attention. An example of this is shown in Figure 1.3.

Tests of between-subject effects

Dependent variable: HADS anxiety score

Source	Type III sum of squares	df	Mean square	F	Sig.	Partial eta squared
Corrected model	1102.107 ^a	5	220.421	16.688	.000	.476
Intercept	5193.219	1	5193.219	393.177	.000	.810
HADSDbase	336.365	2	168.183	12.733	.000	.217
hxinsom	74.793	1	74.793	5.663	.019	.058
HADSDbase * hxinsom	61.310	2	30.655	2.321	.104	.048
Error	1215.169	92	13.208			
Total	8799.000	98				
Corrected total	2317.276	97				

a. R squared = .476 (adjusted R squared = .447)

Figure 1.3 An example of annotated SPSS output

Effect size and power

In addition to reporting statistics, it is important that you state the effect size and power of the outcome. You will learn more about what that means in Chapter 4. Briefly, effect size represents the actual magnitude of an observed difference or relationship; power describes the probability that we will correctly find those effects.

Writing up results

Once you have performed the statistical analyses (and examined effect size and power where appropriate), you need to know how to write up these results. It is important that this is done in a standardised fashion. In most cases you will be expected to follow the guidelines dictated by the British Psychological Society (BPS) (although those rules will vary if you are presenting data in other subject areas). These sections will show you how to report the data using tables, graphs, statistical notation and appropriate wording.

Graphical presentation of data

You will be shown how to draw graphs using the functions available in SPSS, and you will learn when it is appropriate to use them. Drawing graphs with SPSS is much easier than it used to be (compared with earlier versions of the program). In many cases, you can simply drag the variables that you need to measure into a display window and manipulate the type of graph you need. In other cases, you will need to use the menu functions to draw the graphs.

Research example

To illustrate the test you have just examined, it might help to see how this has been applied in real-world research. At the end of each chapter you will find a summary of a recently published research article that uses the relevant statistical tests in its analyses. The papers focus on topics

that may well be related to your own research. While those overviews should be very useful in extending your understanding, you are encouraged to read the full version of that paper. For copyright reasons, we cannot simply give these to you. However, each paper is provided with a link that you can enter into an internet browser. In most cases this will be the '**DOI code**'. These initials stand for 'Digital Object Identifier'. It is an internationally recognised unique character string that locates electronic documents. Most published articles provide the DOI in the document description. Leading international professional bodies, such as the BPS, dictate that the DOI should be stated in reference lists. A typical DOI might be <http://dx.doi.org/10.1080/07420520601085925> (they all start with 'http://dx.doi.org/').

Once you enter the DOI into an internet browser, you are taken directly to the publisher's web page, where you will be given more details about the article, usually including the Abstract (a summary of that paper). If you want to access the full article you will have a series of choices. If you, or your educational institution/employer, have a subscription with that publisher you can download a PDF copy. If not, you can opt to buy a copy. Alternatively, you can give those details to your institutional librarian and ask them to get you a copy. Wherever possible, the DOI will be provided alongside the citation details for the summarised paper; when that is not available an alternative web link will be presented.

Extended learning task

To reinforce your learning, it is useful to undertake some exercises so that you can put this into practice. You will be asked to manipulate a data set according to the instructions you would have learned earlier in the chapter. You will find these extended learning examples at the end of each chapter (or in some cases within the chapter when there are several statistical tests examined). You will be able to check your answers on the web page for this book.

1.4 Take a closer look

Chapter layout



Each statistical chapter will follow a similar pattern, providing you with consistency throughout. This might help you get a better feel of what to expect each time. A typical running order is shown below:

- Learning objectives
- Research question
- Theory and rationale
- Assumptions and restrictions
- Performing manual calculations
- Setting up the data set in SPSS
- Conducting test in SPSS
- Interpretation of output
- Effect size and power
- Writing up results
- Presenting data graphically
- Chapter summary
- Research example
- Extended learning task

Online resources

A series of additional resources is provided on the web page for this book, which you can access at **www.pearsoned.co.uk/mayers**. These resources are designed to supplement and extend your learning. The following list provides a guide to what can expect to find there:

- Data sets:
 - to be used with worked examples and learning exercises
 - available in SPSS and Excel formats.
- Multiple-choice revision tests.
- Answers to all learning exercises.
- Excel spreadsheets for manual calculations of statistical analyses.
- Supplementary guides to SPSS (tasks not covered in the book).
- More extensive versions of distribution tables.

2

SPSS – THE BASICS



Learning objectives

By the end of this chapter you should be able to:

- Understand the way in which data and variables can be viewed in SPSS
- Recognise how to define variables and set parameters
- Enter data into SPSS and navigate menus
 - How to use them to enhance, manipulate and alter data
 - How to transform, recode, weight and select data
- Understand basic concepts regarding syntax

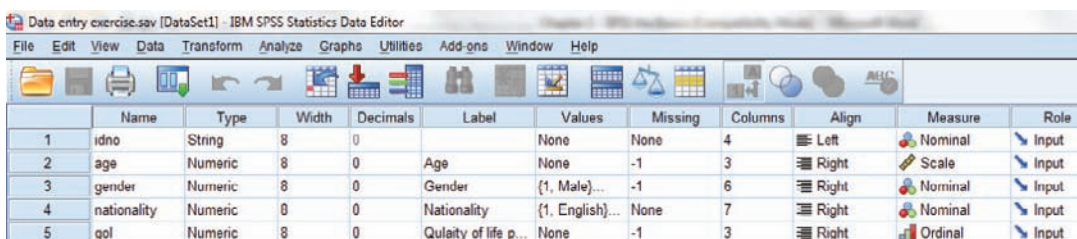
Introduction

SPSS® is one of the most powerful statistical programs available, and probably the most popular. Originally called the ‘Statistical Package for the Social Sciences’, SPSS has evolved to be much more than a program for social scientists, but the acronym remains. Many published studies, in a very wide variety of research fields, include statistics produced with SPSS. To the uninitiated, the program appears daunting and is associated with the horrors of maths and statistics. However, it need not be that scary; SPSS can be easy to learn and manipulate. Most of the tasks are available at the press of a button, and it is a far cry from the days when even the most basic function had to be activated by using programming code. The trick is learning what button to press. Many books report on how to use the functions, but very few provide even the most basic understanding. Some of you may be experienced enough not to need this chapter, in which case, you can happily pass on to the next chapter. However, even if you have been using SPSS for several years, you may benefit from learning about some of the newer functions now available.

We will start by looking at some of the most basic functions of SPSS, such as how to set up new data sets and how to use the main menus. To create a data set, we need to define **variables** – we will learn how to set the parameters according to the type of test we need to perform. We will see that there are two ways that we can view a data set: a ‘**variable view**’, where we define those variable parameters and a ‘**data view**’, where we enter data and manipulate them. Once we have created a data set, it would be useful if we learned how to use important menu functions such as ‘Save’ and ‘Edit’. Then we will proceed to some slightly more advanced stuff. Now, it’s quite likely that some bits about data editing and manipulation will be beyond you at this stage, particularly if you are new to statistics. If that happens, don’t worry. This chapter is not designed to be read in one go; you can return to it again later when you have learned more about statistical analyses themselves. The rationale for this approach is a simple one: it keeps all of the instructions for performing the main functions in one place. In many cases we will revisit the procedures in later chapters, when they become appropriate. However, it is useful to have the most basic instructions all together. We will not explore the data analysis and graphical functions in this chapter, as it is better that we see how to do that within the relevant statistical chapters. But we will (briefly) consider how SPSS uses ‘**syntax**’ language to perform tasks. You will rarely have to use this programming language, but it may be useful for you to see what it is used for. Throughout this book we will be using SPSS for Windows version 19.

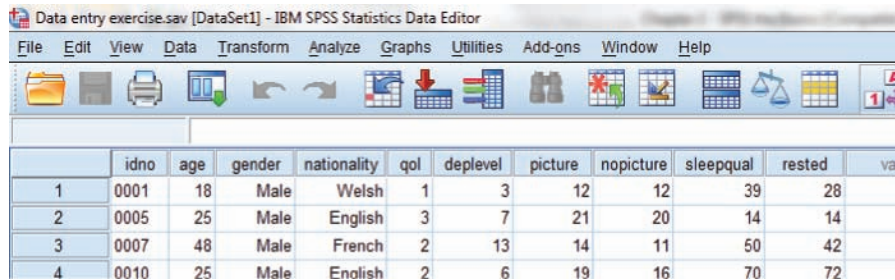
Viewing options in SPSS

One of the first things to note is that there are two editing screens for SPSS (called ‘**Data Editors**’): ‘Variable View’ and ‘Data View’. Variable View is used to set up the data set parameters (such as variable names, type, labels and constraints). Data View is used to enter and manipulate actual data. An example of each is shown in Figure 2.1 and Figure 2.2. Before you enter any data, you should set up the parameters and limits that define the variables (in Variable View). Once you have those variables set up, you can proceed to enter the data; you will do that via Data View.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	idno	String	8	0		None	None	4	Left	Nominal	Input
2	age	Numeric	8	0	Age	None	-1	3	Right	Scale	Input
3	gender	Numeric	8	0	Gender	{1, Male}...	-1	6	Right	Nominal	Input
4	nationality	Numeric	8	0	Nationality	{1, English}...	None	7	Right	Nominal	Input
5	qol	Numeric	8	0	Quality of life p...	None	-1	3	Right	Ordinal	Input

Figure 2.1 SPSS Variable View



	idno	age	gender	nationality	qol	deplevel	picture	nopicture	sleepqual	rested	va
1	0001	18	Male	Welsh	1	3	12	12	39	28	
2	0005	25	Male	English	3	7	21	20	14	14	
3	0007	48	Male	French	2	13	14	11	50	42	
4	0010	25	Male	English	2	6	19	16	70	72	

Figure 2.2 SPSS Data View

Variable View is arranged in columns that relate to the parameters that we will set for each variable. Each row relates to a single variable in the data set.

Data View is arranged in columns that show each of the variables included in the data set (these are the same as the rows in Variable View). Each row represents a single participant or case. Now we should see how we define and enter the information, so that we get the information that is displayed in Figures 2.1 and 2.2.

Defining variable parameters

It might help you understand the functions of SPSS by defining some variables and then entering data. To help us, we are going to use a small data set that will examine participants' age, gender, nationality, perceived quality of life and current level of depression. We will also examine how many words the group can recall (with or without a picture prompt). Finally, we will record the participants' perceptions of sleep quality and how well rested they felt when they woke up that morning.

Starting up a new SPSS data file

Before we start, we need to open a new (blank) SPSS data file. When SPSS is open for the first time, you may be presented with a range of screens. The default view (shown in Figure 2.3) requests options of how to proceed:

Open SPSS 19 from your program menu, or click on the SPSS icon.



In this case, we do not need any of those options, so just click on Cancel; a blank window will open (similar to Figures 2.3 or 2.4). On other occasions, you may wish to perform one of the other functions, but we will look at that later. In Figure 2.3, you will notice that there is a tick-box option saying 'Don't show this dialog in the future'. If that has previously been selected, you will not see Figure 2.3 at all; the program will just open straight into a blank window. Once you have opened a new data file, click on the Variable View button (at the bottom of the page). An example of a brand new Variable View page is shown in Figure 2.4.

Defining variable parameters: rules and limits

When we open Variable View we see a range of parameter descriptions across the column headings. Before we define those we should explore what each of the descriptors means:

Name: Give your variable a name that is relevant to what it measures, but try to keep it short. The limit for SPSS 19.0 is 64 characters, but it is advisable to make it more manageable (you can always provide a fuller description in the 'Label' column). The name should start with a letter; subsequent characters can be any combination of letters, numbers and almost any other character. There are some exceptions, and you will get an error message should you select any of those. You cannot use blanks: 'age of participant' is not acceptable, but 'age_of_participant' is fine. This field is not case sensitive.

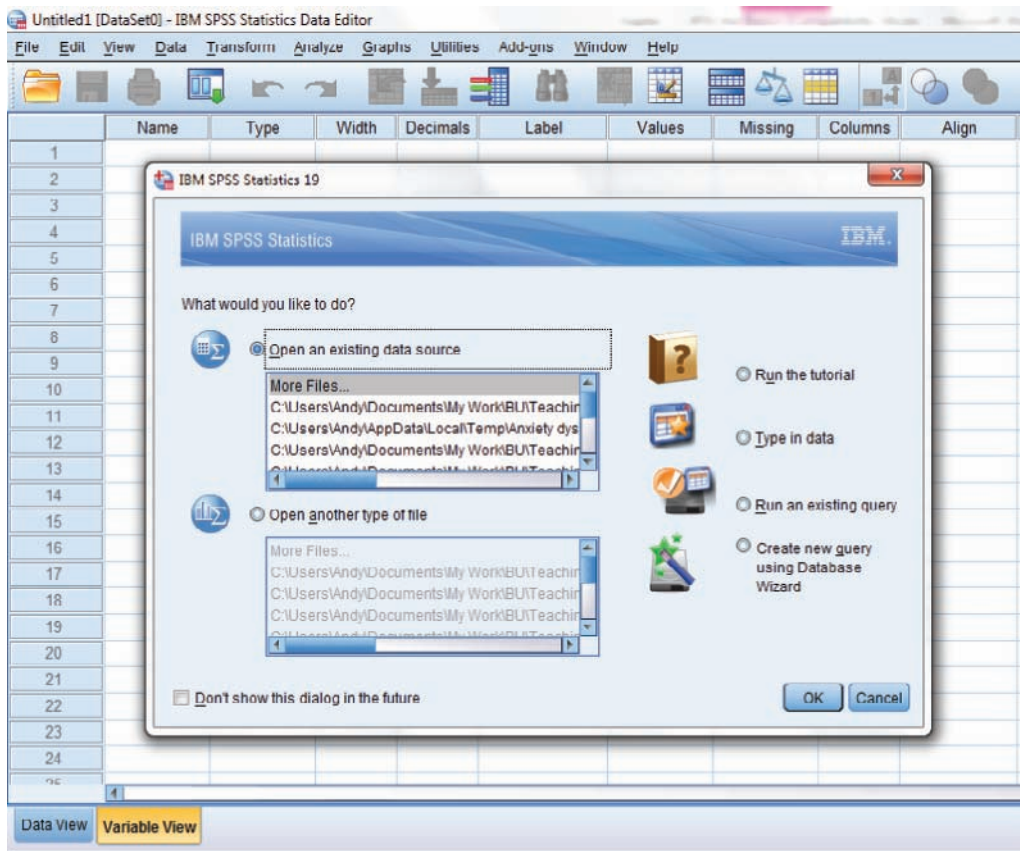


Figure 2.3 SPSS opening view

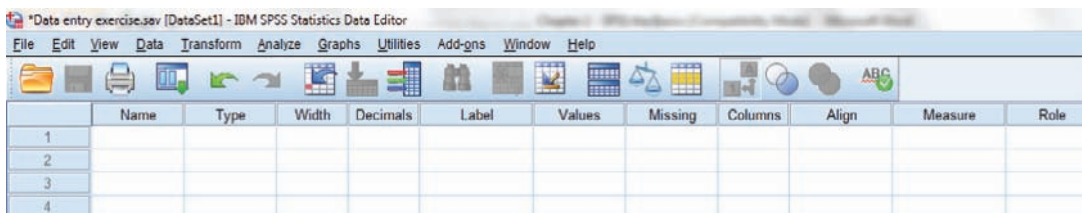
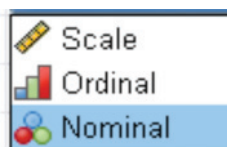


Figure 2.4 Blank Variable View

Type: If you click on the cell for this parameter you will be presented with a row of dots (. . .). Click on that and you will see a list of options (see Figure 2.4). The default is 'Numeric', which you will use most often. The most likely alternative is 'String', which you could use for participant identification. 'Numeric' can be used even when the variable is categorical, such as gender. This is because 'numbers' can be allocated to represent the groups of the variable (see 'Values').

Width: It is unlikely that you will need to change the default on this, unless you expect to require more digits than the default (eight characters). You may need to extend that if you want very large numbers, or if you need to display numbers with several decimal places (see below).

- Decimals:** Setting decimals applies only when using numeric data. You can use this to determine how many decimal places you show (in the data set). The default setting is for two decimal places. For something like age, you may want to change this to '0' (use the arrows to the right of the cell to make changes). For more specific data (such as reaction times) you may want any number of decimal places. This option has no effect on the number of decimal places shown in the results.
- Label:** This is where you can enter something more specific about the nature of the variable, so you can include a longer definition (and there are no limits). For instance, the 'Label' could be 'Depression scores at baseline', while the 'Name' parameter might be 'depbase'. Always put something here, as that label is shown in some parts of the SPSS [output](#).
- Values:** As we will see in later chapters, a categorical variable is one that measures groups (such as gender). So SPSS understands that we are dealing with categorical variables, we need to allocate 'numbers' to represent those groups. For example, we cannot expect SPSS to differentiate between the words 'male' and 'female', but can use the values facility to indicate that '1' represents male and '2' is female. If there are no groups, you would leave the Values cell as 'None' (the default). If you do have groups, you must set these values (you will see how later).
- Missing:** It is always worth considering how you will handle missing data. If there is a response absent from one of your variables, SPSS will count that empty cell as '0'. This will provide a false outcome. For example, the mean (average) score is based on the sum of scores divided by the number of scores. If one of those scores is incorrectly counted as 0, the mean score will be inaccurate. You should include '0' only if it actually represents a zero score. If the data are missing, you can define a specific 'missing variable value'. This will instruct SPSS to skip that cell (a mean score will be based on the remaining values). The missing 'value' indicator must be sensible; it must not be in the range of numbers you might be expecting (otherwise a real number might be ignored). The same applies to numbers used to define groups. A good choice for missing values is - 1: it should cover most scenarios. We will see how to do this later.
- Columns:** This facility determines the width of the column reserved for that variable in the Data View. So long as you can see the full range of digits in the cell, it does not really matter. Set this to be your preference.
- Align:** Data can appear to the left of a cell, the middle, or to the right – the choice is yours.
- Measure:** You need to define what type of variable you are measuring. Click on the arrow ▼ in the Measure cell. The options for Numeric data are [Scale](#), [Ordinal](#) or [Nominal](#). For String the options are Ordinal or Nominal. Select the appropriate one from the pull-down list.



The Scale measure is ruler – representing a range of scores.

The Ordinal measure is step – representing an order of groups.

The Nominal measure is distinct circles – representing categories.

With numeric data, 'Scale' refers to scores such as age, income or numbers that represent ranges and magnitude. These numbers are what we would normally categorise as interval or ratio data. 'Ordinal' data are also 'numerical' but only in the sense that the number represents a range of abstract groups; you will typically find ordinal data in attitude scale (where 1 = strongly agree, through to 5 = strongly disagree). You will learn more about interval, ratio and ordinal data in Chapter 5, so don't worry if that's all a bit confusing right now. 'Nominal' refers to distinct categories such as gender (male or female).

Role: Just use 'Input' for now; you can learn about the rest another time.

Creating new variable parameters

At this stage it would be useful to set up an example set of variables. You will recall that we are creating a data set that examines the participants' age, gender (male or female), nationality (English, Welsh or French), perceived quality of life, current level of depression, how many words they can recall (with and without a picture prompt), perceived sleep quality and how rested the participants felt when they woke up. We will also have a variable called 'participant identifier' (the usefulness of that will become apparent later). Table 2.1 shows the information we are about to enter into our new SPSS data set.

Table 2.1 Data set

SPSS variable									
idno	age	gender	Nationality	qol	deplevel	picture	nopicture	sleepqual	rested
0001	18	Male	Welsh	1	3	12	12	39	28
0002	38	Female	English	4	18	21	20	14	14
0003	30	Female	French	4	?	14	11	50	42
0004	22	Female	English	5	20	19	16	70	72
0005	25	Male	French	3	7	12	12	63	62
0006	40	Female	Welsh	4	19	11	11	39	39
0007	48	Male	English	2	13	21	22	59	39
0008	35	Female	Welsh	5	20	24	20	55	54
0009	45	Female	Welsh	3	10	17	21	39	42
0010	25	Male	English	2	6	18	12	57	60
0011	50	Male	French	5	24	18	11	59	57
0012	35	Male	English	2	11	18	9	74	78
0013	?	Female	Welsh	4	28	14	11	17	27
0014	32	Female	English	5	25	19	14	24	24
0015	40	Male	Welsh	1	12	23	18	50	47
0016	53	Female	Welsh	4	23	15	15	57	61
0017	35	Male	French	3	16	21	12	57	46
0018	30	Male	English	2	13	24	19	61	58
0019	20	Female	French	4	16	17	14	31	24

2.1 Nuts and bolts

SPSS instruction boxes



We will be using instruction boxes throughout this book to show how we perform a function in SPSS. To maintain consistency, fonts will be employed to indicate a specific part of the process:

Black bold: this represents a command or menu options shown within the data window.

Green bold: this indicates the item to select from a list within the menu or variable.

Blue bold: this refers to words and/or numbers that you need to type into a field.

We will now set up the parameters for the variables in this data set. Remember we need a new row for each variable. Go to the blank Variable View window for the new data set.

Participant identifier (Row 1):

We will start with a 'variable' that simply states the participant's identification number. This can be useful for cross-referencing manual files.

In **Name** type **idno** → in **Type** click on the dots... (you will be presented with a new window as shown in Figure 2.5) → select **String** radio button → everything else in this row can remain as default

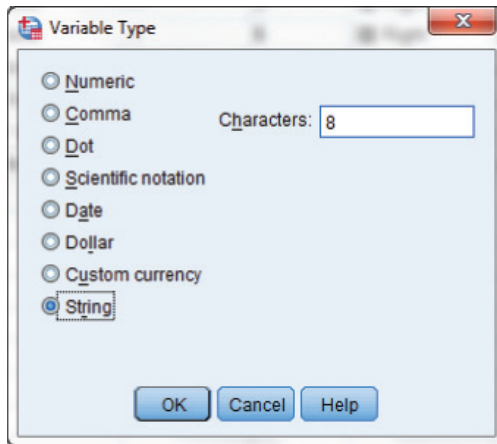


Figure 2.5 Setting type

Age (Row 2):

In **Name** type **age** → set **Type** to **Numeric** → ignore **Width** → change **Decimals** to **0** → in **Label** type **Age** → ignore **Values**

To set the parameter for **Missing** values, click on that cell and then the dots ... (you will be presented with a new window, as shown in Figure 2.6) → select **Discrete missing values** radio button → type **-1** in first box → click **OK** → back in original window, ignore **Columns** → ignore **Align** → click **Measure** → click arrow ▼ → select **Scale**

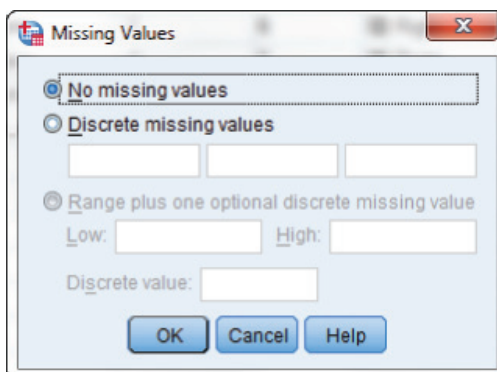


Figure 2.6 Missing values

Gender (Row 3):

In Name type **gender** → set Type to **Numeric** → ignore Width → change Decimals to 0 → in Label type **Gender**

Gender is a 'group' (categorical) variable, so we have to set some **Values** → click on that cell and then the dots ... (you will be presented with a new window, as shown in Figure 2.7) → in Value type **1** → in Label type **Male** → click **Add** → in Value type **2** → in Label type **Female** → click **Add** → click **OK** → back in original window, set **Missing** to **-1** → ignore **Columns** → ignore **Align** → set **Measure** to **Nominal**



Figure 2.7 SPSS value labels

Nationality (Row 4):

In Name type **nationality** → set Type to **Numeric** → ignore Width → change Decimals to 0 → in Label type **Nationality** → set Values as **1 = English**, **2 = Welsh**, and **3 = French** respectively (you saw how just now) → set **Missing** to **-1** → ignore **Columns** → ignore **Align** → set **Measure** to **Nominal**

Quality of life perception (Row 5):

In Name type **qol** → set Type to **Numeric** → ignore Width → change Decimals to 0 → in Label type **Quality of life perception** → ignore Values → set **Missing** to **-1** → ignore **Columns** → ignore **Align** → set **Measure** to **Ordinal**

Current level of depression: (Row 6):

In Name type **deplevel** → set Type to **Numeric** → ignore Width → change Decimals to 0 → in Label type **Current level of depression** → ignore Values → set **Missing** to **-1** → ignore **Columns** → ignore **Align** → set **Measure** to **Scale**

Picture: (Row 7):

In Name type **picture** → set Type to **Numeric** → ignore Width → change Decimals to **0** → in Label type **Words recalled with picture** → ignore Values → set Missing to **–1** → ignore Columns → ignore Align → set Measure to **Scale**

No picture: (Row 8):

In Name type **nopicture** → set Type to **Numeric** → ignore Width → change Decimals to **0** → in Label type **Words recalled without picture** → ignore Values → set Missing to **–1** → ignore Columns → ignore Align → set Measure to **Scale**

Sleep quality: (Row 9):

In Name type **sleepqual** → set Type to **Numeric** → ignore Width → change Decimals to **0** → in Label type **Sleep quality** → ignore Values → set Missing to **–1** → ignore Columns → ignore Align → set Measure to **Scale**

Rested: (Row 10):

In Name type **rested** → set Type to **Numeric** → ignore Width → change Decimals to **0** → in Label type **Rested on waking** → ignore Values → set Missing to **–1** → ignore Columns → ignore Align → set Measure to **Scale**

Entering data

To start entering data, click on the Data View tab and you will be presented with a window similar to the one in Figure 2.8. Remember, each row in Data View will represent a single participant.

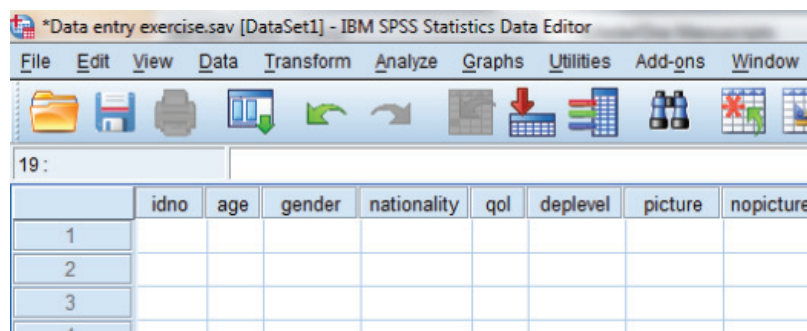


Figure 2.8 Blank Data View

To enter the data, we will use the information from Table 2.1. To get some practice you should enter these data, following the instructions shown below (note that there are some data 'missing').

Using the SPSS data set that we have just created, enter the following information:

Row 1: In **idno** type **0001** → in **age** type **18** → in **gender** type **1** → in **nationality** type **2** → in **qol** type **1** → in **deplevel** type **3** → in **picture** type **12** → in **nopicture** type **12** → in **sleepqual** type **39** → in **rested** type **28**

Row 2: In **idno** type **0002** → in **age** type **38** → in **gender** type **2** → in **nationality** type **1** → in **qol** type **4** → in **deplevel** type **18** → in **picture** type **21** → in **nopicture** type **20** → in **sleepqual** type **14** → in **rested** type **14**

Row 3: In **idno** type **0003** → in **age** type **30** → in **gender** type **2** → in **nationality** type **3** → in **qol** type **4** → in **deplevel** type **-1** (the 'depression score' is missing; so we enter the 'missing value' indicator instead) → in **picture** type **14** → in **nopicture** type **11** → in **sleepqual** type **50** → in **rested** type **42** ... and so on

Perhaps you would like to enter the remaining data (from Table 2.1); there will some further exercises at the end of this chapter.

SPSS menus (and icons)

Now we have created our first data set, we should explore how we use the 'menus' (refer to Figure 2.8 to see the range of menu headings). You will need to use only some of the functions found within these menus, so we will look at the most commonly used. In some cases, a menu function has an icon associated with it (located at the top of the view window). You can click on an icon to save time going through the menus; we look at the most useful of those icons (these are displayed below the menu headings, as shown in Figure 2.9). There are actually many more icons that could be included. You can add and remove the icons that are displayed, but we will not look at how to do that in this section. You can see how to do this in the supplementary facilities supplied in the web features associated with this chapter. The menu structure is the same in Data View and Variable View screens.

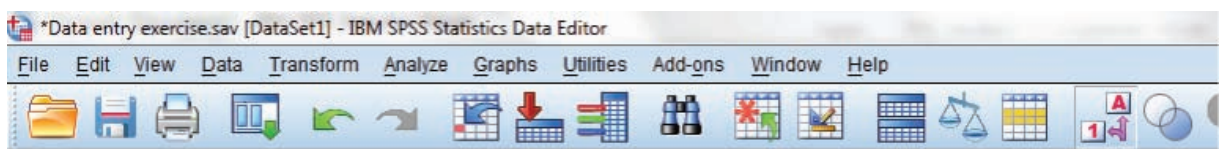


Figure 2.9 SPSS menus and icons

File menu

2.2 Nuts and bolts SPSS files



SPSS uses two main file types: one for data sets (these are illustrated by files that have the extension '.sav') and one for saving the output (the tables of outcome that report the result of a procedure) – these are indicated by files that have the extension '.spv'. A file extension is the letters you see after the final dot in a filename. It determines which program will open the file, and what type of file it is within that program. For example, word-processed files often have the file extension '.doc'. There are other file types in SPSS, such as those used for the syntax programming language. However, most of the time you will use only .sav and .spv files.

When the 'File' menu is selected, a series of options will appear (see Figure 2.10). The file menu is pretty much the same as you will find in most popular software programs, with some exceptions. There are several functions available here. Some of these are more advanced than we need, so we will focus on those that you are most likely to use for now.

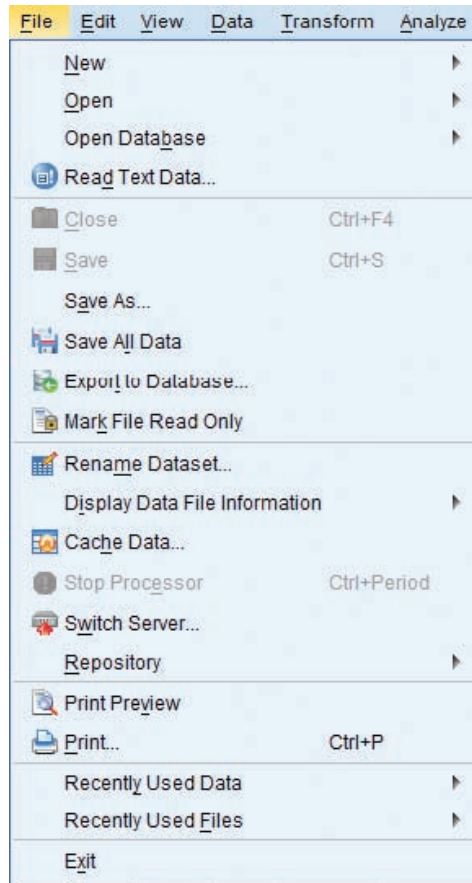


Figure 2.10 File menu options

New: Use this to start a new data file. It is most likely that this will be a new data set, in which case you would follow the route: (click on) **File → New → Data**. However, you might equally choose to start a new **Syntax** or **Output** file.

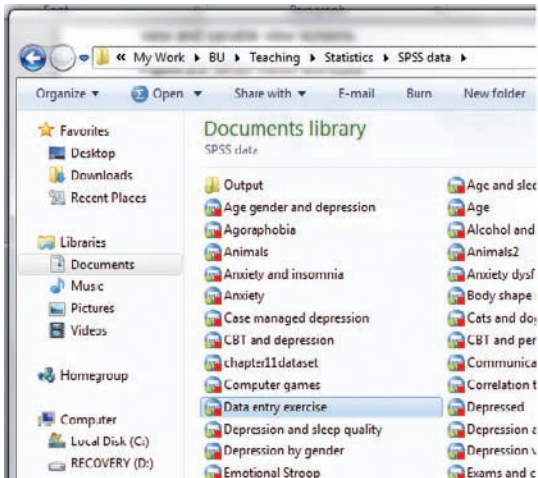
Open: Use this to open an existing file, perhaps one that you have worked on previously. If you want to open a data file, perform **File → Open → Data**. To open a saved output file, perform **File → Open → Output**. There is an icon associated with this function, which you can use just by clicking on it (saving a little time from selecting the menu route):



You can also open a file by clicking on it directly from your own folders (see Figure 2.11).

Save: It is good practice to save data sets and output files frequently, not just when you have finished. If your computer crashes, you might lose everything. To save the file, select **File → Save** (regardless of whether you are saving a data set or an output file). Alternatively, you can click on the icon shown here. If the file has not been saved before, you will be asked to create a name and indicate where you want the file saved. If it is an existing file, it will save any new changes.





Double click on the required file and it will open in the SPSS program.

Figure 2.11 Opening a file from general folders

Save As:

If you make changes to a file but want to keep the original file, use this function to save the changed version to a different file. Select **File → Save As** (regardless of whether you are saving a data set or an output file). Do *not* use the 'file save' function: the details in the file prior to the changes will be overwritten.

Mark File Read Only:

You can protect your file from any further changes being made; new changes can be made to a new file using 'Save As'. Select **File → Mark File Read Only** (you will be reminded to save current unsaved changes).

Print Preview:



You may want to see what a printed copy of your file will look like, without actually printing it (for example, you may want to change margins to make it fit better) – this saves printing costs. Select **File → Print Preview**.

Print:



If you are happy to print the file, send this to a printer of your choice by selecting **File → Print**. You will be given a list of printers that this can be sent to. If you use the 'Print' icon the print will be automatically sent to your default printer.

Exit:

As the name implies, this closes down the file. You will be warned if data have not been saved. You also get a warning if the file is the last SPSS data set still open (it closes the whole program). You can also click on the cross in the top right-hand corner to close the file. Make sure you save before you close anything.

Recently Used Data:

This provides a similar function to 'Open' but will locate the most recently used data sets. This is often quicker because, using 'Open', you may need to trawl through several folders before you find the file you are after. However, this function remembers file names only. If you have moved the file to another folder since it was last used, you will get an error message. Select **File → Recently Used Data** and choose the file you want to open.

Recently Used Files:

This is the same as 'Recently Used Data' but it locates all other files that are not data sets (output files, for example).

Edit menu

The Edit menu also shares properties with other software programs that you may be more familiar with. When this menu is selected, a number of options are displayed (see Figure 2.12).

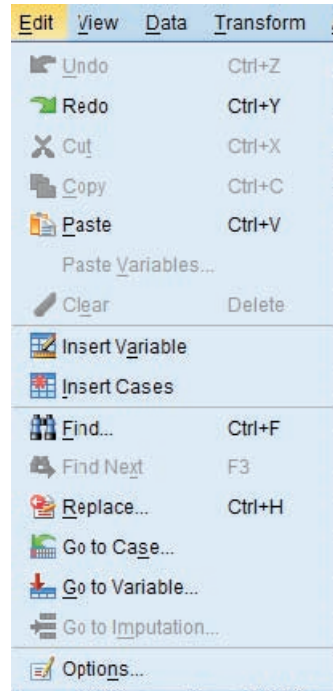


Figure 2.12 Edit menu options

We will explore some of the more common functions here. Where an icon is displayed, this can be selected instead of using the full menu function:

Undo:



Sometimes you may enter data incorrectly, or make some other error that you want to 'undo'. Use this function to do that by selecting **Edit → Undo**.

Redo:



Having undone what you believed to be incorrect, you may decide it was OK after all and want to put the information back in again. You can redo what was undone by selecting **Edit → Redo**.

Cut:



If you want to move information from a current cell and put it somewhere else, you need to use this 'Cut' facility. It's rather like deleting, but the information is saved in a memory cache until you find somewhere else to put it (see 'Paste'). To do this, select **Edit → Cut**.

Copy:



If you want to copy information from the current cell (to somewhere else) but also keep the current information where it is, you need this 'Copy' function. To do this, select **Edit → Copy**. You will need the 'Paste' function to complete the task.

Paste:



Use this to paste information that has been cut or copied from somewhere else into a cell by selecting **Edit → Paste**.

Insert Variable:



You can use this function to insert a new variable in Variable View. In many cases, we would simply start a new row (rather like we did earlier). However, sometimes you might decide to include a new variable but would like to

have it placed next to an existing one (perhaps because it measures something similar). To do this, go to Variable View and click on the row above which you want to insert the new variable. Then select **Edit → Insert Variable**. You would then need to set the parameters as you have been shown.

Insert Case:



You can use this function to insert a new 'case'. In most data sets, a case will be a participant. It is quite likely that it will not matter what order you enter data, but sometimes you may want to keep similar participants together (such as all of the depressed people in one place). In that scenario, you may want to insert a participant into a specific row of your data set. To do that, go to Data View and click on the row above which you want to insert the new case. Then select **Edit → Insert Case**. You can then enter the data for your new participant.

Find:



In larger datasets it can be time consuming to look for specific bits of data. For example, in a data set of 1,000 people you may want to find cases where you have (perhaps mistakenly) used '99' to indicate a missing variable. You can select **Edit → Find** to locate the first example of 99 in your data set. Once you have found the first example, you can use the 'Next' button to locate subsequent examples.

Replace:



Having found the items you are looking for, you may wish to replace them. For example, you have originally chosen to use 99 as your missing value indicator for all variables, including age. Later, you discover that one of your participants is aged 99! If you kept 99 as the missing variable it would not count that person. So you decide to change the missing value indicator to - 1. If there were 50 missing values in all variables across the data set, it would take some time to change them and you might miss some. However, the 'Replace' function will do that for you all at once. Go to **Edit → Replace →** enter 99 in the **Find** box → enter - 1 in the **Replace with** box → click on **Replace All**. However, do be careful that there are not other (valid) cases of 99 - you might replace true data with an invalid missing value. If you are not sure, use the **Find Next** button instead of 'Replace All'.

Options:



This function enables you to change a whole series of default options, including the font display, how tables are presented, how output is displayed, and so on. Much of this is entirely optional and will reflect your own preferences.

View menu

The View menu offers fewer features than the others, but those that are there are very useful.

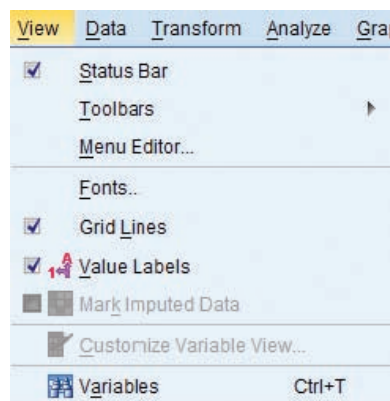


Figure 2.13 View menu options

Three functions can be selected via tick boxes:

- Status bar:** This function confirms current functions at the foot of the display window. This can be quite reassuring that the process is working, so it is a good idea to leave this ticked.
- Grid lines:** This function allows you to show grid lines between cells, or to remove them; it is entirely optional.
- Value Labels:** This is a very useful function. Earlier, we saw how to set up categorical variables that represent groups. For example, we created a Gender variable and used codes of 1 and 2 for 'male' or 'female' respectively. When we display the data set, we can choose whether to show the numbers (such as 1, 2) or the value labels (such as male, female) by ticking that box. Alternatively, you can click on the icon in the toolbar – if you are currently showing numbers it will switch to value labels, and vice versa.



Other functions are selected by clicking on that option and following additional menus:

- Toolbars:** You can use this function to choose which icons to include on the toolbar. Select **View → Toolbars → Customize** (a new window opens) → click **Edit**. From the operations window you can select a menu and choose which icons you can drag onto the toolbar.
- Fonts:** You can use this facility to change the way in which fonts are displayed in the data set. This is entirely your choice. Select **View → Fonts** if you want to change anything.

The next three menus are used to manipulate data. To fully illustrate these functions, we will undertake some of the procedures as well as explain what the menu aims to do.

Data menu

The data menu examines and arranges the data set so that specific information can be reported about those data. In some cases this has an impact on the way in which data are subsequently analysed. There are many functions in this menu, so we will focus on those that are probably most useful to you for the moment.

We can perform these functions on the data set that we created earlier. If you want to see the completed data set, you will find it in the online resources for this book. The file is called '[Data entry exercise](#)'.

- Define Variable Properties:** This function confirms how a variable has been set up and reports basic outcomes, such as the number of cases meeting a certain value. To perform this task, select **Data → Define Variable Properties**.



- Copy Data Properties:** This function enables you to copy the properties of one variable onto another by selecting **Data → Copy Data Properties**.



- Sort cases:** This useful facility allows you to 'sort' one of the columns in the data set in ascending or descending order. For example, using the data set we created, we could sort the 'Current level of depression' column from lowest score to highest score. To illustrate this important function, we will perform that task without data:



Using the SPSS data set **Data entry exercise**

Select **Data** → **Sort Cases** (see Figure 2.14) → (in new window) transfer **Current level of depression** to the **Sort by** window (by clicking on the arrow, or by dragging the variable to that window) → select radio button by **Ascending** → click **OK** (as shown in Figure 2.15).

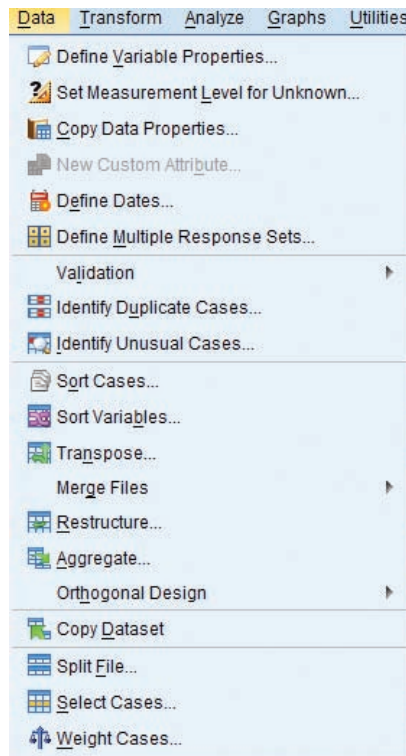


Figure 2.14 Data menu options

Return to the data set and you will notice the column for 'Current level of depression' is now in order, from the lowest to the highest.

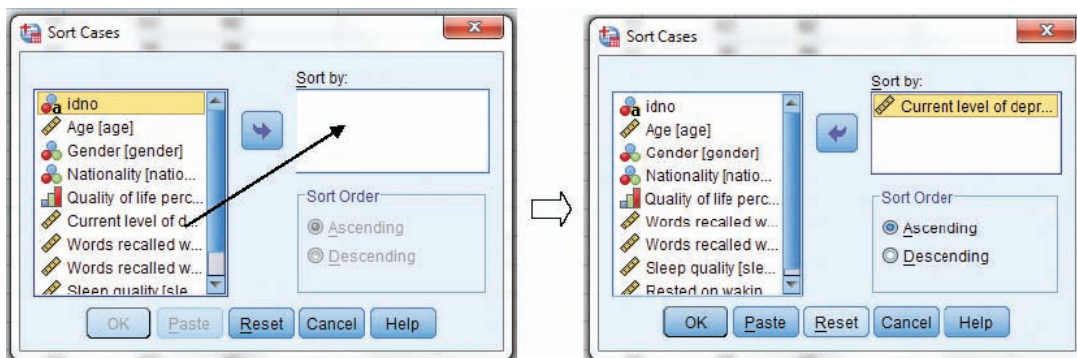


Figure 2.15 Sort cases function

Split File: This is another extremely useful facility. It enables you to split the data set according to one of the (categorical variable) groups. This can be used to report outcomes across remaining variables but separately in respect of those groups.



We will use this function in very important analyses later in the book, notably for multi-factorial ANOVAs (Chapters 11 and 13). However, we can illustrate this function with a simple example now. In the data set that we created, we have two variables that measure ‘word recall’. These measure how words can be recalled by the participants when they are given a picture prompt to aid recall (‘Words recalled with picture’) and when they are not (‘Words recalled without picture’). If we examine our entire sample across those two variables, we can compare the outcomes. We call that a within-group study (we will encounter these often throughout the book). We might find that people recall more when they are given the picture prompt. This is all very well, but we might also want to know whether that outcome differs according to gender. We can do this with the split file. Before we split the file, we should look at some basic outcome regarding the word recall across the group.

		Mean	N	Std. deviation	Std. error mean
Pair 1	Words recalled with picture	17.79	19	4.008	.920
	Words recalled without picture	14.74	19	4.067	.933

Figure 2.16 Mean number of words recalled in each condition

Figure 2.16 appears to show that more words are recalled when the group are given the picture prompt (mean [average] words remembered = 17.79) than when no picture is given (mean = 14.74). We should analyse that statistically, but we will leave all of that for later chapters, when you have learned more about such things. For now, let’s see what happens when we ‘split the file’ by gender:

Select **Data → Split File** (see Figure 2.14) → (in new window) click radio button for **Compare groups** → transfer **Gender** to **Groups Based on:** window → click **OK** (as shown in Figure 2.17). Choosing the ‘Compare Groups’ option here will result in output that directly compares the groups. This is probably better than selecting the ‘Organize output by groups’ option, which would produce separate reports for each group.

Now we can examine the difference in word recall across the picture conditions, now according to gender. We can see some fundamental differences between the groups on these outcomes. Figure 2.18 suggests that there is very little difference in mean words recalled between conditions for men, but women appear to recall far more words when prompted with the picture than with no picture.

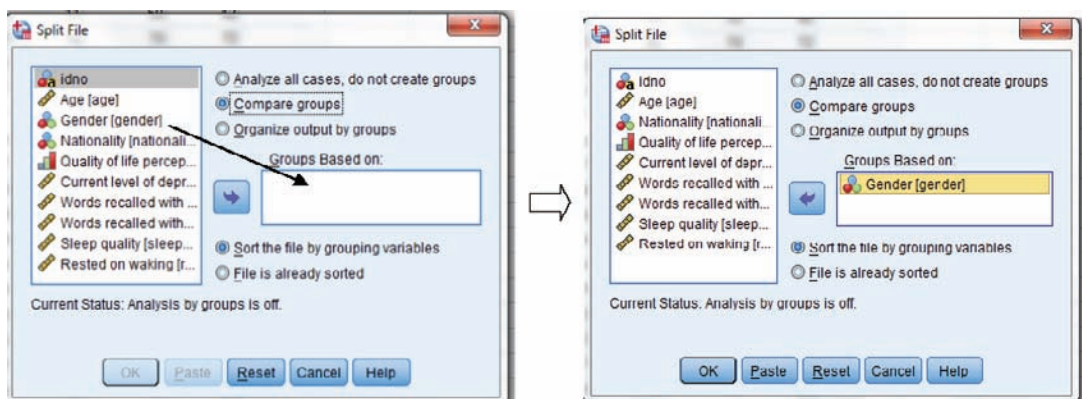


Figure 2.17 Split File function, step 2

Gender			Mean	N	Std. deviation	Std. error mean
Male	Pair 1	Words recalled with picture	16.78	9	4.738	1.579
		Words recalled without picture	16.11	9	4.676	1.559
Female	Pair 1	Words recalled with picture	18.70	10	3.199	1.012
		Words recalled without picture	13.50	10	3.171	1.003

Figure 2.18 Mean number of words recalled in each condition (by gender)

You must remember to return the data set to a state where there is no 'split' – otherwise all subsequent analyses will be affected.

Select **Data** → **Split File** → click radio button for **Analyze all cases, do not split groups** → click **OK**

Select Cases: This function allows you to explore certain sections of the data. In some respects it is similar to what we saw for the 'Split File' facility, but there are several more options. For example, you can exclude a single group from the data set and report outcomes on the remaining groups. In our data set, we could decide to analyse only English and Welsh participants, excluding French people. In effect, we 'switch off' the French participants from the data. This is how we do it:



Select **Data** → **Select Cases** (see Figure 2.14) → (in new window) select **If condition is satisfied** radio button → click on **If ...** box (as shown in Figure 2.19)

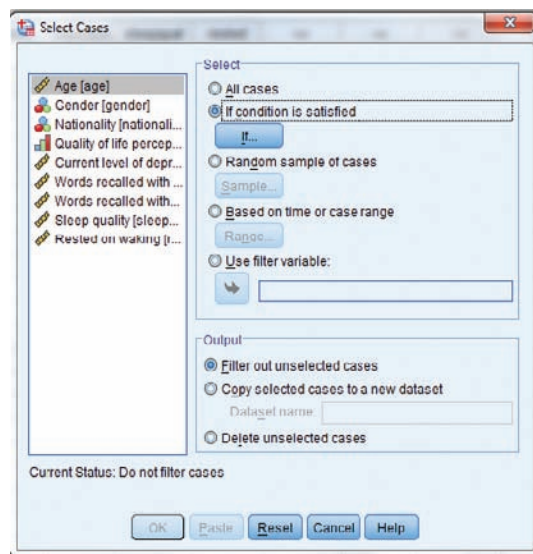


Figure 2.19 Select cases function, step 1

In that new window (see Figure 2.20), transfer **Nationality** to blank window to the right ('Nationality' will now appear in that window) → click on **~** (this means 'does not equal') → Type **3** (because 'Nationality = 3' represents French people (who we want to deselect)) → click **Continue** → click OK (see Figure 2.21 to see completed action)

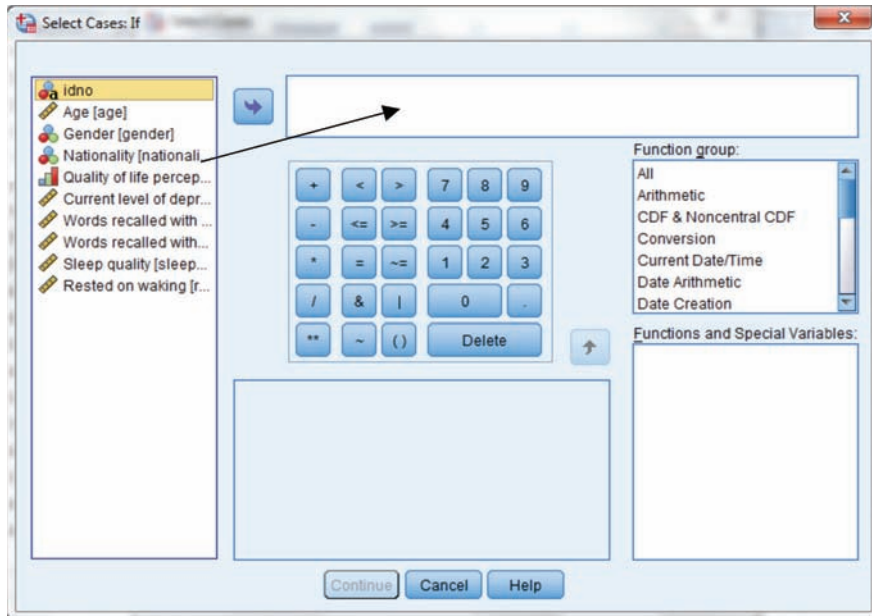


Figure 2.20 Select cases function, step 2

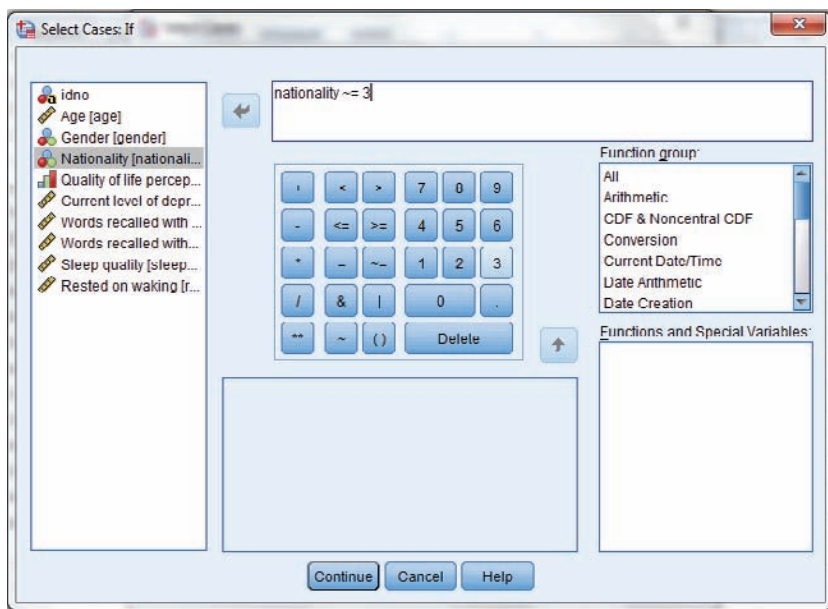


Figure 2.21 Select cases function, step 2 (completed)

When you return to Data View you will notice that all of the cases referring to French people are now crossed out. You would now be able to perform your analyses just based on English and Welsh people. Before you can use the data for other functions, you will need to remove the selected cases and return to the full data set:

Select **Data** → **Select Cases** → click **All cases** → click **OK**

Weight cases: This facility has a couple of useful functions. First, it can be used to count the number of cases that match a combination of scenarios. Or, second, we can 'control' a single variable in the data set so that the remaining variables are 'equal' in respect of that controlling variable. To illustrate how we can use this function to count cases we need a much larger data set. In this scenario, we have a sample of 200 people, for whom we measure two variables: gender (males/females) and whether they watch football on TV (yes/no). Now imagine how long it would take to enter data for 200 participants. Thankfully, there is a shortcut. We can count the number of times we find the combination of the following: males who watch football on TV, males who do not, females who do and females who do not. The data set might look something like Figure 2.22.



	gender	football	count	value
1	Male	Yes	31	
2	Male	No	19	
3	Female	Yes	12	
4	Female	No	38	
5				

Figure 2.22 SPSS data set: watching TV by gender

However, as it stands, the 'count' is simply another variable. To use it to count the number of cases that match the scenarios in the first two columns, we need to use the 'weight' function.

Open SPSS data set **Football**

Select **Data** → **Weight Cases** (see Figure 2.14) → select **Weight cases by** radio button → transfer **Count** to **Frequency Variable** window → click **OK** (see Figure 2.23)

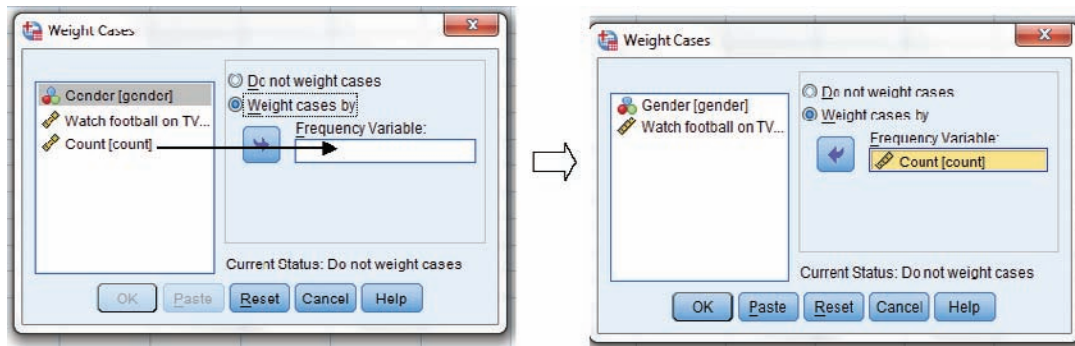


Figure 2.23 Weight Cases function

Now that the data are 'weighted' by count, analyses can be performed to explore how men and women differ in watching football.

We can also use the 'weight' function to 'normalise' data. In social science research (including psychology) it is difficult to control all of the variables. Using the data set that we created earlier, we might choose to explore 'current level of depression' by gender. We might find that women score more highly (poorly) on depression scores than men. However, what if we also notice that depression scores increase with age? How can we be sure that the observed outcome is not the result of age rather than gender? To be confident that we are measuring just depression scores by gender, we need to 'control' for age. By using the 'weight' function, we can adjust the depression scores so that everyone is equal in age. As we will see later in this book, there are more sophisticated tests that can do this (see ANCOVA, Chapter 15). However, the weight function provides one fairly easy way of exploring a simple outcome. This is how we do it:

Select **Data** → **Weight Cases** → select **Weight cases by** radio button → transfer **Age** to **Frequency Variable** window → click **OK**

Before you can use the data for other functions, you will need to remove the weighting function:

Select **Data** → **Weight Cases** → click on **Do not weight cases** → click **OK**

Transform menu

The transform menu undertakes a series of functions that can change the properties of variables, or create new variables based on the manipulation of existing variables. Once again, we will focus on the ones that you are most likely to use. To illustrate those important facilities, we will perform the functions using example data.

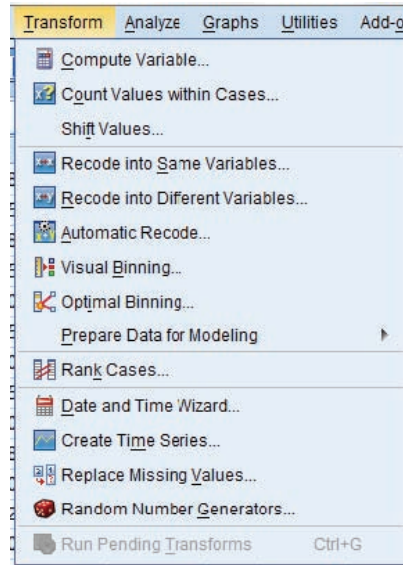


Figure 2.24 Transform menu options

Compute Variable: You can use this to perform calculations on your variables, perhaps to adjust them or create new variables. For example, you might have several variables that measure similar concepts, so you decide to create a new variable that is the sum of those added together. In the data set that we created earlier, we had one variable for 'Sleep quality' and one for 'Rested on waking'. We could combine those into a new variable called 'Sleep_perceptions'. Here's how we do that:



Using the SPSS data set **Data entry exercise**

Select **Transform** → **Compute variable** (see Figure 2.24) → (in new window, as shown in Figure 2.25), for **Target Variable** type **Sleepperceptions** → transfer **Sleep quality** to **Numeric Expression** window → click on + (the 'plus' sign shown in keypad section below the **Numeric Expression** window) → transfer **Rested on waking** to **Numeric Expression** window → click **OK** (see Figure 2.26 for completed action)

Go back to the data set. You will see that a new variable (sleepperceptions) has been included.

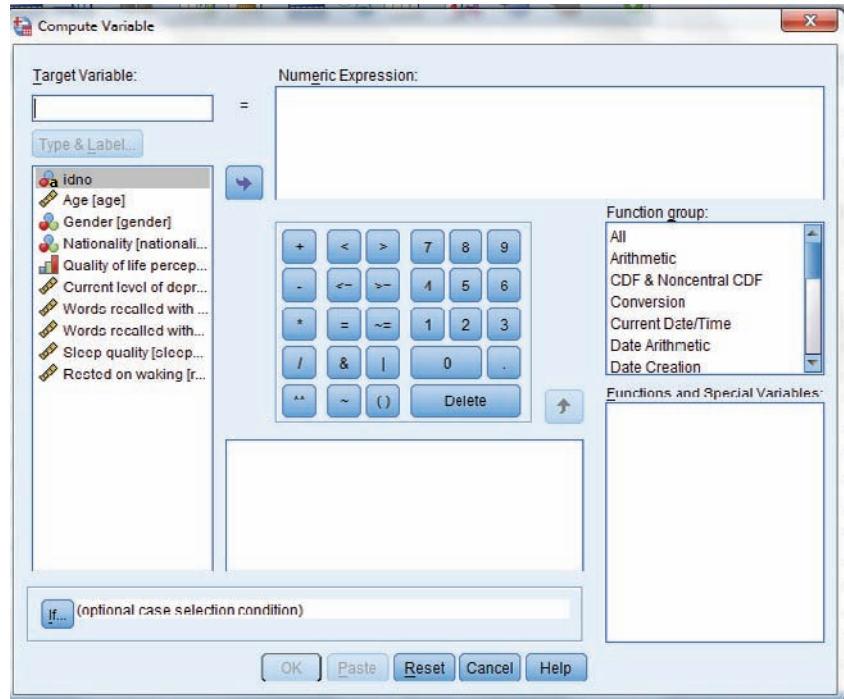


Figure 2.25 Transform Compute Variable

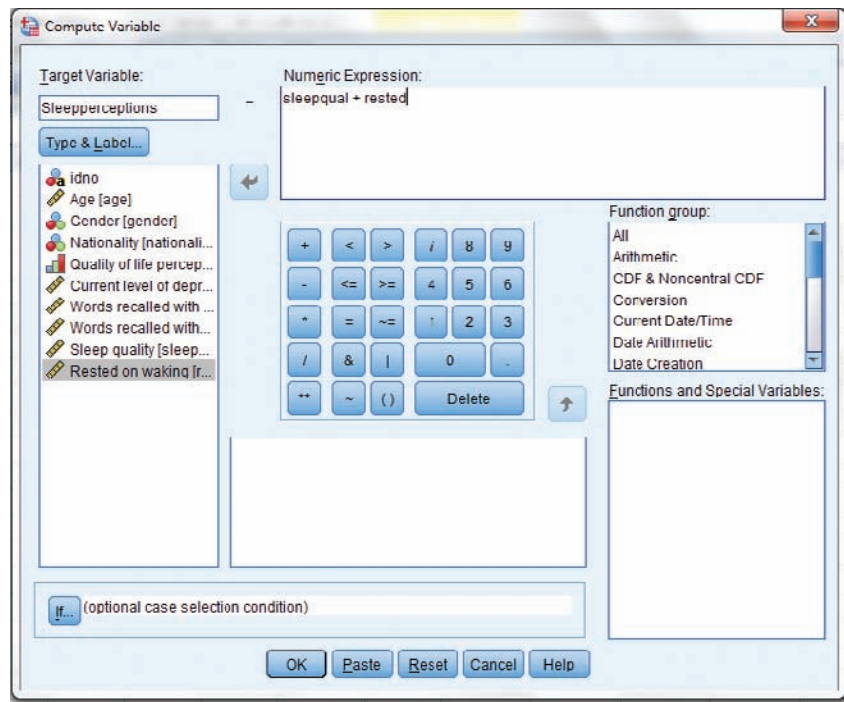


Figure 2.26 Transform Compute Variable (completed)

Recode into Same Variables: Sometimes you may need to recode the values of your variables. For example, when we created our data set, we input the values for gender as 1 (male) and 2 (female). However, as we will see in later chapters, some statistical procedures (such as linear



regression – Chapter 16) require that categorical variables can have only two groups and must be coded as 0 and 1 (don't worry about why for the moment). This is how we make those changes (this procedure will overwrite the values that we set up before):

Select **Transform** → **Recode into Same Variables** (see Figure 2.24) → in new window (as shown in Figure 2.27) transfer **Gender** to **Variables** window (which becomes renamed as 'NumericVariables') → click **Old and New Values...**

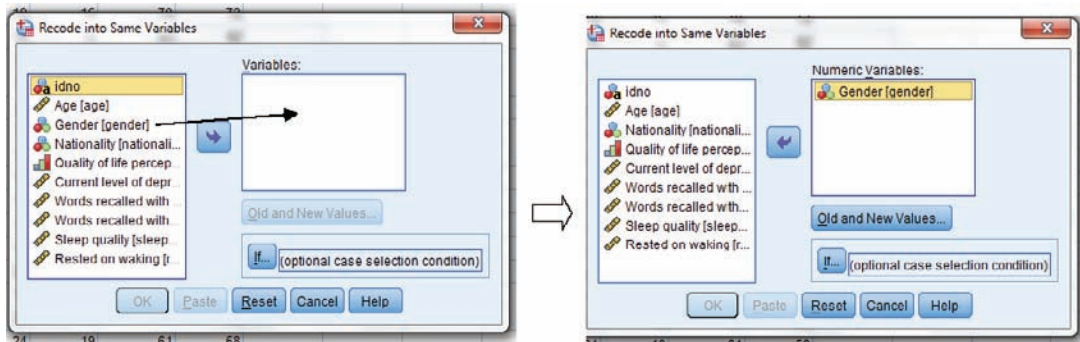


Figure 2.27 Recode into Same Variables function – step 1

In new window (as shown in Figure 2.28), under **Old Value**, select **Value** radio button → type **1** in box → under **New Value**, select **Value** radio button → type **0** in box → click **Add** (1 --> 0 appears in **Old --> New** box) → for **Old Value**, type **2** → for **New Value**, type **1** → click **Add** (2 --> 1 appears in **Old --> New** box) → click **Continue** → (in original window) click **OK**

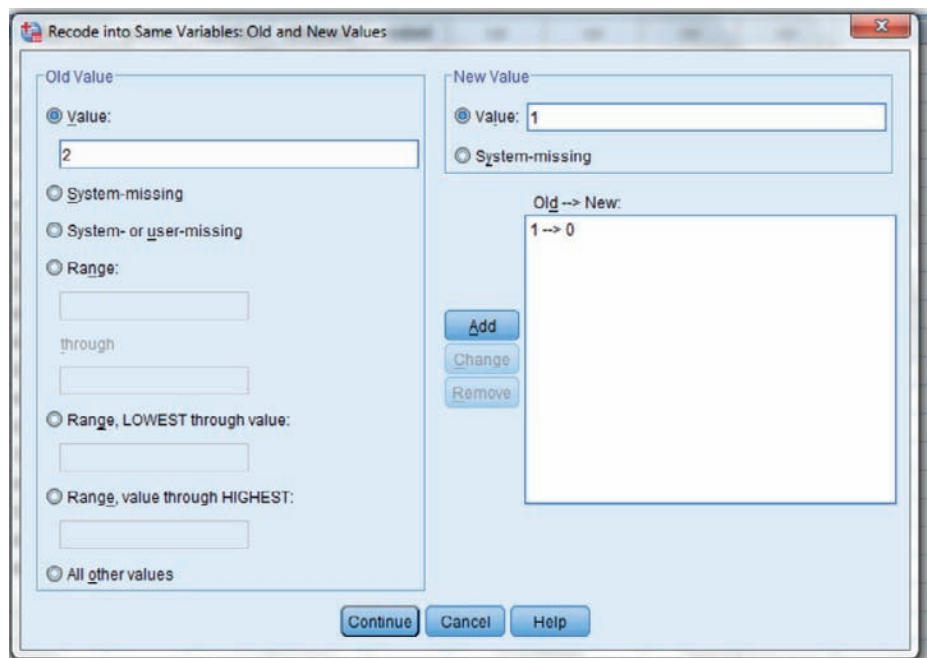
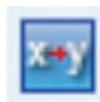


Figure 2.28 Recode into Same Variables function – step 2

If you look at the data set you will see that the gender data now show '0 and 1' where '1 and 2' used to be. But now, the variable is coded incorrectly. You must go to Variable View and change the value codes to show males = 0, females = 1.

Recode in Different Variables:



This is the same as what we have just seen, but a new variable is created rather than changing the existing one (it will not overwrite the original variable information).

Analyze menu

This menu contains the statistical techniques that we can use to analyse and manipulate data. We will be exploring how to analyse data in the statistical chapters later, so we do not need to look at this in too much detail here. This menu permits a wide range of statistical analyses, each with different rules of operation so we will leave that for now.

Direct Marketing menu

This menu is more likely to be useful for market researchers. According to SPSS, it 'provides a set of tools designed to improve the results of direct marketing campaigns by identifying demographic, purchasing, and other characteristics that define various groups of consumers and targeting specific groups to maximize positive response rates'.

Graphs menu

Once you have reported your results, you may want to represent the outcome graphically. This menu provides a wide range of graphs that can be used. However, we will explore that in more detail when we get to the statistical chapters.

Utilities menu

We will not dwell on this menu – the facilities are more likely to be attractive to advanced users. It offers further opportunities to view the properties of variables (how they are defined in the program, including the programming language parameters). Perhaps the most useful facility is one where you can change the output format so that it can be sent to another medium (such as Word, PDF, etc.). You can append comments to SPSS files, which may be useful if you are sharing data. Other facilities are much more advanced and might be useful only to those who understand the more technical aspects of programming (so, not me then).

Add-ons menu

This menu highlights a number of additional products that SPSS would like you to be aware of, such as supplementary programs or books about using SPSS. There is then a link to a website that invites you to buy these products. Enough said.

Window menu

This is simply a facility whereby you change the way in which the program windows are presented, such as splitting the screen to show several windows at once.

Help menu

This does exactly as it says on the tin: it helps you find stuff. You can search the index for help on a topic, access tutorials on how to run procedures, and scan contents of help files. This can be very useful even for the most experienced user.

Syntax

Syntax is the programming language that SPSS uses (mostly in the background). For the most part, you will not need to use this, as the functions are performed through menus and options. However, there are times when using syntax is actually much quicker than entering all of the required information using the main menus. We may need to run the same statistical test many times, particularly if we are collecting data on an ongoing basis. Running tests in SPSS can be relatively straightforward (such as an independent t-test), while others are rather more complex (such as a mixed multi-factorial ANOVA or a multiple regression). Using syntax can save a lot of time and energy in setting up the parameters for those tests. As you will see when you run each statistical test, the SPSS output includes a few lines of syntax code (just before the main outcome tables). If we want to run a subsequent test on this data set, we can cut and paste the code into the syntax operation field. The test will run without having to redefine the test parameters. There may also be some occasions when you will need to write some syntax to perform a task that is not available through the normal menus (see Chapter 11 for an example).

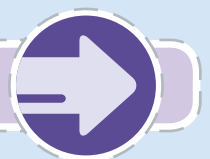
Chapter summary

In this chapter we have explored some of the basic functions of SPSS. At this point, it would be good to revisit the learning objectives that we set at the beginning of the chapter.

You should now be able to:

- Recognise that SPSS presents data sets in two 'views': the Variable View where variables are defined and parameters are set, and the Data View where the raw data are entered.
- Understand that there are a number of limits that we must observe when setting up those parameters.
- Appreciate the need to correctly define 'missing variables' so that blank spaces in the data set are not treated as '0'.
- Perform basic data entry in SPSS.
- Understand the purpose of the SPSS menus, and the function of the more popular sub-menus, including basic data manipulation and transformation.

Extended learning task



Following what we have learned about setting up variables, data input and data manipulation, perform the following exercises. Your task will be to create an SPSS data set that will explore outcome regarding mood, anxiety and body shape satisfaction in respect of gender and age group. The variable parameters are as follows:

Gender: male (1), female (2)

Age group: under 25 (1), 25–40 (2), 41–55 (3)

Outcome measures: anxiety, mood (measured on an interval scale), body shape satisfaction (measured on an ordinal scale)

We have some raw data in respect of eight participants, shown in Table 2.2.

Table 2.2 Raw data

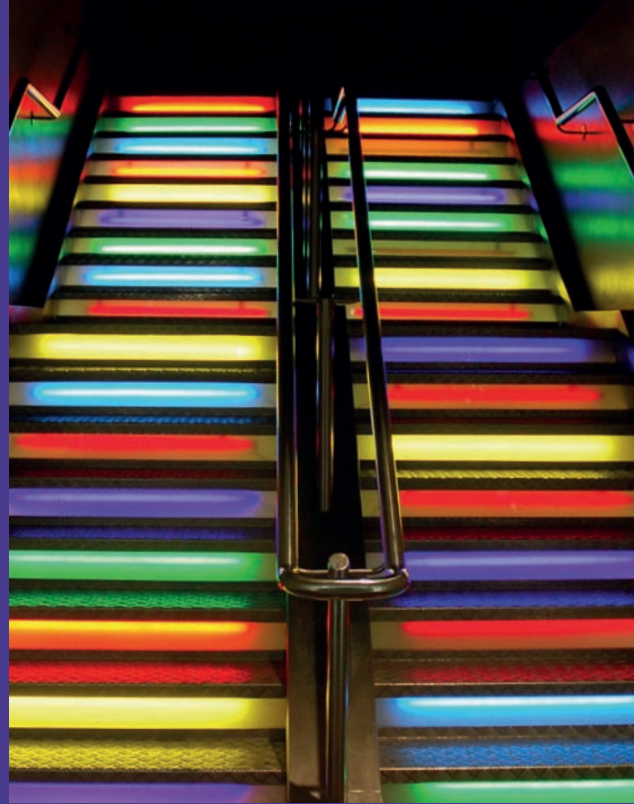
Gender	Age group	Anxiety	Mood	Body shape satisfaction
Male	<25	87	74	11
Female	25–40	54	61	23
Female	41–55	31	38	?
Male	25–40	43	39	34
Male	<25	69	82	8
Female	41–55	18	12	51
Female	25–40	38	77	29
Male	<25	74	65	16

Open a new SPSS data set.

1. Create the variables, using the parameters shown above.
2. Enter the data, using the raw data from Table 2.2.
3. Create a new variable that measures a combination of 'Anxiety' and 'Mood' scores added together (called 'Affect').
 - a. Format the variable parameters for the new variable.
4. Recode the Gender variable (using values of 0 = male and 1 = female).
 - a. Format the variable parameters for the new variable.

3

NORMAL DISTRIBUTION



Learning objectives

By the end of this chapter you should be able to:

- Understand the importance of normal distribution
- Recognise the effects of skew and kurtosis, and what we mean by 'outliers'
- Appreciate how to measure and interpret normal distribution graphically and statistically
- Recognise ways in which we can deal with potential violations in normal distribution
- Understand how to adjust outliers and transform variables
- Recognise what we mean by homogeneity and sphericity of variances

What is normal distribution?

Normal distribution describes the way in which data are ‘spread’. Imagine that we collected some information about the age for a group of 30 people, aged between 18 and 50. Some of those people would be younger, some older, others somewhere in between. **Probability** statistics describe the likelihood of something happening based on what we know about previous outcomes. In probability, we expect things to happen in a predictable, uniform way. If our group was representative of the general population, we would expect the ages of our group to be pretty evenly spread out. However, there may be circumstances that might cause those ages to be not so even. If this were a group of university students, we might expect most of the ages to tend towards being younger; if the group were members of a crown green bowls club, the ages might be somewhat older. In normal distribution, we start with the assumption that the data we collect represent something close to the general population. In our example, we could plot the ages in a graph: the range of ages would be placed in ascending order along the horizontal (x) axis and we would count the number of people matching that age along the vertical (y) axis. The graph might look something like the one shown in Figure 3.1.

The bars in Figure 3.1 represent a group of age bands, with the height of the bar showing how many people are in that group of ages. We have added a curve that shows the trend of the ages (we will see how to draw this graph, including adding the curve, later on). That curve is useful in two respects: it shows how ‘evenly spread’ the ages are across the group and it provides some information on what the average age of the group is likely to be. The peak of that curve approximately indicates the average age (around 35 in this example). Overall, we appear to have a gentle ‘bell-shaped’ curve where the distribution of ages is roughly equal either side of the mean. It is this type of ‘normal distribution’ that we should be aiming for. In this chapter we will explore exactly how we quantify that.

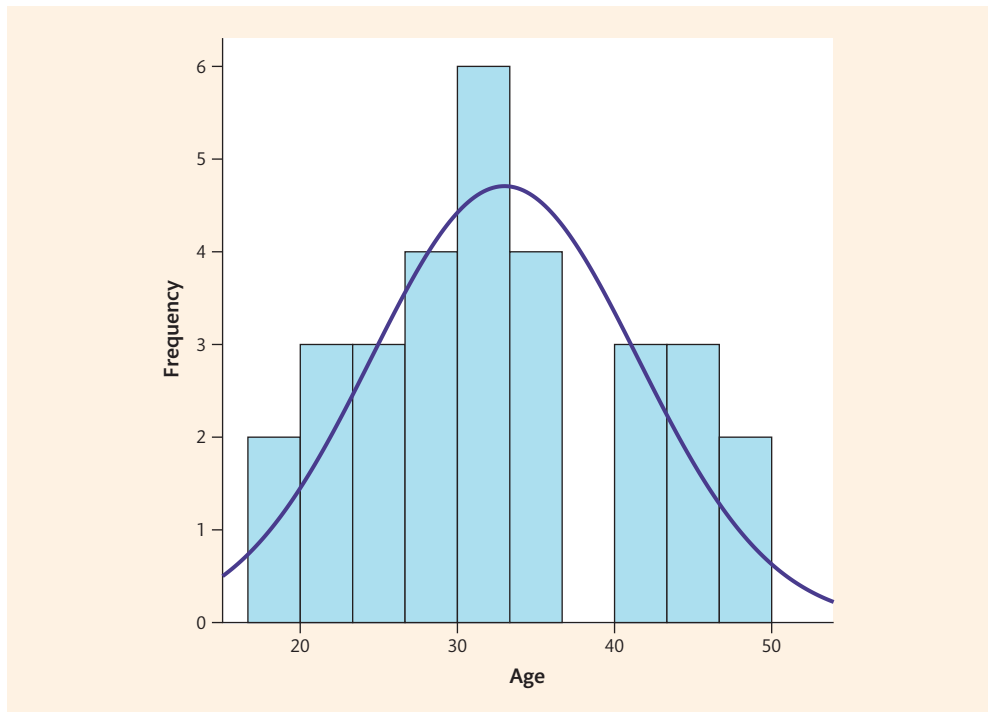


Figure 3.1 Distribution of ages (n = 50)

What does normal distribution look like?

Looking at Figure 3.1, we have some idea of what normal distribution looks like. For data to be 'normally distributed' we expect them to be 'evenly' distributed either side of the mean, illustrated by a smooth, bell-shaped pattern, and where the 'peak' of that distribution is neither 'too pointed' nor 'too flat'. Graphically, we often draw a curve through the data to indicate the trend in those data; we call these '**histograms**'. To illustrate a good example of normal distribution, compared with examples where normal distribution may have been compromised, we need to look at a series of histograms. We need to compare the curves in these histograms to appreciate how they differ. However, before we start, we need to learn some basic terms about how we measure data (see Box 3.1).

3.1 Nuts and bolts

Basic units of measurement



- Mean:** This is the average number in a data set. We add up all of the numbers in the data set and divide the answer by the number of cases (or people).
- Median:** This is the middle number in a data set, when those numbers have been ordered numerically from lowest to highest (or vice versa).
- Mode:** This is the most common number in a data set.

We will start with an example of a normal distribution. Table 3.1 shows what some normally distributed data might look like.

Table 3.1 Example data for normal distribution

Ages															Mean	Median	Mode
20	23	28	28	32	32	35	35	35	38	38	42	42	47	50	35.0	35	35

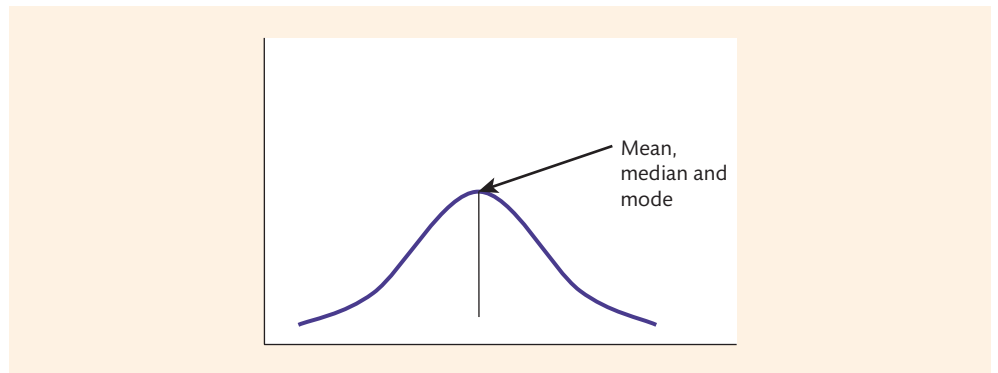


Figure 3.2 Normal distribution

Figure 3.2 is an example of a normal distribution. It is signified by a smooth, bell-shaped curve. The mean and median are identical.

Skewed data

By definition, normal distribution describes a range of data where the scores at either end of the distribution are the same distance to the mean. In our example, the eldest person is 15 years older than the mean age; the youngest is 15 years younger than the mean age. If there are extreme scores at one end of the distribution it is likely to ‘skew’ the mean score away from the median. We call those extreme scores ‘outliers’. If the data are skewed, this can distort the mean score and can bias any test that depends on it (as we will see later).

Positively skewed data

When the data are positively skewed, there are extreme (outlier) scores at the higher end of the range of data. This might cause the mean score to be overstated (see Table 3.2).

Table 3.2 Example of positively skewed data

Ages															Mean	Median	Mode
20	23	28	28	32	32	35	35	35	38	38	42	42	55	60	36.2	35	35

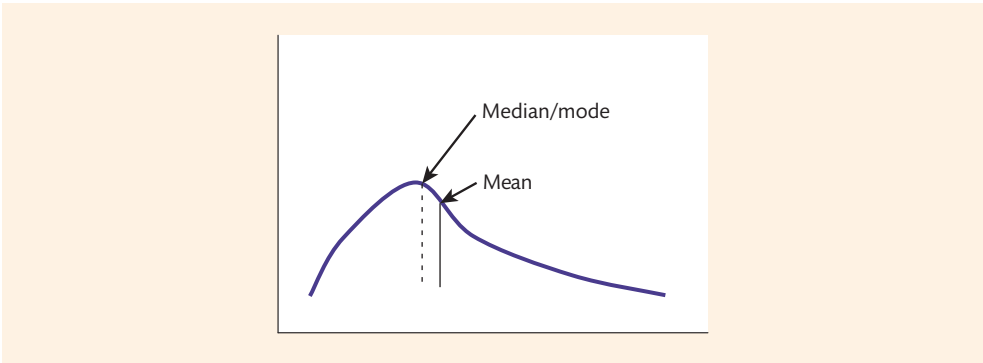


Figure 3.3 Positively skewed distribution

Figure 3.3 shows data that are positively skewed. One tip of the curve points towards the right-hand side of the distribution. The mean is drawn to the right of the median and mode. The high extreme scores may have artificially inflated the mean score.

Negatively skewed data

When the data are negatively skewed, there are extreme scores at the lower end of the range of data. This might cause the mean score to be understated (see Table 3.3).

Table 3.3 Example of negatively skewed data

Ages															Mean	Median	Mode
9	10	28	28	32	32	35	35	35	38	38	42	42	47	50	33.4	35	35

Figure 3.4 presents an example of negative skew. One tip of the curve points towards the left-hand side of the distribution. The mean is drawn to the left of the median and mode. The low extreme scores may have artificially deflated the mean score.

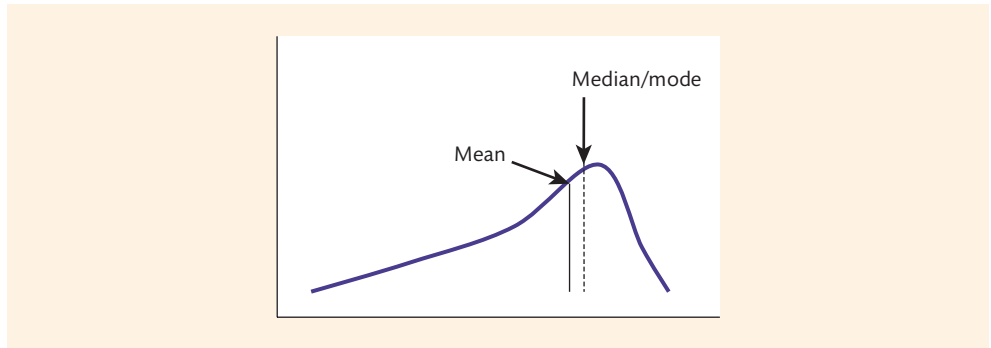


Figure 3.4 Negatively skewed distribution

Kurtosis

In addition to skew, we need to measure **kurtosis**. This describes the 'peakedness' of the curve. A normal distribution is often referred to as being 'mesokurtic', which is another reference to the 'bell shape' that we are aiming for. However, we may encounter problems with curves that are too 'peaked', or ones that are too 'flat'.

Leptokurtic distributions

A **leptokurtic** distribution describes a curve that is 'peaked', like a pointed hat (see Table 3.4).

Table 3.4 Example of leptokurtic data

Ages														Mean	Median	Mode	
31	31	32	32	34	34	35	35	35	36	36	38	38	39	39	35	35	35

Although the mean and median are the same, there is very little variation in the data, making analyses difficult. Graphically, the data distribution might look like Figure 3.5.

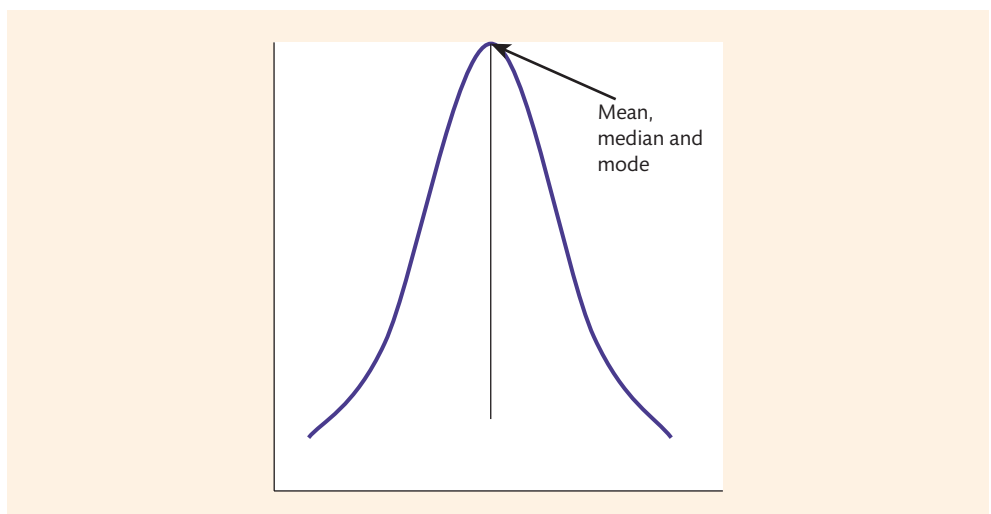


Figure 3.5 Leptokurtic distribution

Platykurtic distributions

A **platykurtic** distribution describes a curve that is flat (see Table 3.5).

Table 3.5 Example of platykurtic data

Ages																Mean	Median	Mode
20	22	24	26	28	30	34	35	36	40	42	44	46	48	50		35	35	None

Once again, the mean and median are the same, but now there is too much variation in the data to make analyses viable. Graphically, the data distribution might look like Figure 3.6.

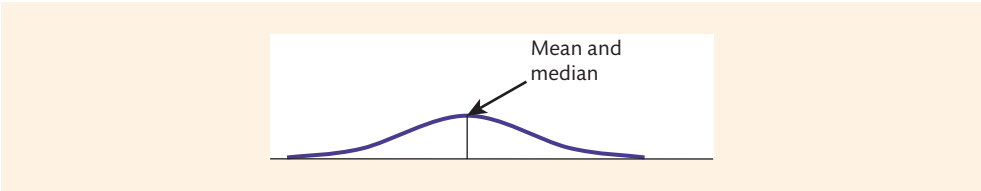


Figure 3.6 Platykurtic distribution

3.2 Take a closer look

Terms used in measuring normal distribution



- Normally distributed:** data are evenly distributed either side of the mean (see Figure 3.2)
- Positive skew:** where there are outliers at the higher end of a data set (see Figure 3.3)
- Negative skew:** where there are outliers at the lower end of a data set (see Figure 3.4)
- Kurtosis:** describes the peakedness of a normal distribution curve
- Mesokurtic:** a 'normal' curve, as demonstrated by the bell shape (see Figure 3.2)
- Leptokurtic:** very 'peaked' distribution, with little variation in the data (see Figure 3.5)
- Platykurtic:** very 'flat' distribution, with data widely dispersed across the data set (see Figure 3.6)

What happens when data are not normally distributed?

As we have just seen, data may not be normally distributed if there are problems with skew and kurtosis. Data that are positively skewed may cause the mean score to be artificially inflated. This may have occurred because there are some extreme high scores. Without those outliers, a more realistic mean score might have been somewhat lower. Similarly, data that are negatively skewed might lead to an artificially deflated mean because of some extreme low scores. Either way, the mean score in skewed data may not be reliable. We also saw that deviations in kurtosis may cause a problem. Leptokurtic distributions may offer too little variation in the data, while platykurtic distributions may have too much variation. But why might all of this be a problem? Many of the statistical procedures that we will explore in this book depend on measuring differences in mean scores. We will come to know these as parametric tests (we will explore this in more depth in Chapter 4). Normal distribution is a major determinant in deciding whether we can classify our data as parametric. If normal distribution has been compromised, we may no longer be able to

trust the mean score as truly reflecting the data. If we cannot trust the mean score, we may have less confidence in the outcome produced by parametric tests. In short, if we lack normal distribution we may need to choose alternative tests (such as those examined in Chapter 18).

Measuring normal distribution

So how can we check that our data are normally distributed? We can get SPSS to help us here. This can be achieved through the production of graphs (such as histograms, **box plots** or **stem-and-leaf plots**), or we can employ statistical procedures. We will look at each of these in turn.

Graphical procedures

Histograms

In Figure 3.1, we saw a graphical representation of normal distribution. This type of graph is called a histogram. It is a bar chart, where bars represent individual cases or groups, and where the height of the bar indicates the frequency of that outcome. We can add a curve to the display to illustrate normal distribution. We can get SPSS to draw this histogram:

Open the SPSS file **Age and sleep quality**

Select **Analyze** → **Descriptive Statistics** → **Frequencies** (as shown in Figure 3.7)

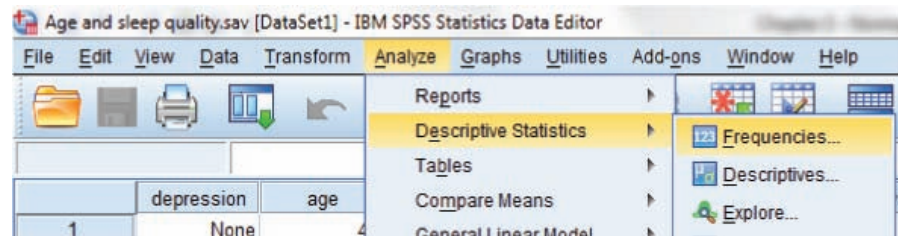


Figure 3.7 Creating histograms – step 1

In new window (see Figure 3.8) transfer **Age** to **Variable(s)** window (by clicking on the arrow to the left of that window, or by 'dragging' the variable there) → click **Statistics**

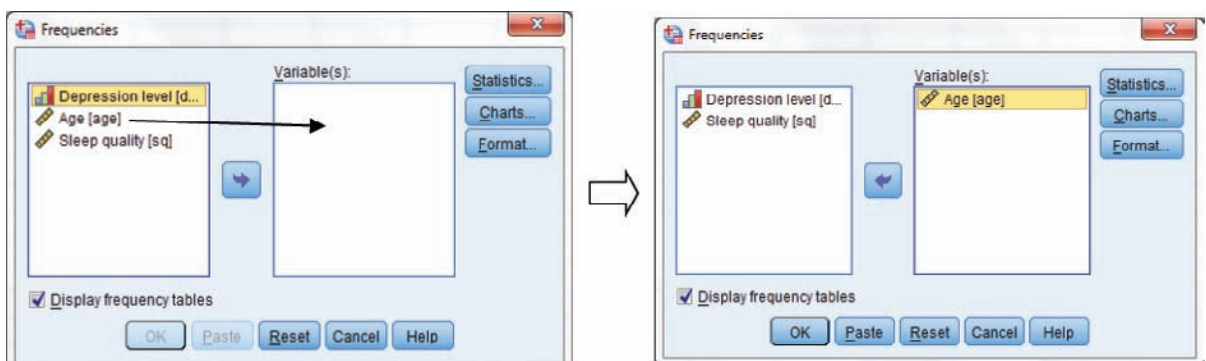


Figure 3.8 Creating histograms – step 2

In new window (see Figure 3.9) select **Mean**, **Median**, and **Mode** radio buttons → click **Continue** → (in original window) click **Charts**

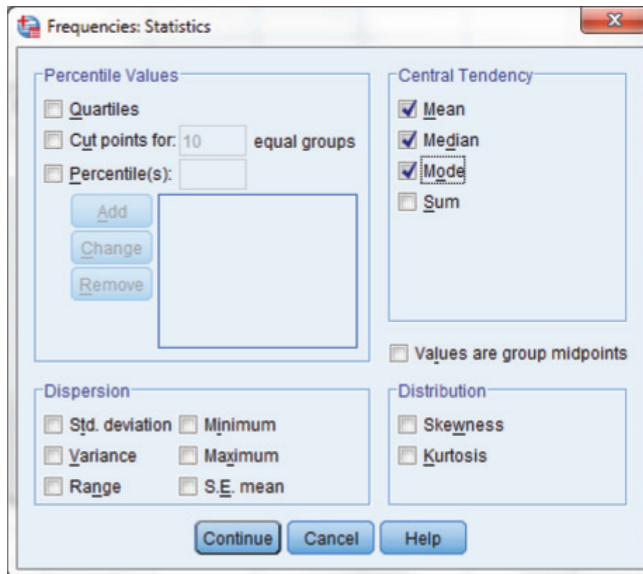


Figure 3.9 Creating histograms – step 3

In new window (see Figure 3.10) click **Histogram** radio button → tick **Show normal curve on histogram** box → click **Continue** → (in original window) click **OK**

If you need further guidance on these procedures, you can visit the website for this book and follow the video guides for SPSS

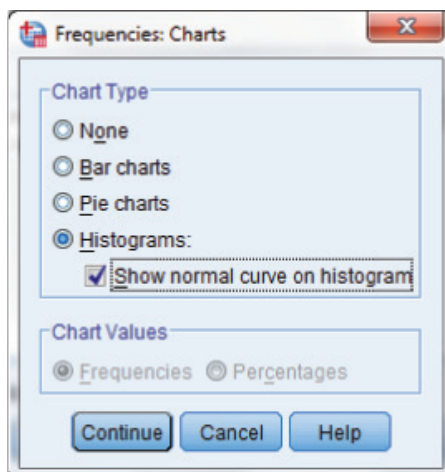


Figure 3.10 Creating histograms – step 4

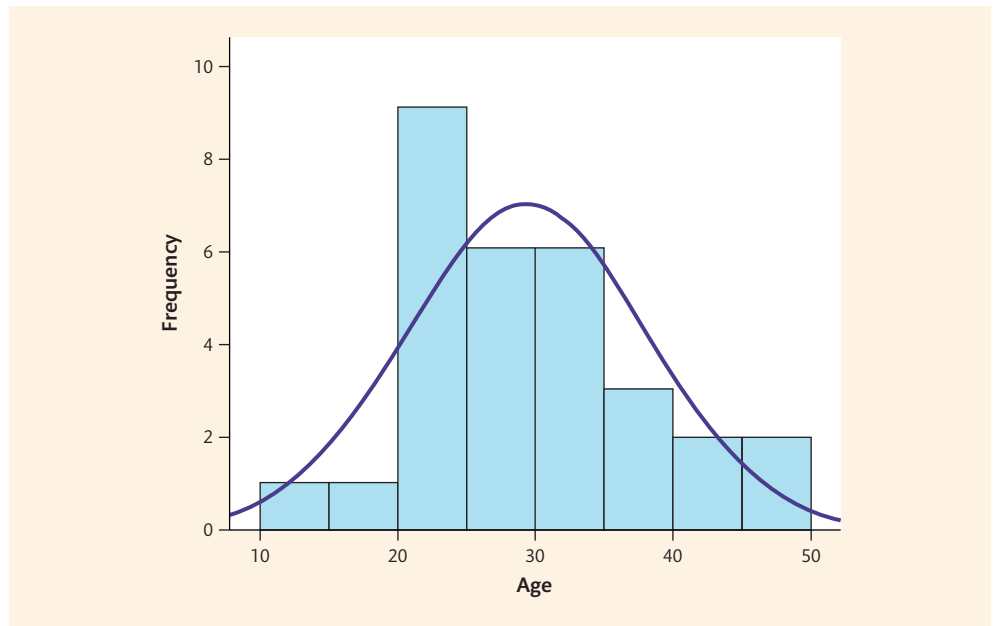


Figure 3.11 Completed histogram

This *appears* to be a pretty good example of a normal distribution, at least according to the curve that has been added to the graph (see Figure 3.11). However, we may feel that the bars suggest slightly positively skewed data. To help us here, we can refer to the descriptive statistics that we asked for (see Table 3.6).

Table 3.6 Descriptive data

	Mean	Median	Mode
Age	29.30	27	24

Table 3.6 suggests that there are some differences in the mean, median and mode. These differences might cause us to question whether the data are normally distributed after all. This illustrates a drawback of graphical displays: they can be a little subjective. However, we can supplement the graphs with formal statistics, which is something we will look at shortly. Nevertheless, these graphical displays are useful in providing some initial indications about normal distribution, so we should look at a few more examples.

Box plots

Another graphical display that we can use is called a box plot (also known as a box and whisker plot, for reasons that are about to become obvious). Some examples of box and whisker plots are shown in Figure 3.12.

Box plots show how the data are spread around the median (the thick line through the box, representing the middle point of the data). The inter-quartile ranges are represented by the 'hinges' at either end of the box. The bottom hinge is equivalent to the (lower) 25 per cent data point; the higher hinge symbolises the (upper) 75% data point. The 'whiskers', either side of the boxes, approximately represent the lowest and highest scores (unless there are outliers – see later).

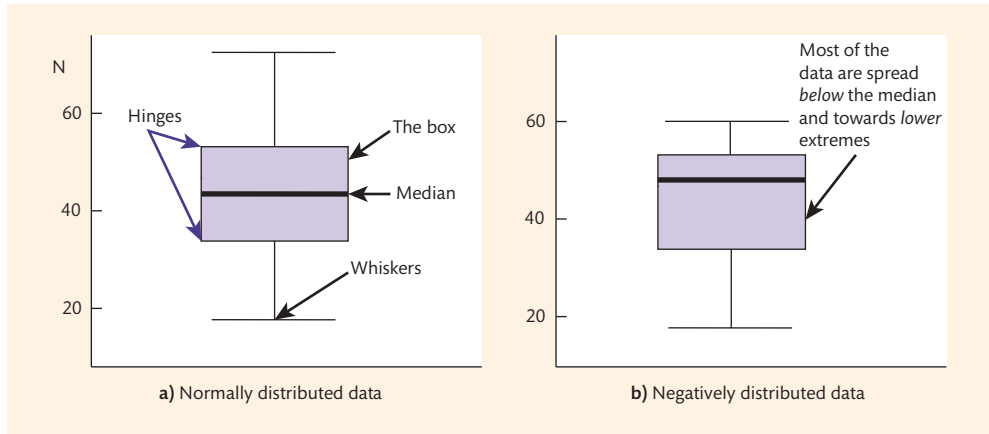


Figure 3.12 Box and whisker plot

Figure 3.12a shows an example of a normal distribution – data are evenly spread either side of the median, with whiskers at equal length above and below the box. Figure 3.12b illustrates some negatively skewed data – there is a larger shaded area below the median line than above it, and there is a disproportionately longer whisker below the box than above it. Positively skewed data will show the opposite of this. This is how we can request a box plot in SPSS (using the same data as we examined with a histogram):

Select **Analyze** → **Descriptive Statistics** → **Explore** (see Figure 3.7) → (in new window, as shown in Figure 3.13) transfer **Age** to **Dependent List** window → select **Plots** radio button → click **Plots** box

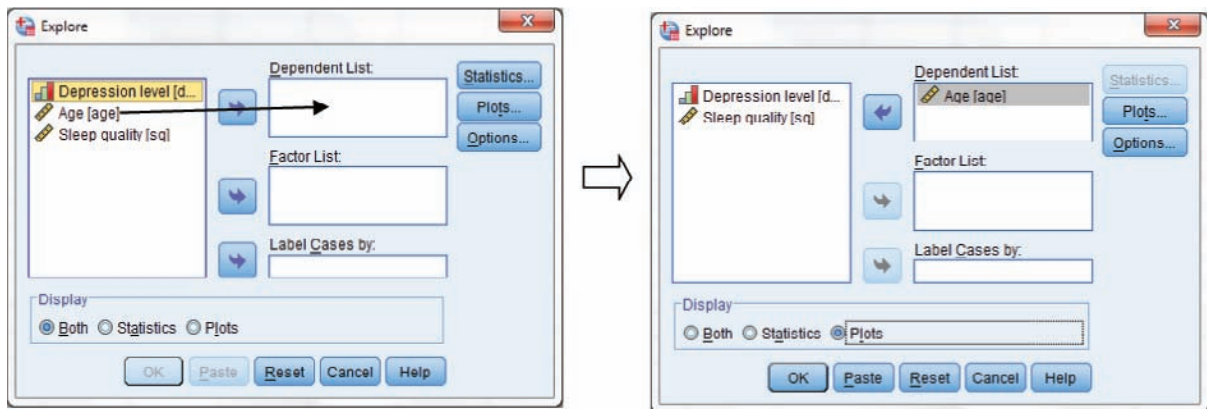


Figure 3.13 Creating box plots – step 1

In new window (as shown in Figure 3.14), click **Factor levels together** radio button (under **Boxplot**) → make sure that **Stem-and-leaf** and **Histogram** (under **Descriptive**) are unchecked (for now) → click **Continue** → (in original window) click **OK**

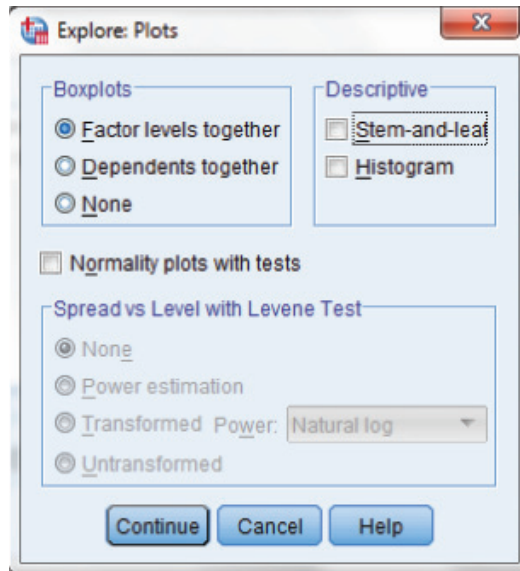


Figure 3.14 Creating box plots – step 2

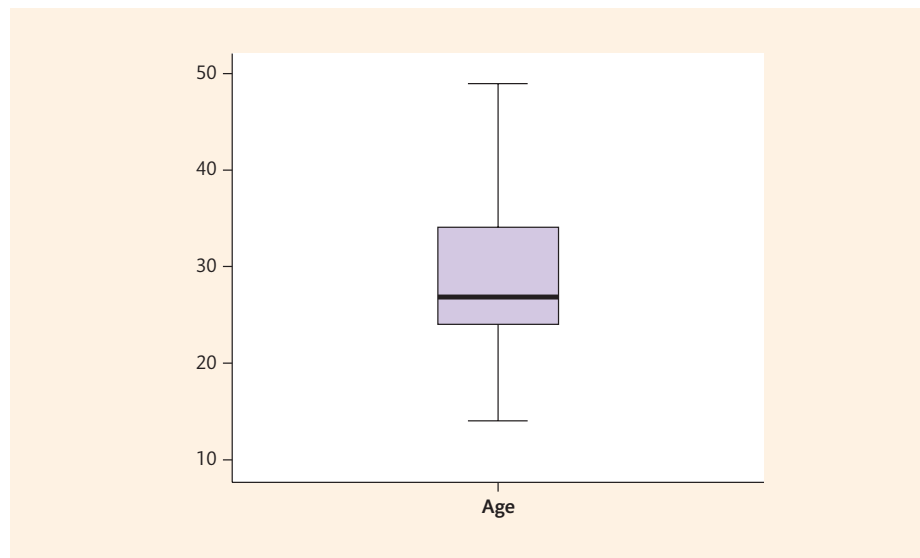


Figure 3.15 Completed box plot

Given what we saw in Figure 3.12b, we might conclude that the data appear to be positively skewed (an outcome potentially supported by the data in Table 3.6).

Stem-and-leaf plots

Stem-and-leaf plots are another way in which we can present data to visually examine normal distribution. The style of presentation is similar to histograms but has the added advantage of retaining the actual numbers within the graphical display. The 'stem' refers to a group of data (usually tens, hundreds, thousands, etc.) and the 'leaf' refers to units within that group. An example is shown in Figure 3.16.

<u>Stem</u>	<u>Leaf</u>	
Tens	Units	
0	3 3 3 5	The red bold number in this row represents 3
1	2 4 6	The red bold number in this row represents 16
2	0 0 2 4	The red bold number in this row represents 20

Figure 3.16 Simple stem-and-leaf plot

Larger data sets are arranged in a similar fashion, but can be more easily assessed to establish whether those data are normally distributed. A larger set of numbers is shown in Figure 3.17.

<u>Stem</u>	<u>Leaf</u>
Tens	Units
0	
1	
2	
3	1,3,3
4	2,3,3,4,4,8
5	0,0,2,4,5,6,8
6	1,1,4,4,4,4,5,8,8,9
7	0,2,3,3,4,6,7
8	1,1,4,4,5,6
9	4,4,6
10	
11	

Figure 3.17 Normally distributed stem-and-leaf plot

The data in Figure 3.17 appear to be normally distributed, because the numbers are evenly spread either side of those in the 60s range. If we rotated the display 90° (anticlockwise), we would see the bell-shaped curve typical of normal distributions presented by histograms (as shown by Figure 3.2). However, this ‘histogram’ has actual numbers in it.

Figure 3.18 presents a stem-and-leaf plot where the data may be positively skewed. If we were to rotate this 90° (anticlockwise), we would see a distribution similar to the positively skewed histogram we saw in Figure 3.3. The tail tends towards the higher numbers. A negatively skewed

<u>Stem</u>	<u>Leaf</u>
Tens	Units
0	
1	
2	
3	1,3,3,5
4	2,3,3,4,4,8,8
5	0,0,2,4,5,6,7,9
6	1,1,4,4,4,4,5,8,8
7	0,2,3,3,4,6
8	1,1,4,4,6
9	4,4,6
10	1,7
11	3
12	5

Figure 3.18 Positively distributed stem-and-leaf plot