

Statistics for Research

With a Guide to SPSS

© George Argyrous 2005

First edition published 2000
Second edition published 2005

SPSS and associated proprietary computer software are the trademarks of SPSS Inc.

Statistics for Research is not sponsored or approved or connected with SPSS Inc. All references in the text of this book to SPSS products are to the trademarks of SPSS Inc.

All names of computer programs are hereby acknowledged as trademarks (registered or otherwise), whether or not the symbol appears in the text.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may not be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licenses issued by the Copyright Licensing Agency. Inquiries concerning reproduction outside those terms should be sent to the publishers.



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

SAGE Publications
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
32, M-Block Market
Greater Kailash – I
New Delhi 110 048

British Library Cataloguing in Publication data

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

ISBN 1 4129 1947 9
ISBN 1 4129 1948 7 (pbk)

Library of Congress Control Number: 2005929626

Typeset in Times New Roman
Printed on paper from sustainable sources
Printed in Great Britain by The Cromwell Press, Townbridge, Wiltshire

Contents

Preface

Part 1 An introduction to statistical analysis

- 1 **Variables and their measurement** 3
 - The conceptualization and operationalization of variables 4
 - Scales of measurement 7
 - Levels of measurement 8
 - Univariate, bivariate, and multivariate analysis 11
 - Descriptive statistics 14
 - Exercises 15
- 2 **Setting up an SPSS data file** 17
 - Obtaining a copy of SPSS 17
 - Alternatives to SPSS 17
 - Options for data entry in SPSS 18
 - The SPSS Data Editor 19
 - Assigning a variable name 21
 - Setting the data type 22
 - Setting the data width and decimal places 23
 - Defining variable labels 23
 - Defining value labels 24
 - Setting missing values 25
 - Setting the column format and alignment 26
 - Specifying the level of measurement 27
 - Generating variable definitions in SPSS 28
 - The SPSS Viewer window 32
 - Saving a data file 32
 - Data entry 33
 - Checking for incorrect values: Data cleaning 35
 - Summary 35
 - Exercises 35

Part 2 Descriptive statistics: Graphs and tables

- 3 **The graphical description of data** 39
 - Some general principles 39
 - Pie graphs 40
 - Bar graphs 42
 - Histograms and polygons 44
 - Interpreting a univariate distribution 46
 - Graphing two variables 47
 - Common problems and misuses of graphs 50
 - Exercises 53

- 4 The tabular description of data 55**
 Listed data tables 55
 Simple frequency tables 55
 Relative frequency tables: percentages and proportions 57
 Cumulative frequency tables 60
 Class intervals 61
 Percentiles 64
 Frequency tables using SPSS 65
 Valid cases and missing values 67
 Improving the look of tables 67
 Exercises 68
- 5 Using tables to investigate the relationship between variables: Crosstabulations 70**
 Crosstabulations as descriptive statistics 70
 Types of data suitable for crosstabulations 72
 Crosstabulations with relative frequencies 73
 Crosstabulations using SPSS 74
 Interpreting a crosstabulation: The pattern and strength of a relationship 75
 Interpreting a crosstabulation when both variables are at least ordinal 76
 Summary 78
 Exercises 78
- 6 Measures of association for crosstabulations: Nominal data 81**
 Measures of association as descriptive statistics 81
 Measures of association for nominal scales 83
 Properties of lambda 86
 Lambda using SPSS 87
 Limitations on the use of lambda 90
 Standardizing table frequencies 92
 Exercises 93
- 7 Measures of association for crosstabulations: Ranked data 95**
 Data considerations 95
 Concordant pairs 96
 Discordant pairs 97
 Measures of association for ranked data 98
 Gamma 99
 Somers' d 101
 Kendall's tau- b 102
 Kendall's tau- c 102
 Measures of association using SPSS 102
 Summary 107
 Exercises 107
- 8 Multivariate analysis of crosstabs: Elaboration 110**
 Direct relationship 110
 Elaboration of crosstabs using SPSS 112
 Partial gamma 113
 Spurious or intervening relationship? 114
 Conditional relationship 115
 Summary 117
 Exercises 118

Part 3 Descriptive statistics: Numerical measures

- 9 Measures of central tendency 123**
 Measures of central tendency 123
 The mode 124
 The median 125
 The mean 126
 Choosing a measure of central tendency 128
 Measures of central tendency using SPSS: Univariate analysis 129
 Measures of central tendency using SPSS: Bivariate and multivariate analysis 132
 Summary 133
 Exercises 134
- 10 Measures of dispersion 136**
 The range 136
 The interquartile range 137
 The standard deviation 138
 Coefficient of relative variation 140
 Index of qualitative variation 141
 Measures of dispersion using SPSS 145
 Summary 145
 Exercises 146
- 11 The normal curve 147**
 The normal distribution 147
 Using normal curves to describe a distribution 150
 z-scores 151
 Normal curves on SPSS 157
 Exercises 159
- 12 Correlation and regression 161**
 Scatter plots 161
 Linear regression 162
 Pearson's product moment correlation coefficient 169
 Explaining variance: The coefficient of determination 170
 Plots, correlation, and regression using SPSS 172
 The assumptions behind regression analysis 177
 Spearman's rank-order correlation coefficient 179
 Spearman's rho using SPSS 180
 Correlation where the independent variable is categorical: Eta 182
 Summary 183
 Exercises 183
- 13 Multiple regression 187**
 Introduction to multiple regression 188
 Multiple regression with SPSS 190
 Testing for the significance of the multivariate model 193
 Alternative methods for selecting variables in the regression model 193
 Stepwise regression 194
 Extending the basic regression analysis: Adding categorical independent variables 197
 Further extensions to the basic regression analysis: Hierarchical regression 198
 The assumptions behind multiple regression 198
 Exercises 199

Part 4 Inferential statistics: Tests for a mean**14 Sampling distributions 203**

- Random samples 204
- The sampling distribution of a sample statistic 205
- The central limit theorem 210
- Generating random samples using SPSS 210
- Summary 212
- Exercises 212

15 Introduction to hypothesis testing and the one sample z -test for a mean 214

- Step 1: State the null and alternative hypotheses 217
- Step 2: Choose the test of significance 219
- Step 3: Describe the sample and derive the p -score 220
- Step 4: Decide at what alpha level, if any, the result is statistically significant 222
- Step 5: Report results 224
- What does it mean when we 'fail to reject the null hypothesis'? 226
- What does it mean to 'reject the null hypothesis'? 226
- A two-tail z -test for a single mean 227
- The debate over one-tail and two-tail tests of significance 228
- A one-tail z -test for a single mean 229
- Summary 230
- Appendix: Hypothesis testing using critical values of the test statistic 230
- Exercises 231

16 The one sample t -test for a mean 233

- The Student's t -distribution 233
- The one sample t -test for a mean 234
- The one sample t -test using SPSS 238
- Summary 239
- Exercises 240

17 Inference using estimation and confidence intervals 242

- The sampling distribution of sample means 242
- Estimation 243
- Changing the confidence level 247
- Changing the sample size 250
- Estimation using SPSS 250
- Confidence intervals and hypothesis testing 252
- Exercises 253

18 The two samples t -test for the equality of means 255

- Dependent and independent variables 256
- The sampling distribution of the difference between two means 257
- The two samples t -test for the equality of means 259
- The two samples t -test using SPSS 261
- Exercises 264

19 The F -test for the equality of more than two means: Analysis of variance 266

- The one-way analysis of variance F -test 269
- ANOVA using SPSS 272
- Summary 277
- Exercises 279

20 The two dependent samples t -test for the mean difference 280

- Dependent and independent samples 280
- The two dependent samples t -test for the mean difference 281
- The two dependent samples t -test using SPSS 283
- Exercises 286

Part 5 Inferential statistics: Tests for frequency distributions**21 One sample tests for a binomial distribution 291**

- Data considerations 291
- The sampling distribution of sample percentages 292
- The z -test for a binomial percentage 293
- The z -test for a binomial percentage using SPSS 295
- Estimating a population percentage 297
- The runs test for randomness 299
- The runs test using SPSS 302
- Exercises 303

22 One sample tests for a multinomial distribution 305

- The chi-square goodness-of-fit test 305
- Chi-square goodness-of-fit test using SPSS 308
- The chi-square goodness-of-fit test for normality 312
- Summary 313
- Exercises 314

23 The chi-square test for independence 316

- The chi-square test and other tests of significance 316
- Statistical independence 317
- The chi-square test for independence 317
- The distribution of chi-square 322
- The chi-square test using SPSS 323
- Problems with small samples 328
- Problems with large samples 329
- Appendix: hypothesis testing for two percentages 331
- Exercises 333

24 Frequency tests for two dependent samples 335

- The McNemar chi-square test for change 335
- The McNemar test using SPSS 337
- The sign test 338
- Summary 340
- Exercises 340

Part 6 Inferential statistics: Other tests of significance**25 Rank-order tests for two or more samples 343**

- Data considerations 343
- The rank sum and mean rank as descriptive statistics 344
- The z -test for the rank sum for two independent samples 348
- Wilcoxon's rank sum z -test using SPSS 352
- The Wilcoxon signed-ranks z -test for two dependent samples 353
- The Wilcoxon signed-ranks test using SPSS 356

Other non-parametric tests for two or more samples	357
Appendix: the Mann-Whitney U test	358
Exercises	359
26 The t-test for a correlation coefficient	362
The t -test for Pearson's correlation coefficient	362
Testing the significance of Pearson's correlation coefficient using SPSS	364
The t -test for Spearman's rank-order correlation coefficient	365
Testing the significance of Spearman's correlation coefficient using SPSS	366
Testing for significance in multiple regression	367
Exercises	368
Appendix	369
Table A1 Area under the standard normal curve	369
Table A2 Critical values for t -distributions	370
Table A3 Critical values for F -distributions ($\alpha = 0.05$)	371
Table A4 Critical values for chi-square distributions	372
Table A5 Sampling errors for a binomial distribution (95% confidence level)	373
Table A6 Sampling errors for a binomial distribution (99% confidence level)	373
Key equations	374
Glossary	379
Answers	383
Index	397

Preface

This book is aimed at students and professionals who do not have any existing knowledge in the field of statistics. It is not unreasonable to suggest that most people who fit that description come to statistics reluctantly, if not with hostility. It is usually regarded as 'that course we had to get through'. I suspect that a sense of dread is also shared by many instructors when confronted with the prospect of having to teach the following material.

This book will hopefully ease some of these problems. It is written by a non-statistician for non-statisticians, for students who are new to the subject, and for professionals who may use statistics occasionally in their work. It is certainly not the only book available that attempts to do this. One might in fact respond with the statement 'not another stats book!' There are important respects, however, in which this book is different to the other numerous books in the field. When this book was first published by Macmillan Education as *Statistics for Social Research* it differentiated itself from other texts in three ways, each of which have been carried into this edition:

Communication of ideas. This book is written with the aim of communicating the basic ideas and procedures of statistical analysis to the student and user, rather than as a technical exposition of the fine points of statistical theory. The emphasis is on the explanation of basic concepts and especially their application to 'real-life' problems, using a more conversational tone than is often the case. Such an approach may not be as precise as others in dealing with statistical theory, but it is often the mass of technical detail that leaves readers behind, and turns potential users of statistical analysis away.

Integrated use of SPSS. This book integrates the conceptual material with the use of the main computer software package, SPSS. The development and availability of this software has meant that for most people 'doing stats' equals using a computer. The two tasks have converged. Unfortunately, most books have not caught up with this development and adequately integrated the use of computer packages with statistical analysis. They concentrate instead on the logic and formulas involved in statistical analysis and the calculation 'by hand' of problem solutions. At best other books have appendices that give brief introductions and guides to computer packages, but this does not bridge the gap between the hand calculations and the use of computer software. This book builds the use of SPSS into the text. The logic and application of various statistical techniques are explained, and then the examples are reworked on SPSS. Readers can link explicitly the traditional method of working through problems 'by hand' and working through the same problems on SPSS. Exercises also explicitly attempt to integrate the hand calculations with the use and interpretation of computer output.

To help readers along, a CD with all the data necessary to generate the results in the following chapters is included with this book, so that all the procedures described there can be replicated. You will need your own copy of SPSS to perform these procedures, and Chapter 2 lists a number of means by which you can obtain SPSS.

It is necessary, however, to point out that this is not a complete guide to SPSS. This book simply illustrates how SPSS can be used to deal with the basic statistical techniques that most researchers commonly encounter. It does not exhaust the full range of functions and options available in SPSS. For the advanced user, nothing will replace the *User's Guide* published by SPSS Inc. But for most people engaged in research, the following text will allow them to handle the bulk of the problems they will encounter.

For users of other statistics packages, the files are also saved in ASCII and Excel format so that they can be imported to these programs, along with a **Readme** file that explains the data definitions. All the files, and periodic minor updates and corrections, can be obtained at the following web site:

www.sagepub.co.uk/argyrous

Clear guide to choosing the appropriate procedures. This book is organized around the individual procedures (or sets of procedures) needed to deal with the majority of problems people encounter when analyzing quantitative data. Other texts flood the reader with procedure after procedure, which can be overwhelming. How to choose between the options? This book concentrates on just the most widely used techniques, and sorts through them by building the structure of the book around these options. Entire chapters are devoted to individual tests so that the situations in which a particular test is applied will not be confused with situations that call for other tests. Thus after working through the text, readers can turn to individual chapters as needed in order to address the particular problems they encounter.

The first version of this text proved to be popular in disciplines outside the social sciences, especially in the health sciences. As a result, the next version, published by Sage, UK, as *Statistics for Social and Health Research*, broadened its appeal to the health sciences through the inclusion of examples and exercises suited to their interests, but which were still intelligible to a non-specialist.

The second version of this text, as with the first, also found a broader audience than suggested by its title. This broad appeal suggested to me that a comprehensible 'generic' statistics textbook is of value to researchers in any field, and also desperately needed. Thus, this version drops from its title any reference to a specific discipline; its appeal is to all researchers who need some basic understanding of quantitative methods and the use of SPSS. Some specialized topics that are normally covered in certain fields and not others, such as the greater interest in small sample problems in the health sciences than in the social sciences, are not covered as a result. I have found, however, that instructors or students can supplement the basic techniques covered in this text with such specialized topics as required, especially given the vast amount of material now available on the internet.

In developing this new edition, I have also made some substantial changes (improvements!), while still retaining the three broad objectives set for the first version.

Reordering chapters around classes of descriptive techniques rather than levels of measurement. The previous editions were criticized, rightly I believe, for being too rigid in their emphasis on the limits placed on analysis by levels of measurement. When people analyze data they usually think in terms of classes of statistics first, such as central tendency, frequency tables, or correlation. The level at which variables are measured is an important consideration, but does not correspond to the way researchers 'think' about the problems they want to address. To accommodate this, chapters have been organized around the mainly used descriptive techniques, with data considerations (including levels of measurement) forming an element in the exposition of those techniques.

Reference to material available on the internet. The material now available on the internet is extensive and growing all the time. The lack of 'quality control', however, can make the use of such material fraught with perils. I have drawn on internet tools where appropriate and where I have been able to assess the quality of the information and resources presented. I have given the address for these internet sources in the text, but the reader should be aware that the maintenance of these sites is out of the control of myself or Sage, UK.

Streamlining of the five-step hypothesis testing procedure. I have dramatically altered the five-step hypothesis testing procedure, eliminating the calculation of critical scores and the pre-setting of alpha levels. This is in response to the current trends in academic journals, especially in the health sciences, which seek a less prescriptive approach to decision-making than has been the case in the past. It also reduces the calculations needed to arrive at a conclusion; a great relief to many students.

Greater emphasis on reporting results. I have found that researchers are often at a loss as to how to communicate their findings. I therefore have built into the five-step hypothesis testing procedure an explication of how to report findings. Getting results is one thing, but unless these can be communicated, especially to a general audience, their importance is lost.

Reference to the literature on statistical methods. Textbooks are always a lie. They present a field of knowledge as uncontroversial, when in fact it is usually a terrain of hot debate. This is no less the case with statistics textbooks, including previous incarnations of this one. Rather than continue the lie, I have introduced at various places some important points of debate and references to the literature where those interested can pursue the debates further.

Having noted the main features of this book as compared to others in the field, it is also worth noting what this book is not. This book looks at the analysis of quantitative data, and only the analysis of quantitative data. It makes no pretence to being a comprehensive guide to social or health research. Issues relating to the selection of research problems, the design of research methods, and the procedures for checking the validity and reliability of results are not covered. Such a separation of statistics from more general considerations in the design of research is a dangerous practice since it may give the impression that statistical analysis *is* research. Yet, nothing could be further from the truth. Statistical analysis is one way of processing information, and not always the best. Nor is it a way of proving anything (despite the rhetorical language it employs). At best it is evidence in an ongoing persuasive argument. The separation of statistics from the research process in general may in fact be responsible for the over-exalted status of statistics as a research tool.

Why then write a book that reinforces this separation? First, there is the simple fact that no single book can do everything. Indeed, other books exist which detail the issues involved in research, and the place of statistical analysis in the broader research process. Rather than duplicating such efforts this book is meant to sit side by side with such texts, and provide the methods of statistical analysis when required. Second, statistical analysis is hard. It raises distinct issues and problems of its own which warrant a self-contained treatment.

To the researcher or students using this book I have included other material on the CD that accompanies this book, especially chapters on detailed SPSS procedures that were too specialized for the actual text but which may be of interest. I have also placed this material on the website www.sagepub.co.uk/argyrous, to which I will be adding more material (including a list of corrections to any errors that may be discovered) over time, so you may wish to check this site periodically for such new material.

To the instructor, I have a wealth of material available for you at your request. This includes PowerPoint slides, Flash presentations, complete web pages for use in on-line courses, and a database of over 500 WebCT quiz questions that can be used for testing students and also for providing tutorial exercises. Please feel free to contact me at the address below and I will forward to you any (or all) of this material.

In the preparation of this edition I have been greatly assisted by the comments of many people who read the previously published versions of this text, and my thanks go to them. I do wish to specifically thank Ji In Lee for a thorough reading of the previous edition and

suggestions, to Punitha Arjunan for compiling the index and comments on the manuscript, and to Paul Francis and Peta Kennedy for comments on the manuscript of this edition. I am indebted to the Longman Group UK Ltd, on behalf of the Literary Executor of the late Sir Ronald Fisher and Dr Frank Yates FRS, for permission to reproduce Tables III, IV, and V from *Statistical Tables for Biological, Agricultural, and Medical Research*, 6/e (1974) in the Appendix, and to Professor A. Hald for permission to reproduce in amended form Table 1 of *Statistical Tables and Formulas 1952* in the Appendix.

Lastly, to the reader, I welcome any comments and criticisms, which can be passed on to me at the following address:

School of Social Science and Policy
University of New South Wales
NSW 2052, Australia
email: g.argyrous@unsw.edu.au

PART 1

An introduction to statistical analysis